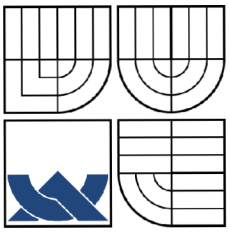


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ
ÚSTAV MATEMATIKY
FACULTY OF MECHANICAL ENGINEERING
INSTITUTE OF MATHEMATICS

STATISTICKÉ KLASIFIKAČNÍ METODY
STATISTICAL CLASSIFICATION METHODS

DIPLOMOVÁ PRÁCE
DIPLOMA THESIS

AUTOR PRÁCE
AUTHOR
Bc. OLDŘICH BARVENČÍK

VEDOUCÍ PRÁCE
SUPERVISOR
doc. RNDr. JAROSLAV MICHÁLEK, CSc.

BRNO 2010

Abstrakt

Práce se zabývá vybranými klasifikačními metodami. Jsou zde popsány základy shlukové analýzy, diskriminační analýzy a teorie klasifikačních stromů. Použití metod je ukázáno při klasifikaci simulovaných dat, výpočet je proveden v programu STATISTICA. V praktické části práce pak následuje porovnání metod při klasifikaci reálných datových souborů různých rozsahů. Klasifikačními metodami je také řešena reálná úloha – predikce znečištění ovzduší na základě předpovědi počasí.

Summary

The thesis deals with selected classification methods. The thesis describes the basis of cluster analysis, discriminant analysis and theory of classification trees. The usage is demonstrated by classification of simulated data, the calculation is made in the program STATISTICA. In practical part of the thesis there is the comparison of the methods for classification of real data files of various extent. Classification methods are used for solving of the real task – prediction of air pollution based of the weather forecast.

klíčová slova

klasifikační metody, shluková analýza, diskriminační analýza, klasifikační stromy

key words

classification methods, cluster analysis, discriminant analysis, classification trees.

BARVENČÍK, O.: *Statistické klasifikační metody*, Brno, Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2010. 63 s. Vedoucí diplomové práce doc. RNDr. Jaroslav Michálek, CSc.

Prohlašuji, že jsem diplomovou práci *Statistické klasifikační metody* vypracoval samostatně pod vedením doc. RNDr. Jaroslava Michálka, CSc. s použitím materiálů uvedených v seznamu literatury.

Bc. Oldřich Barvenčík

Děkuji doc. RNDr. Jaroslavu Michálkovi, CSc. za četné rady a připomínky během vedení mé diplomové práce. Také bych chtěl poděkovat své rodině, že mi umožnila studovat, a své přítelkyni za neustálou podporu.

Bc. Oldřich Barvenčík

Obsah

1 Úvod

Cílem této diplomové práce je uvést základní přehled statistických klasifikačních metod a jejich následné porovnání. Tyto metody mají široké využití, například v lékařství, kdy má lékař na základě několika symptomů rozhodnout, kterou nemocí pacient onemocněl. Nebo v biologii, když je třeba určit příslušnost pozorovaného exempláře k jednomu ze známých druhů. Klasifikační metody jsou rozvíjeny od 30. let minulého století a jsou spojeny například se jmény Fisher, K. Pearson, E. C. Pearson nebo Rao.

Ve druhé kapitole, která následuje po úvodu, je čtenář seznámen s formulací problému, jímž se budeme zabývat, a se základní charakteristikou dále popisovaných metod.

Třetí kapitola, zpracovaná podle [6], se již věnuje metodám shlukové analýzy. Jsou zde definovány základní pojmy, princip jednotlivých metod je ilustrován příklady. V závěrečné části kapitoly je ukázka výpočtu shlukování pomocí počítačového programu STATISTICA.

Následující čtvrtá kapitola popisuje metody diskriminační analýzy. Použití je demonstrováno příkladem, který je opět řešen pomocí počítače. Jsou zde však uvedeny vztahy, které umožňují i analytické řešení. Tato část práce byla zpracována podle [1] s přihlédnutím k [4] a [5].

V páté kapitole je pak popsána klasifikace pomocí klasifikačních stromů. Jsou zde definovány základní pojmy z této oblasti, dále se čtenář seznámí se dvěma metodami konstrukce klasifikačních stromů. Kapitola je zakončena řešeným příkladem, zpracována byla podle [3] a [7].

Šestá kapitola se věnuje aplikaci klasifikačních metod na reálná data. Klasifikace simulovaných dat je již součástí předcházejících kapitol. Zde je nejprve klasifikován datový soubor malého rozsahu převzatý z [5]. Výsledky ilustrované obrázky jsou pak porovnávány. Následuje klasifikace datového souboru většího rozsahu. Dále je modelována reálná úloha – predikce znečištění ovzduší pomocí předpovědi počasí, řešena je užitím vybraných klasifikačních metod.

V sedmé kapitole – Závěr jsou pak shrnuty dosažené cíle.

2 Úvod do klasifikačních metod

Statistické klasifikační metody patří mezi vícerozměrné statistické metody. Je dána množina objektů, označme ji $\Omega = \{\omega_1, \omega_2, \ldots\}$, k různých typů, $k \geq 2$. Množinu Ω tak tvoří k po dvou disjunktních skupin. Na každém objektu pozorujeme dvojici statistických znaků $\mathbf{X}$ a Z. Vektor $\mathbf{X}$ popisuje daný objekt a veličina Z určuje, jakého je typu, $Z = i$ právě tehdy, když je objekt i-tého typu. Složky vektoru $\mathbf{X}$ nazýváme prediktory.

Definice 2.1. Soubor dat na N objektech $(\mathbf{x}_1^\top, z_1)^\top, \ldots, (\mathbf{x}_N^\top, z_N)^\top$, kde $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^n$ a $z_i \in \mathcal{C} = \{1, 2, \ldots, k\},\ i = 1, 2, \ldots, N$, nazveme učebním souborem $\mathcal{S}$.

Může nastat situace, kdy vektory $\mathbf{x}_i$, $i = 1, 2, \ldots, N$, nemají stejnou dimenzi. Pokud tomu tak není, mluvíme o souboru dat se standardní strukturou.

Protože vektor $\mathbf{X}$ nabývá hodnot z $\mathcal{X} \subset \mathbb{R}^n$, klasifikační pravidlo, které každý objekt zařadí do jedné z k skupin dostaneme rozkladem $\mathcal{X}$ na k tříd $A_1, A_2, \ldots, A_k$, pro které požadujeme, aby tvořily rozklad $\mathcal{X}$, tedy

$$\mathcal{X} = \bigcup_{i=1}^{k} A_i, \qquad A_i \cap A_j = \emptyset, \quad i \neq j.$$

Padnou-li hodnoty vektoru $\mathbf{X}$ klasifikovaného objektu do třídy A_i, pak tento objekt zařadíme do i-té skupiny (prohlásíme, že je i-tého typu). Toto klasifikační pravidlo označíme $\mathbb{A}$, množinu všech klasifikačních pravidel pak $\mathcal{A}$. Jedním z hlavních cílů, které si kladou metody popsané v následujícím textu, je vybrat z množiny $\mathcal{A}$ takovový rozklad $\mathcal{X}$, aby rozhodování bylo optimální. Tím máme zjednodušeně řečeno na mysli, aby byla pravděpodobnost, že bude objekt i-tého typu zařazen do i-té třídy co nejvyšší.

Metody shlukové analýzy se zabývají úlohou, jak rozložit základní množinu objektů Ω na její podmnožiny, kde by byly objekty ω sobě podobné. Shluková analýza tak objekty přímo neklasifikuje.

V teorii diskriminační analýzy tvoří vektor $\mathbf{X}$ a veličina Z náhodný vektor definovaný na pravděpodobnostním prostoru $(\Omega, \mathfrak{A}, P)$. Vycházíme z učebního souboru $\mathcal{S}$, pomocí něj formulujeme klasifikační pravidla. Další objekty pak klasifikujeme na základě zjištěných hodnot vektorů $\mathbf{x}$.

Podobně postupujeme při klasifikaci klasifikačním stromem, který rovněž konstruujeme pomocí učebního souboru. Výhodou je, že klasifikační stromy lze použít i v situaci, kdy jsou prediktory kategoriální i spojité.

3 Shluková analýza

3.1 Základní pojmy

Pojem shluku definujeme intuitivně, jedná se o množinu objektů, podmnožinu základní množiny Ω. Na shluk lze také nahlížet jako na výsledek aplikace shlukovacího algoritmu.

Zaměřme se na pojetí vzájemné podobnosti jednotlivých objektů a kvantitativní vyjádření této podobnosti, to je jeden ze základních problémů shlukové analýzy.

Definice 3.1. Je dána možina $\Omega = \{\omega_1, \omega_2, \ldots\}$. Zobrazení π přiřazující každé dvojici objektů (ω_i, ω_j) z Ω číslo $\pi(\omega_i, \omega_j)$ nazveme mírou podobnosti objektů, jestliže splňuje

$$\begin{aligned} &\pi(\omega_i, \omega_j) \geq 0, \quad \forall i, j \\ &\pi(\omega_i, \omega_j) = \pi(\omega_j, \omega_i), \quad \forall i, j \end{aligned}$$

Dále má smysl požadovat, aby π nabývalo své maximální hodnoty pro $\omega_i = \omega_i$, $\forall i = = 1, 2, \ldots$.

Míra podobnosti objektů je konstruována tak, že čím je větší hodnota $\pi(\omega_i, \omega_j)$, tím je větší vzájemná podobnost objektů ω_i a ω_j. Podobnostní vztah mezi objekty můžeme vyjádřit například tak, že přiřadíme objektům charakterizovaným vektorem $\mathbf{X}$ jako modely body n rozměrného euklidovského prostoru $\mathbb{E}_n$. Pak lze jako míru podobnosti objektů využít metriku, objekty jsou si pak tím podobnější, čím je vzdálenost jejich bodů v prostoru $\mathbb{E}_n$ menší.

Metrika ϱ je zobrazení $\varrho\colon \mathbb{E}_n \times \mathbb{E}_n \longrightarrow \mathbb{R}$, splňující pro libovolné body $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{E}_n$ následující čtyři axiomy:

$$\begin{aligned} &\varrho(\mathbf{a}, \mathbf{b}) = 0 \quad \Longleftrightarrow \quad \mathbf{a} = \mathbf{b}, \\ &\varrho(\mathbf{a}, \mathbf{b}) \geq 0, \\ &\varrho(\mathbf{a}, \mathbf{b}) = \varrho(\mathbf{b}, \mathbf{a}), \\ &\varrho(\mathbf{a}, \mathbf{c}) \leq \varrho(\mathbf{a}, \mathbf{b}) + \varrho(\mathbf{b}, \mathbf{c}). \end{aligned}$$

Čtvrtá podmínka je známá jako trojúhelníková nerovnost.

Poznámka. Předpokládejme body $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ a $\mathbf{b} = (b_1, b_2, \ldots, b_n)$, Euklidovská metrika $\varrho(\mathbf{a}, \mathbf{b})$ je dána vztahem

$$\varrho(\mathbf{a}, \mathbf{b}) = \left[\sum_{i=1}^{n} (a_i - b_i)^2 \right]^{\frac{1}{2}}.$$

Dále lze užít metriku $\varrho_1(\mathbf{a}, \mathbf{b})$, ta je dána předpisem

$$\varrho_1(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{n} |a_i - b_i|.$$

Obě uvedené metriky jsou zvláštním případem Minkovského metriky:

$$\varrho_p(\mathbf{a}, \mathbf{b}) = \left[\sum_{i=1}^{n} (a_i - b_i)^p \right]^{\frac{1}{p}}, \quad p \geq 1.$$

U většiny shlukovacích postupů je však vhodnější vycházet z duálního pojmu míry podobnosti objektů.

Definice 3.2. Zobrazení d přiřazující každé dvojici objektů (ω_i, ω_j) z množiny objektů $\Omega = \{\omega_1, \omega_2, \ldots\}$ číslo $d(\omega_i, \omega_j)$ nazveme koeficientem nepodobnosti objektů, jestliže splňuje

$$\begin{aligned} &d(\omega_i, \omega_j) = 0 \quad \Longleftrightarrow \quad \omega_i = \omega_j, \\ &d(\omega_i, \omega_j) \geq 0, \\ &d(\omega_i, \omega_j) = d(\omega_j, \omega_i). \end{aligned}$$

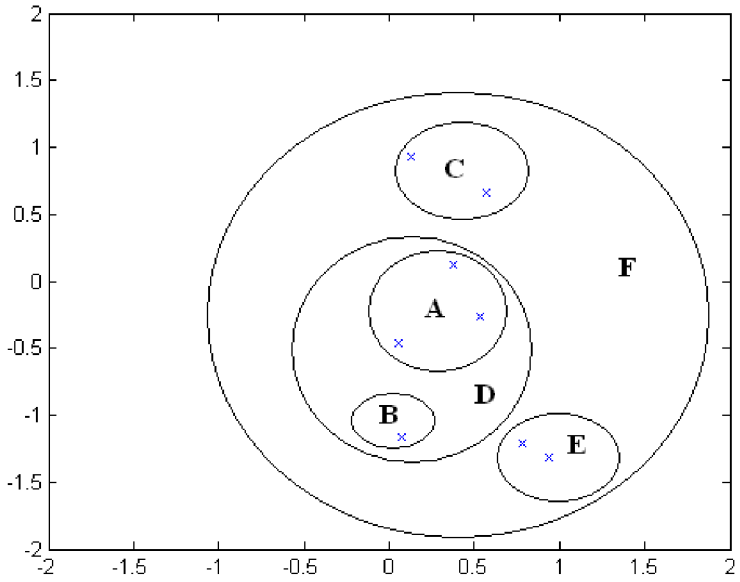
Nepožadujeme zde splnění trojúhelníkové nerovnosti, koeficient nepodobnosti d je tedy semimetrikou.

Jako koeficienty nepodobnosti objektů lze opět využít výše uvedené metriky.

3.2 Hierarchické aglomerativní shlukování

Metody shlukové analýzy je možno rozdělit na hierarchické a nehierarchické metody, z nichž první směřují k hierarchické klasifikaci, druhé k nehierarchické klasifikaci.

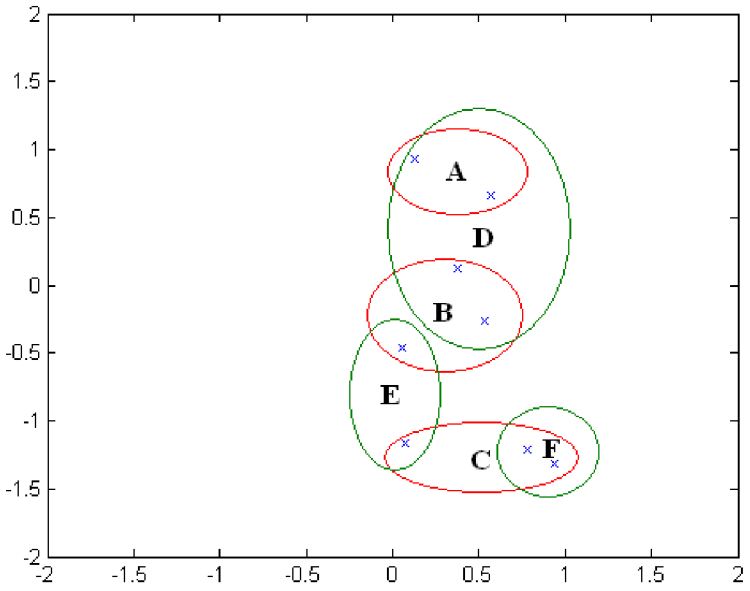
Definice 3.3. Hierarchickým shlukováním nazveme systém navzájem různých neprázdných podmnožin množiny objektů Ω, v němž průnikem každých dvou podmnožin je buď jedna z nich, nebo prázdná množina a v němž existuje alespoň jedna dvojice podmnožin, jejichž průnikem je jedna z nich.



Obrázek 1: Příklad hierarchického shlukování.

Hierarchické shlukování má tak zpravidla charakter posloupnosti rozkladů množiny objektů, v níž každý rozklad je zjemněním rozkladu následujícího.

Příklad shlukování, které není hierarchické, je znázorněn na Obrázku 2. Barevně jsou zde vyznačeny dva rozklady množiny objektů Ω, rozklad $\mathbb{A}_1$ obsahuje shluky A, B a C,



Obrázek 2: Příklad nehierarchického shlukování.

rozklad $\mathbb{A}_2$ pak shluky D, E a F. V dalším textu se budeme zabývat výhradně shlukováním hierarchickým.

Metody hierarchického shlukování lze dále rozdělit podle způsobu shlukování. Pokud provádíme shlukování tak, že vycházíme z množiny objektů Ω jako celku a jejím postupným rozdělováním získáváme hierarchický systém podmnožin, aplikujeme divizní přístup. Vycházíme-li naopak z jednotlivých objektů a jejich postupným seskupováním docházíme až ke konečnému stavu, kterým je spojení všech objektů do jedné množiny, jedná se o hierarchické aglomerativní shlukování.

Při algoritmu shlukování metodou tohoto druhu tvoří na počátku každý objekt dané množiny Ω jednoprvkový shluk. Tento počáteční rozklad množiny objektů na N jednoprvkových shluků označíme jako nultý rozklad $\mathbb{A}_0$ množiny Ω. K postupnému vytváření dalších rozkladů množiny Ω je třeba předem definovat způsob kvantitativního hodnocení podobnostních vztahů mezi shluky (viz dále). V prvním kroku shlukování pak vybereme ty dva shluky, které jsou si ve smyslu naší definice podobnosti shluků nejpodobnější. Tyto dva shluky sloučíme a vytvoříme tak nový shluk. Tento nově vytvořený shluk pak spolu se zbývajícími shluky o jednom objektu tvoří první rozklad $\mathbb{A}_1$ množiny Ω. Dále postupujeme analogicky. Nechť $\mathbb{A}_s$ je s-tý rozklad ($s \geq 1$) množiny Ω. Každý z dosud vytvořených rozkladů snížil počet shluků o jeden, s-tý rozklad obsahuje tedy $(N - s)$ shluků. Tento rozklad se skládá z nového shluku vzniklého sloučením dvou vzájemně si nejpodobnějších shluků předcházejícího $(s-1)$-ního rozkladu $\mathbb{A}_{s-1}$ a z ostatních nezměněných shluků rozkladu $\mathbb{A}_{s-1}$. Proto k porovnání vzájemných podobností shluků stačí určit podle dané definice hodnoty podobností nového shluku s ostatními shluky rozkladu a zbývající hodnoty vzájemných podobností nezměněných shluků převzít z předcházejícího $(s-1)$-ního shlukovacího kroku. Na základě porovnání hodnot vzájemných podobností všech $(N - s)$ shluků s-tého rozkladu $\mathbb{A}_s$ vybereme opět dva shluky, které jsou si nejpodobnější, a sloučíme je v jeden shluk. Ten pak spolu s ostatními nezměněnými shluky s-tého rozkladu $\mathbb{A}_s$ tvoří $(s+1)$-ní rozklad množiny Ω. Tento postup opakujeme, až dospějeme k poslednímu rozkladu $\mathbb{A}_{N-1}$ o jediném shluku obsahujícím všechny shlukované objekty.

Definice 3.4. Jestliže při hierarchickém shlukování obsahuje posloupnost rozkladů rozklady $\mathbb{A}_0 = \big\{\{\omega_1\}, \{\omega_2\}, \ldots, \{\omega_N\}\big\}$ a $\mathbb{A}_{N-1} = \{\Omega\}$, nazveme toto shlukování úplným hiererchickým shlukováním.

Pro výše popsaný postup hierarchického aglomerativního shlukování je nezbytné kvantitativní ohodnocení podobnostních vztahů mezi shluky.

Definice 3.5. Zobrazení D přiřazující každé dvojici shluků (A_i, A_j) v daném rozkladu $\mathbb{A}_r = \{A_1, A_2, \ldots\},\ r = 0, \ldots, N-2$, číslo $D(A_i, A_j)$, které splňuje pro všechny shluky $A_i, A_j \in \mathbb{A}_r$

$$D(A_i, A_i) = 0, \tag{3.1}$$
$$D(A_i, A_j) \geq 0, \tag{3.2}$$
$$D(A_i, A_j) = D(A_j, A_i), \tag{3.3}$$

nazveme koeficientem nepodobnosti shluků daného rozkladu.

Shlukování je možno znázornit také graficky. Diagram, který o průběhu shlukování obsahuje úplnou a jednoznačnou informaci, se nazývá dendrogram. Ve dvou vzájemně kolmých směrech jsou zde zaznamenány monotónní posloupnost koeficientů nepodobnosti shluků, při nichž došlo ke sloučení některých shluků, a pořadová čísla objektů seřazena tak, aby bylo možno znázornit postupné slučování shluků (viz například Obrázek 3).

Nyní se již můžeme zaměřit na nejznámější metody zavedení koeficientu nepodobnosti shluků.

3.2.1 Metoda nejbližšího souseda

Definice 3.6. Nechť d je libovolný koeficient nepodobnosti objektů a A, B jsou shluky rozkladu $\mathbb{A}$. Pak

$$D_{nn}(A, B) = \min_{\substack{\omega_i \in A \\ \omega_j \in B}} \{d(\omega_i, \omega_j)\}$$

nazveme koeficientem nepodobnosti shluků definovaným na základě koeficientu nepodobnosti objektů d metodou nejbližšího souseda (nearest neighbour).

Metoda bývá v anglicky psané literatuře označována „single linkage“. Koeficient D_{nn} splňuje podmínky (3.1-2.3) pro každý koeficient nepodobnosti objektů d. Princip shlukování metodou nejbližšího souseda ilustrujeme na jednoduchém příkladu.

Příklad 3.7. Je dána množina množina objektů $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$. Dále předpokládejme, že koeficienty nepodobnosti jednotlivých objektů jsou dány následující tabulkou

$d(\omega_i, \omega_j)$	ω_1	ω_2	ω_3	ω_4	ω_5
ω_1	0				
ω_2	9	0			
ω_3	3	7	0		
ω_4	6	5	9	0	
ω_5	11	10	2	8	0

Proveďme hierarchické aglomerativní shlukování metodou nejbližšího souseda.

Rozklad $\mathbb{A}_0$ tvoří jednotlivé objekty, které považujeme za jednoprvkové shluky. Koeficient D_{nn} je tedy totožný s d.

$$\min_{i,j} d(\omega_i, \omega_j) = d(\omega_3, \omega_5) = 2.$$

Objekty ω_3, ω_5 tak utvoří první shluk, dostáváme $\mathbb{A}_1 = \big\{\{\omega_3, \omega_5\},\, \omega_1,\, \omega_2,\, \omega_4\big\}$. Určíme nové koeficienty nepodobnosti shluků z $\mathbb{A}_1$:

$$\begin{aligned}
D_{nn}\big(\{\omega_3, \omega_5\},\, \{\omega_1\}\big) &= \min\big(d(\omega_1, \omega_3), d(\omega_1, \omega_5)\big) = \min(3, 11) = 3,\\
D_{nn}\big(\{\omega_3, \omega_5\},\, \{\omega_2\}\big) &= \min\big(d(\omega_2, \omega_3), d(\omega_2, \omega_5)\big) = \min(7, 10) = 7,\\
D_{nn}\big(\{\omega_3, \omega_5\},\, \{\omega_4\}\big) &= \min\big(d(\omega_3, \omega_4), d(\omega_4, \omega_5)\big) = \min(9, 8) = 8.
\end{aligned}$$

Dostáváme tak

D_{nn}	$\{\omega_3, \omega_5\}$	$\{\omega_1\}$	$\{\omega_2\}$	$\{\omega_4\}\}$
$\{\omega_3, \omega_5\}$	0			
$\{\omega_1\}$	3	0		
$\{\omega_2\}$	7	9	0	
$\{\omega_4\}$	8	6	5	0

Nejnižší hodnotu má koeficient $D_{nn}\big(\{\omega_3, \omega_5\},\, \{\omega_1\}\big) = 3$, další rozklad má tvar $\mathbb{A}_2 = \big\{\{\omega_3, \omega_5, \omega_1\},\, \{\omega_2\},\, \{\omega_4\}\big\}$. Opět určíme nové koeficienty nepodobnosti shluků, dostáváme

$$\begin{aligned}
D_{nn}\big(\{\omega_3, \omega_5, \omega_1\},\, \{\omega_2\}\big) &= \min\big(d(\omega_1, \omega_2), d(\omega_3, \omega_2), d(\omega_5, \omega_2)\big) = \min(9, 7, 10) = 7,\\
D_{nn}\big(\{\omega_3, \omega_5, \omega_1\},\, \{\omega_4\}\big) &= \min\big(d(\omega_1, \omega_4), d(\omega_3, \omega_4), d(\omega_5, \omega_4)\big) = \min(6, 9, 8) = 6.
\end{aligned}$$

Vypočtené koeficienty dosadíme do tabulky

D_{nn}	$\{\omega_3, \omega_5, \}$	$\{\omega_2\}$	$\{\omega_4\}\}$
$\{\omega_3, \omega_5, \omega_1\}$	0		
$\{\omega_2\}$	7	0	
$\{\omega_4\}$	6	5	0

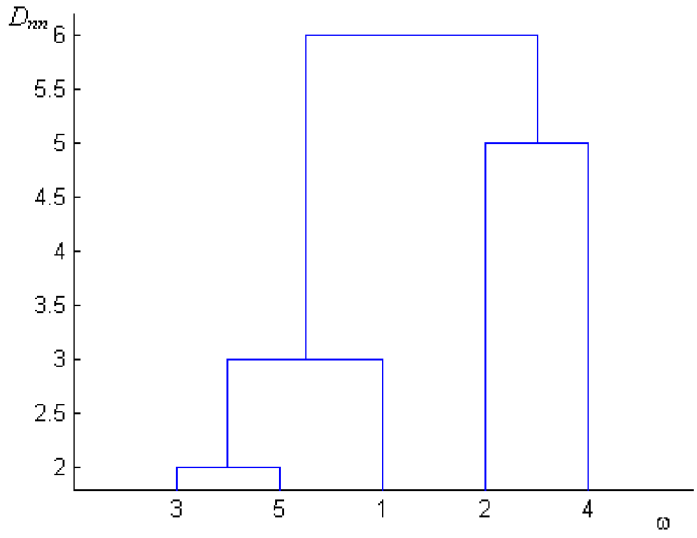
Z koeficientů nepodobnosti shluků v tabulce je má nejnižší hodnotu $D_{nn}\big(\{\omega_2\}, \{\omega_4\}\big) = 5$, dostáváme rozklad $\mathbb{A}_3 = \big\{\{\omega_3, \omega_5, \omega_1\},\, \{\omega_2, \omega_4\}\big\}$. Určíme koeficient nepodobnosti shluků pro shluky rozkladu $\mathbb{A}_3$:

$$D_{nn}\big(\{\omega_3, \omega_5, \omega_1\},\, \{\omega_2, \omega_4\}\big) = \min_{\substack{i\in\{1,3,5\}\\ j\in\{2,4\}}} \big(d(\omega_i, \omega_j)\big) = \min(9, 6, 7, 9, 10, 8) = 6.$$

Tabulka koeficientů D_{nn} tak dostává tvar

D_{nn}	$\{\omega_3, \omega_5, \omega_1\}$	$\{\omega_2, \omega_4\}$
$\{\omega_3, \omega_5, \omega_1\}$	0	
$\{\omega_2, \omega_4\}$	6	0

V posledním kroku utvoří všechny objekty původní množiny Ω jeden shluk, tedy $\mathbb{A}_4 = \big\{\{\omega_3, \omega_5, \omega_1, \omega_2, \omega_4\}\big\}$. Průběh shlukování si můžeme prohlédnout na dendrogramu (Obrázek 3).



Obrázek 3: Dendrogram, shlukování metodou nejbližšího souseda.

3.2.2 Metoda nejvzdálenějšího souseda

Definice 3.8. Nechť d je libovolný koeficient nepodobnosti objektů a A, B jsou shluky rozkladu $\mathbb{A}$. Pak

$$D_{fn}(A,B) = \max_{\substack{\omega_i \in A \\ \omega_j \in B}} \{d(\omega_i, \omega_j)\}, \quad \text{pro } A \neq B,$$
$$D_{fn}(A,A) = 0$$

nezveme koeficientem nepodobnosti shluků definovaným na základě koeficientu d nepodobnosti objektů metodou nejvzdálenějšího souseda (furthest neighbour).

V anglicky psané literatuře bývá metoda nazývána „complete linkage“. Takto zavedené zobrazení D_{fn} opět splňuje podmínky (3.1-2.3) pro libovolný koeficient nepodobnosti d. Nyní opět ilustrujme princip shlukování metodou nejvzdálenějšího souseda, zadání je stejné jako v předchozím příkladu.

Příklad 3.9. Je dána množina množina objektů $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$. Opět předpokládejme, že koeficienty nepodobnosti jednotlivých objektů jsou dány tabulkou

$d(\omega_i, \omega_j)$	ω_1	ω_2	ω_3	ω_4	ω_5
ω_1	0				
ω_2	9	0			
ω_3	3	7	0		
ω_4	6	5	9	0	
ω_5	11	10	2	8	0

Proveďme hierarchické aglomerativní shlukování metodou nejvzdálenějšího souseda.

Začátek postupu je stejný jako v předchozím případě, rozklad $\mathbb{A}_0$ tvoří jednotlivé objekty, které považujeme za jednoprvkové shluky. Koeficient D_{fn} je totožný s d.

$$\min_{i,j} d(\omega_i, \omega_j) = d(\omega_3, \omega_5) = 2.$$

Objekty ω_3, ω_5 tak utvoří první shluk, pak $\mathbb{A}_1 = \big\{\{\omega_3, \omega_5\}, \omega_1, \omega_2, \omega_4\big\}$. Určíme nové koeficienty nepodobnosti shluků z $\mathbb{A}_1$:

$$D_{fn}\big(\{\omega_3, \omega_5\}, \{\omega_1\}\big) = \max\big(d(\omega_1, \omega_3), d(\omega_1, \omega_5)\big) = \max(3, 11) = 11,$$
$$D_{fn}\big(\{\omega_3, \omega_5\}, \{\omega_2\}\big) = \max\big(d(\omega_2, \omega_3), d(\omega_2, \omega_5)\big) = \max(7, 10) = 10,$$
$$D_{fn}\big(\{\omega_3, \omega_5\}, \{\omega_4\}\big) = \max\big(d(\omega_3, \omega_4), d(\omega_4, \omega_5)\big) = \max(9, 8) = 9.$$

Dosazením obdržíme

D_{fn}	$\{\omega_3, \omega_5\}$	$\{\omega_1\}$	$\{\omega_2\}$	$\{\omega_4\}\}$
$\{\omega_3, \omega_5\}$	0			
$\{\omega_1\}$	11	0		
$\{\omega_2\}$	10	9	0	
$\{\omega_4\}$	5	6	5	0

Nejnižší hodnotu má nyní koeficient $D_{fn}\big(\{\omega_2\}, \{\omega_4\}\big) = 5$, další rozklad má tak tvar $\mathbb{A}_2 = \big\{\{\omega_3, \omega_5\}, \{\omega_2, \omega_4\}, \{\omega_1\}\big\}$. Dále postupujeme analogicky:

$$D_{fn}\big(\{\omega_3, \omega_5\}, \{\omega_2, \omega_4\}\big) = \max_{\substack{i\in\{3,5\}\\ j\in\{2,4\}}} \big(d(\omega_i, \omega_j)\big) = \max(7, 10, 9, 8) = 10,$$
$$D_{fn}\big(\{\omega_2, \omega_4\}, \{\omega_1\}\big) = 9.$$

Sestavíme tabulku koeficientů D_{fn} a další rozklad:

D_{fn}	$\{\omega_3, \omega_5\}$	$\{\omega_2, \omega_4\}$	$\{\omega_1\}$
$\{\omega_3, \omega_5\}$	0		
$\{\omega_2, \omega_4\}$	10	0	
$\{\omega_1\}$	11	9	0

$\mathbb{A}_3 = \big\{\{\omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_1\}\big\}$. V dalším kroku dostáváme

$$D_{fn}\big(\{\omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_1\}\big) = 11,$$

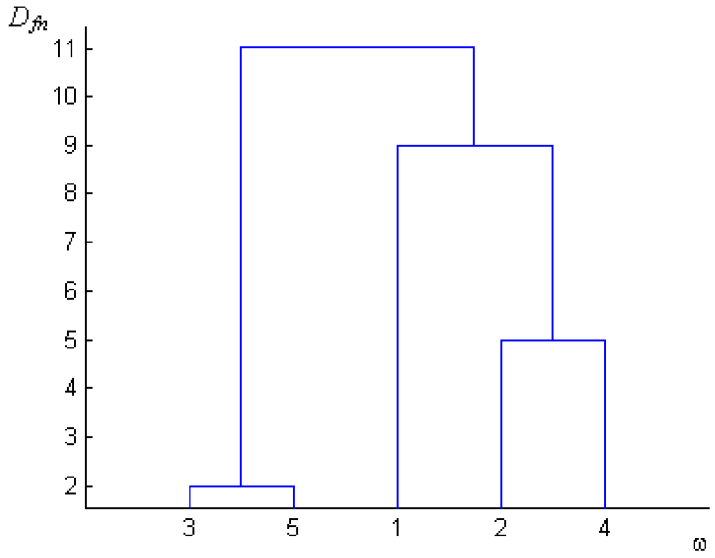
D_{fn}	$\{\omega_3, \omega_5, \}$	$\{\omega_2, \omega_4, \omega_1\}$
$\{\omega_3, \omega_5\}$	0	
$\{\omega_2, \omega_4, \omega_1\}$	11	0

V posledním kroku všechny objekty původní množiny Ω tvoří opět jeden shluk: $\mathbb{A}_4 = \big\{\{\omega_3, \omega_5, \omega_2, \omega_4, \omega_1\}\big\}$. Na závěr ještě zkonstruujeme dendrogram, viz Obrázek 4.

3.2.3 Metoda průměrné nepodobnosti objektů

Definice 3.10. Nechť d je libovolný koeficient nepodobnosti objektů. Dále předpokládejme, že $A = \{\omega_1, \omega_2, \ldots, \omega_k\}$ a $B = \{\omega'_1, \omega'_2, \ldots, \omega'_l\}$ jsou shluky rozkladu $\mathbb{A}$. Pak

$$D_{av}(A, B) = \frac{1}{kl} \sum_{i=1}^{k} \sum_{j=1}^{l} d(\omega_i, \omega'_j), \quad \text{pro } A \neq B,$$
$$D_{av}(A, A) = 0$$



Obrázek 4: Dendrogram, shlukování metodou nejvzdálenějšího souseda.

nezveme koeficientem nepodobnosti shluků definovaným na základě koeficientu d nepodobnosti objektů metodou průměrné nepodobnosti objektů.

Také koeficient D_{av} splňuje podmínky (3.1-3.3). Princip metody ilustrujeme na stejném příkladu.

Příklad 3.11. Je dána množina množina objektů $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$. Koeficienty nepodobnosti jednotlivých objektů jsou dány tabulkou

$d(\omega_i, \omega_j)$	ω_1	ω_2	ω_3	ω_4	ω_5
ω_1	0				
ω_2	9	0			
ω_3	3	7	0		
ω_4	6	5	9	0	
ω_5	11	10	2	8	0

Proveďme hierarchické aglomerativní shlukování metodou průměrné nepodobnosti objektů.

Postupujeme podobně jako v předchozích případech, $\mathbb{A}_0$ tvoří jednotlivé objekty, které považujeme za jednoprvkové shluky. Koeficient D_{av} je totožný s d.

$$\min_{i,j} d(\omega_i, \omega_j) = d(\omega_3, \omega_5) = 2.$$

Objekty ω_3, ω_5 opět utvoří první shluk. $\Omega_1 = \big\{\{\omega_3, \omega_5\}, \omega_1, \omega_2, \omega_4\big\}$, určíme nové koeficienty nepodobnosti shluků z $\mathbb{A}_1$.

$$D_{av}\big(\{\omega_3, \omega_5\}, \{\omega_1\}\big) = \frac{1}{2}(d(\omega_3, \omega_1) + d(\omega_5, \omega_1)) = \frac{1}{2}(3 + 11) = 7,$$
$$D_{av}\big(\{\omega_3, \omega_5\}, \{\omega_2\}\big) = \frac{1}{2}(d(\omega_3, \omega_2) + d(\omega_5, \omega_2)) = \frac{1}{2}(7 + 10) = 8,5,$$
$$D_{av}\big(\{\omega_3, \omega_5\}, \{\omega_4\}\big) = \frac{1}{2}(d(\omega_3, \omega_4) + d(\omega_5, \omega_4)) = \frac{1}{2}(9 + 8) = 8,5,$$

dostáváme tak

D_{av}	$\{\omega_3,\omega_5\}$	$\{\omega_1\}$	$\{\omega_2\}$	$\{\omega_4\}\}$
$\{\omega_3,\omega_5\}$	0			
$\{\omega_1\}$	7	0		
$\{\omega_2\}$	8,5	9	0	
$\{\omega_4\}$	8,5	6	5	0

Nejnižší hodnotu má nyní koeficient $D_{av}\big(\{\omega_2\},\{\omega_4\}\big) = 5$, další rozklad má tvar $\mathbb{A}_2 = \big\{\{\omega_3,\omega_5\},\ \{\omega_2,\omega_4\},\ \{\omega_1\}\big\}$. Analogickým postupem dostáváme

$$D_{av}\big(\{\omega_3,\omega_5\},\ \{\omega_2,\omega_4\}\big) = \frac{1}{4}(7+9+10+8) = 8,5,$$
$$D_{av}\big(\{\omega_2,\omega_4\},\ \{\omega_1\}\big) = \frac{1}{2}(9+6) = 7,5.$$

D_{av}	$\{\omega_3,\omega_5\}$	$\{\omega_2,\omega_4\}$	$\{\omega_1\}$
$\{\omega_3,\omega_5\}$	0		
$\{\omega_2,\omega_4\}$	8,5	0	
$\{\omega_1\}$	7	7,5	0

Další rozklad je tedy tvaru $\mathbb{A}_3 = \big\{\{\omega_3,\omega_5\},\ \{\omega_2,\omega_4,\omega_1\}\big\}$. Určíme poslední koeficient nepodobnosti shluků a dosadíme jej do tabulky.

$$D_{av}\big(\{\omega_3,\omega_5,\omega_1\},\{\omega_2,\omega_4\}\big) = \frac{1}{6}(7+10+9+9+8+6) = 8,2,$$

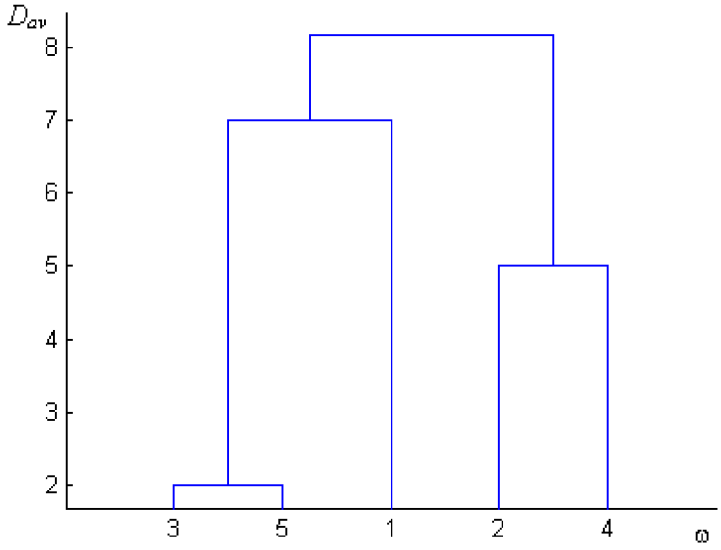
D_{av}	$\{\omega_3,\omega_5,\omega_1\}$	$\{\omega_2,\omega_4\}$
$\{\omega_3,\omega_5,\omega_1\}$	0	
$\{\omega_2,\omega_4\}$	8,2	0

Na konci shlukování obsahuje rozklad opět jen jeden shluk tvořený všemi objekty, $\mathbb{A}_4 = \big\{\{\omega_3,\omega_5,\omega_2,\omega_4,\omega_1\}\big\}$. Příslušný dendrogram je na Obrázku 5.

Z uvedených příkladů je patrné, že hierarchické aglomerativní shlukování nemusí být jednoznačné. V některém kroku může nastat situace, kdy minimální hodnoty dosáhne koeficient D pro více dvojic shluků. Je tedy nutné předem stanovit pravidlo, podle nějž se vybere jediná. Nejčastěji se volí dvojice shluků s nejvyššími (nejnižšími) pořadovými čísly. Každé takovéto pravidlo však způsobuje závislost výsledného shlukování na počátečním uspořádání objektů. Tato nejednoznačnost výsledku je závažným nedostatkem metod hierarchického aglomerativního shlukování.

3.3 Výpočet shlukování v programu STATISTICA

Na závěr této kapitoly ukážeme výpočet shlukování pomocí statistického softwaru. Je dán datový soubor (Obrázek 6, viz také příloha P1 – data_sim.sta, data_sim.xls), který vznikl složením tří náhodných výběrů rozsahu 30 z dvojrozměrného normálního rozdělení s různými vektory středních hodnot a se stejnou varianční maticí, data byla vygenerována v systému MATLAB. Máme tedy soubor o 90 pozorováních a pomocí shlukování se pokusíme určit, z kterého náhodného výběru je které pozorování.



Obrázek 5: Dendrogram, shlukování metodou průměrné nepodobnosti objektů.

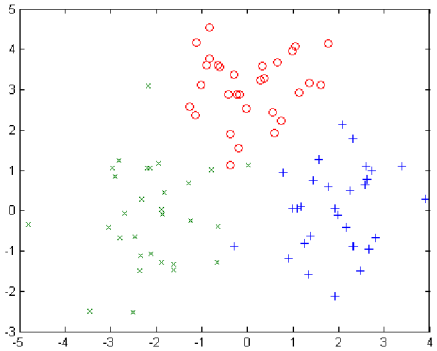
Na Obrázku 6 jsou modře znázorněna pozorování z $\mathrm{N}_2(\boldsymbol{\mu}_1, \mathbf{V})$, zeleně pozorování z $\mathrm{N}_2(\boldsymbol{\mu}_2, \mathbf{V})$ a konečně červeně data z $\mathrm{N}_2(\boldsymbol{\mu}_3, \mathbf{V})$, kde

$$\boldsymbol{\mu}_1 = (2,0)^\top, \quad \boldsymbol{\mu}_2 = (-2,0)^\top, \quad \boldsymbol{\mu}_3 = (0,3)^\top.$$

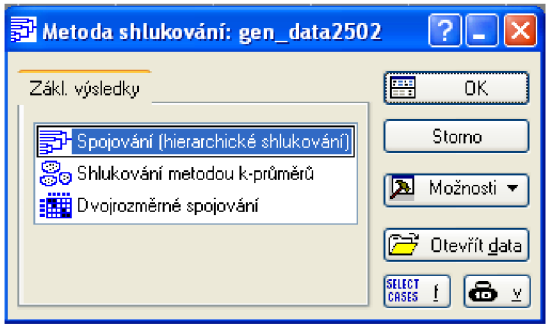
Pro každé pozorování definujeme také veličinu Z, která určuje, z jakého rozdělení je které pozorování. Jednotlivá pozorování jsou označena P_1 až P_{90}, přičemž

$$\begin{aligned} \mathrm{P}_1, \ldots, \mathrm{P}_{30} &\sim \mathrm{N}_2(\boldsymbol{\mu}_1, \mathbf{V}), \ Z = 1, \\ \mathrm{P}_{31}, \ldots, \mathrm{P}_{60} &\sim \mathrm{N}_2(\boldsymbol{\mu}_2, \mathbf{V}), \ Z = 2, \\ \mathrm{P}_{61}, \ldots, \mathrm{P}_{90} &\sim \mathrm{N}_2(\boldsymbol{\mu}_3, \mathbf{V}), \ Z = 3. \end{aligned}$$

Pokud v programu STATISTICA zvolíme z dostupných metod *Shluková analýza*, objeví se úvodní nabídka, Obrázek 7.

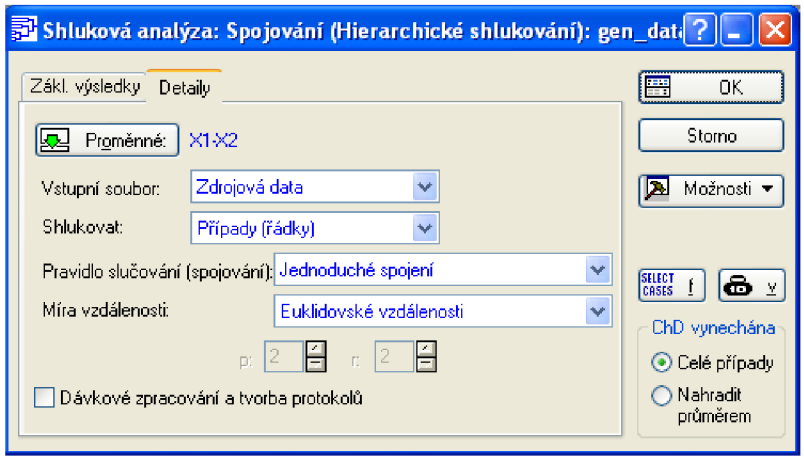


Obrázek 6: Datový soubor zakreslený do roviny.



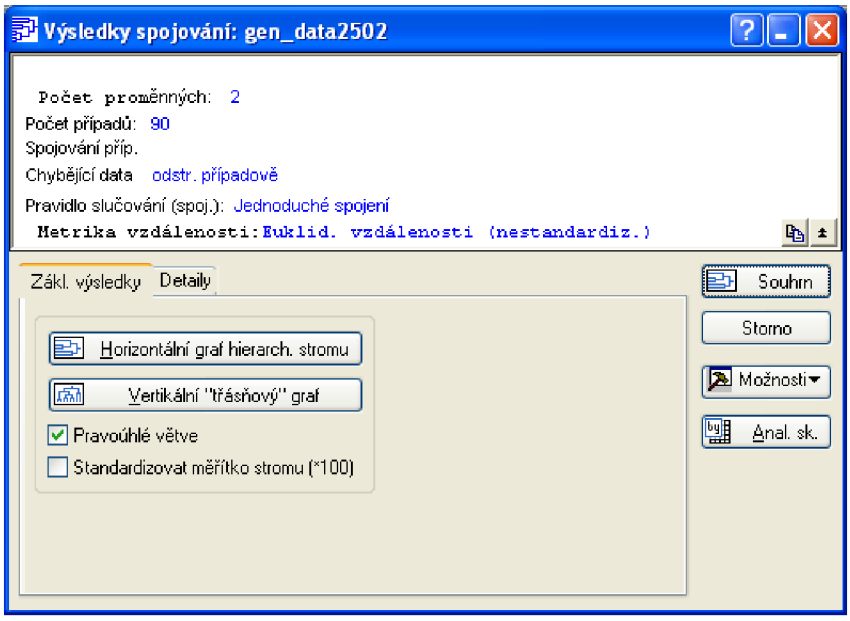
Obrázek 7: Úvodní nabídka.

Zde volíme *Spojování (hierarchické shlukování)*. Dále máme možnost stanovit parametry shlukovací procedury, Obrázek 8.



Obrázek 8: Volba parametrů shlukování.

Předně určíme prediktory, podle kterých se bude shlukování provádět. Vybíráme také koeficient nepodobnosti objektů (*Euklidovské vzálenosti*) a koeficient nepodobnosti shluků, *Jednoduché spojení* zde představuje shlukování metodou nejbližšího souseda. Nyní již obdržíme výsledky, viz Obrázek 9. Průběh shlukování si můžeme prohlédnout na dendro-

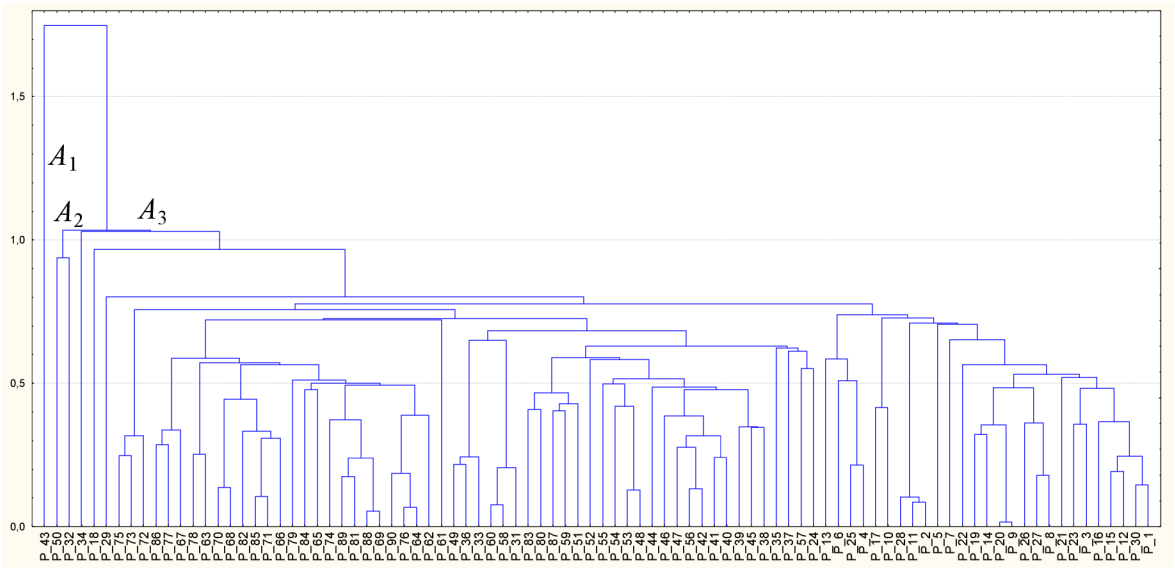


Obrázek 9: Výsledky.

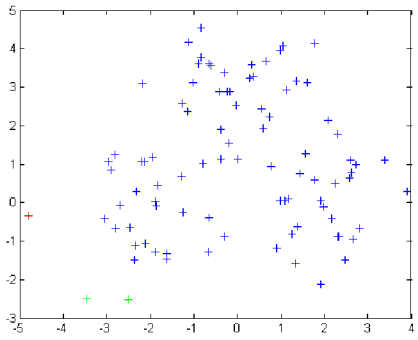
gramu (volba *Vertikální „třásňový“ graf*) – Obrázek 10. Protože jsme skupinu jednotlivých objektů chtěli rozdělit na tři podskupiny, shlukování přerušíme ve chvíli, kdy máme tři shluky A_1, A_2 a A_3, dostáváme tak tento výsledek

$$A_1 = \{\mathrm{P}_{43}\},$$
$$A_2 = \{\mathrm{P}_{32}, \mathrm{P}_{50}\},$$
$$A_3 = \{\mathrm{P}_1, \ldots, \mathrm{P}_{31}, \mathrm{P}_{33}, \ldots, \mathrm{P}_{42}, \mathrm{P}_{44}, \ldots, \mathrm{P}_{49}, \mathrm{P}_{51}, \ldots, \mathrm{P}_{90}, \}.$$

Shluky rozlišené barvami jsou na Obrázku 11.



Obrázek 10: Dendrogram, metoda nejbližšího souseda.



Obrázek 11: Shlukování metodou nejbližšího souseda.

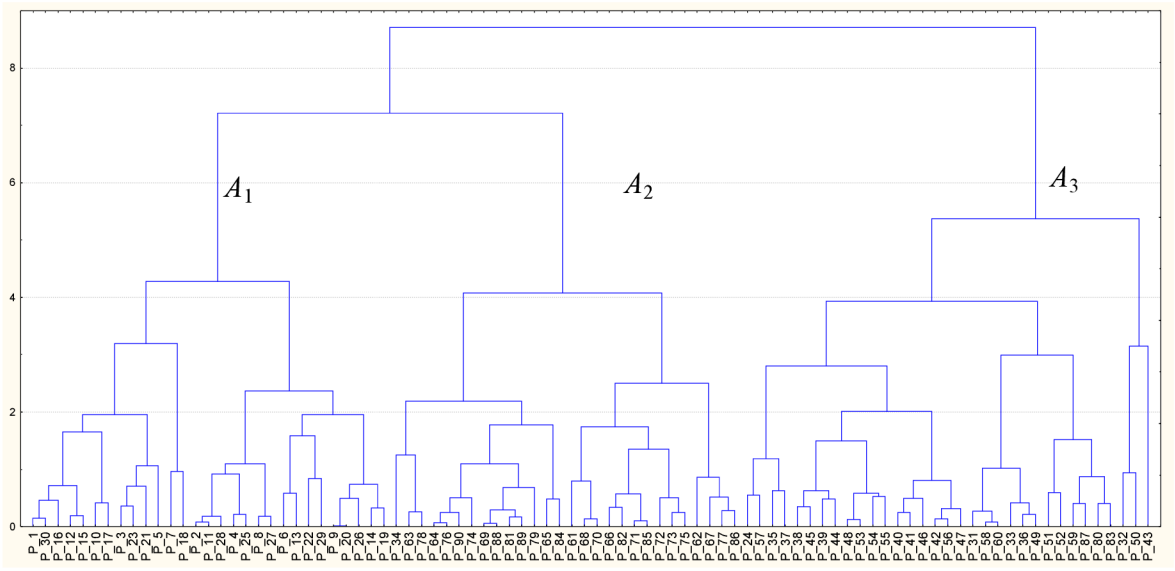
Postup nyní zopakujeme, jen koeficient nepodobnosti objektů zvolíme metodou nejvzdálenějšího souseda, v programu STATISTICA zadáme *Úplné spojení*. Výsledky se podstatně liší, dostáváme shluky

$$A_1 = \{\mathrm{P}_1, \ldots, \mathrm{P}_{23}, \mathrm{P}_{25}, \ldots, \mathrm{P}_{30}\},$$
$$A_2 = \{\mathrm{P}_{34}, \mathrm{P}_{61}, \ldots, \mathrm{P}_{79}, \mathrm{P}_{81}, \mathrm{P}_{82}, \mathrm{P}_{84}, \ldots, \mathrm{P}_{86}, \mathrm{P}_{88}, \ldots, \mathrm{P}_{90}\},$$
$$A_3 = \{\mathrm{P}_{24}, \mathrm{P}_{31}, \ldots, \mathrm{P}_{33}, \mathrm{P}_{35}, \ldots, \mathrm{P}_{60}, \mathrm{P}_{80}, \mathrm{P}_{83}, \mathrm{P}_{87}\}.$$

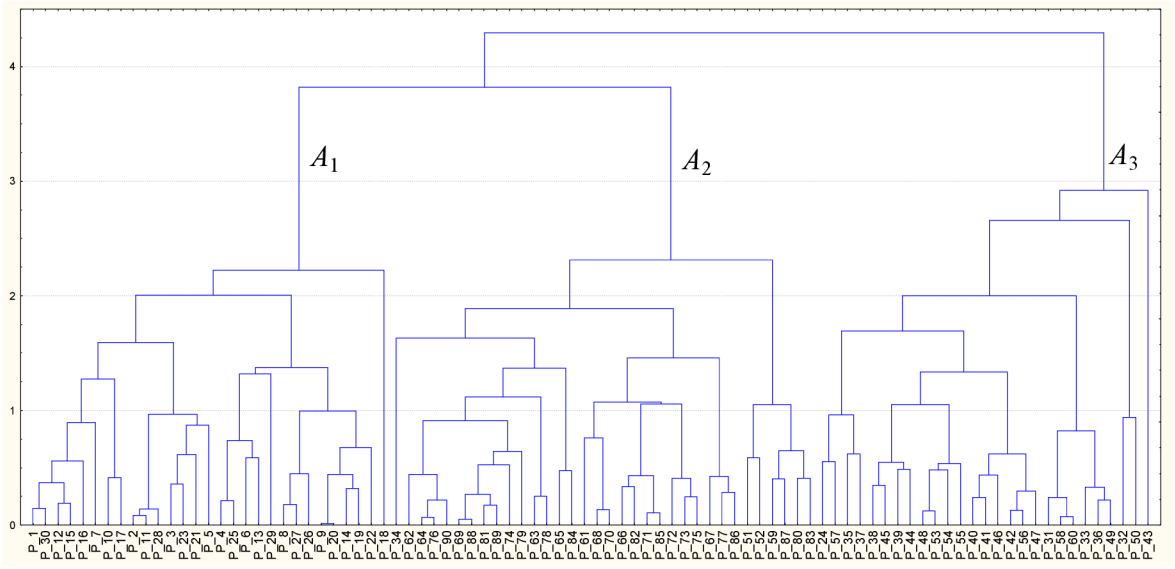
Dendrogram pro tuto metodu je na Obrázku 12. Shlukování podobným postupem spočteme i pro koeficient nepodobnosti shluků určený na základě průměrné nepodobnosti objektů. Zde je pod názvem *Nevážený průměr skupin dvojic*. Obdržíme tyto shluky:

$$A_1 = \{\mathrm{P}_1, \ldots, \mathrm{P}_{23}, \mathrm{P}_{25}, \ldots, \mathrm{P}_{30}\},$$
$$A_2 = \{\mathrm{P}_{34}, \mathrm{P}_{51}, \mathrm{P}_{52}, \mathrm{P}_{59}, \mathrm{P}_{61}, \ldots, \mathrm{P}_{90}\},$$
$$A_3 = \{\mathrm{P}_{24}, \mathrm{P}_{31}, \ldots, \mathrm{P}_{33}, \mathrm{P}_{35}, \ldots, \mathrm{P}_{50}, \mathrm{P}_{53}, \ldots, \mathrm{P}_{58}, \mathrm{P}_{60}\},$$

Příslušný dendrogram je na Obrázku 13.



Obrázek 12: Dendrogram, metoda nejvzdálenějšího souseda.



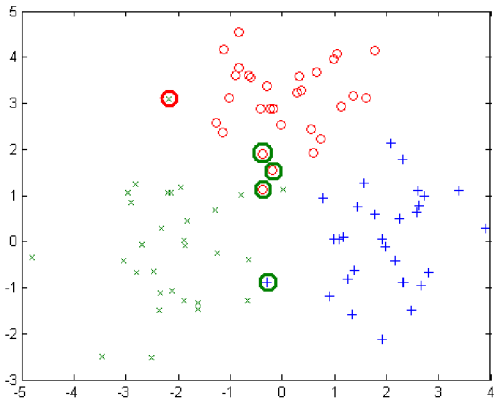
Obrázek 13: Dendrogram, metoda průměrné nepodobnosti objektů.

Výsledky jednotlivých shlukování můžeme nyní porovnat. Jako nejméně vhodné se pro tento případ jeví shlukování metodou nejbližšího souseda. Počty pozorování se zde v jednotlivých shlucích velmi liší, nelze tak ani těmto shlukům přiřadit rozdělení pravděpodobnosti, z kterého by měla daná pozorování být. Shlukování zbývajícími dvěma metodami však dopadlo o poznání lépe. Pozorování byla do třech shluků rozdělena téměř rovnoměrně a při shlukování metodou nejvzdálenějšího souseda i metodou průměrné nepodobnosti objektů padlo do „nesprávných“ shluků shodně 5 pozorování, což můžeme blíže analyzovat pomocí následujících tabulek, kde porovnáváme hodnoty veličiny Z a jejího odhadu $\widehat{Z}$ (tabulka vlevo je pro shlukování metodou nejvzdálenějšího souseda, vpravo pak pro shlukování metodou průměrné nepodobnosti objektů).

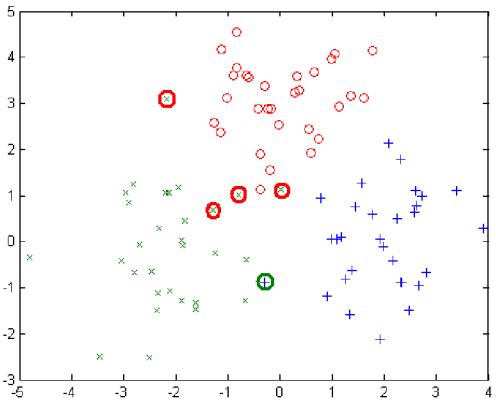
	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$
$Z=1$	29	1	0
$Z=2$	1	29	0
$Z=3$	0	3	27

	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$
$Z=1$	29	1	0
$Z=2$	0	26	4
$Z=3$	0	0	30

Tento výsledek je ilustrován dalšími dvěma obrázky. Vyznačena jsou zde nesprávně klasifikovaná pozorování (barva vyznačení je volena podle toho, ke kterým datům byla tato pozorování zařazena).



Obrázek 14: Metoda nejvzdálenějšího souseda.



Obrázek 15: Metoda průměrné nepodobnosti.

4 Diskriminační analýza

4.1 Základní pojmy

V této kapitole předpokládáme, že vektor $\mathbf{X}$ tvoří spolu s veličinou Z náhodný vektor $\mathbf{Y} = (\mathbf{X}^\top, Z)^\top$ definovaný na pravděpodobnostním prostoru $(\Omega, \mathfrak{A}, P)$ s hustotou $f_{\mathbf{X}Z}(\mathbf{x}, j)$ vzhledem k nějaké součinové míře $\mu = \mu_{\mathbf{X}} \times \nu_Z$, kde $\mu_{\mathbf{X}}$ je Lebesquova a ν_Z sčítací míra. Předpokládáme, že pro hustotu $f_{\mathbf{X}Z}$ platí

$$f_{\mathbf{X}Z}(\mathbf{x}, j) = p_j f_j(\mathbf{x}), \quad j = 1, 2, \ldots, k, \ \mathbf{x} \in \mathbb{R}^n,$$

kde

$$p_1 + \ldots + p_k = 1, \quad p_j > 0, \ j = 1, 2, \ldots, k,$$

nazveme apriorními pravděpodobnostmi. Dále předpokládáme, že $f_j(\mathbf{x}), \ j = 1, 2, \ldots, k$, $\mathbf{x} \in \mathbb{R}^n$, jsou hustoty pravděpodobností vzhledem k Lebesqueově míře. Pak dostaneme následující vztah pro hustotu marginálního rozdělení pravděpodobnosti náhodného vektoru $\mathbf{X}$

$$f_{\mathbf{X}}(\mathbf{x}) = \sum_{j=1}^{k} f_{\mathbf{X}Z}(\mathbf{x}, j) = \sum_{j=1}^{k} p_j f_j(\mathbf{x}).$$

Pro pravděpodobnostní funkci marginálního rozdělení náhodné veličiny Z platí

$$P(Z = j) = \int_{\mathbb{R}^n} f_{\mathbf{X}Z}(\mathbf{x}, j)\,\mathrm{d}\mathbf{x} = p_j \int_{\mathbb{R}^n} f_j(\mathbf{x})\,\mathrm{d}\mathbf{x} = p_j, \quad j = 1, 2, \ldots, k.$$

Dále lze vyjádřit podmíněnou hustotu pravděpodobnosti pro podmíněné rozdělení vektoru $\mathbf{X}$ pro $Z = j$

$$f_{\mathbf{X}|Z}(\mathbf{x}|Z = j) = \frac{f_{\mathbf{X}Z}(\mathbf{x}, j)}{f_Z(j)} = \frac{p_j f_j(\mathbf{x})}{p_j} = f_j(\mathbf{x}), \quad j = 1, 2, \ldots, k.$$

A podmíněnou pravděpodobnostní funkci pro podmíněné rozdělení náhodné veličiny Z pro $\mathbf{X} = \mathbf{x}$

$$f_{Z|\mathbf{X}}(j|\mathbf{X} = \mathbf{x}) = \frac{f_{\mathbf{X}Z}(\mathbf{x}, j)}{f_{\mathbf{X}}(\mathbf{x})} = \frac{p_j f_j(\mathbf{x})}{\sum_{i=1}^{k} p_i f_i(\mathbf{x})} = p_{j\mathbf{x}}, \ j = 1, 2, \ldots, k, \ \mathbf{x} \in \mathbb{R}^n.$$

Hodnoty $p_{j\mathbf{x}}, \ j = 1, 2, \ldots, k$, nazveme aposteriorními pravděpodobnostmi.

Definice 4.1. Řekneme, že $C(i, j) \in \mathbb{R}$ je chyba špatné klasifikace objektu ze třídy i do třídy j, jestliže platí

1. $C(i, j) \geq 0, \ i \neq j,$
2. $C(i, j) = 0, \ i = j, \ i, j = 1, 2, \ldots, k.$

Pokud objekt patří do i-té skupiny, dostáváme střední hodnotu chyby špatné klasifikace klasifikačním pravidlem $\mathbb{A}$

$$R(i) = E\big(C(i, j)\big) = \sum_{j=1}^{k} C(i, j) \int_{A_j} f_i(\mathbf{x})\,\mathrm{d}\mathbf{x}, \quad i = 1, 2, \ldots, k.$$

Pomocí apriorních pravděpodobností p_i, $i = 1, 2, \ldots, k$, můžeme nyní vyjádřit střední hodnotu chyby špatné klasifikace (již nepodmíněnou) klasifikačním pravidlem $\mathbb{A}$,

$$R(\mathbb{A}) = \sum_{i=1}^{k} p_i R(i).$$

Definice 4.2. Nechť $\mathbb{A}^* = \{A_1^*, A_2^*, \ldots, A_k^*\} \in \mathcal{A}$ je klasifikační pravidlo a $R(\mathbb{A}^*)$ jemu příslušná střední hodnota chyby špatné klasifikace. Pokud je splněna podmínka

$$R(\mathbb{A}^*) = \min_{\mathbb{A} \in \mathcal{A}} R(\mathbb{A}),$$

řekneme, že $\mathbb{A}^*$ je optimální klasifikační pravidlo.

Dále označme

$$q_j(\mathbf{x}) = \sum_{i=1}^{k} C(i, j) p_i f_i(\mathbf{x}),$$

tuto funkci nazveme j-tý skór vektoru $\mathbf{X}$. Pokud použijeme toto značení, dostáváme vztah pro střední hodnotu chyby špatné klasifikace klasifikačním pravidlem $\mathbb{A}$

$$R(\mathbb{A}) = \sum_{i=1}^{k} \int_{A_i} q_i(\mathbf{x}) \, \mathrm{d}\mathbf{x}.$$

Věta 4.3. *(Bayesovské klasifikační pravidlo). Nechť $\mathbb{A}^* = \{A_1^*, A_2^*, \ldots, A_k^*\} \in \mathcal{A}$ je takové klasifikační pravidlo, že objekt s pozorovanou hodnotou znaku $\mathbf{X} = \mathbf{x}$ zařadíme j-té třídy, pokud*

$$q_j(\mathbf{x}) \leq q_i(\mathbf{x}), \quad i = 1, \ldots, k.$$

Pak $\mathbb{A}^*$ je optimální klasifikační pravidlo a platí

$$R(\mathbb{A}) \geq R(\mathbb{A}^*), \quad \forall \mathbb{A} \in \mathcal{A}.$$

Důkaz.

$$\begin{aligned} R(\mathbb{A}) = \sum_{i=1}^{k} \int_{A_i} q_i(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \sum_{i=1}^{k} \sum_{j=1}^{k} \int_{A_i \cap A_j^*} q_i(\mathbf{x}) \, \mathrm{d}\mathbf{x} \geq \\ \geq \sum_{i=1}^{k} \sum_{j=1}^{k} \int_{A_i \cap A_j^*} q_j(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \sum_{j=1}^{k} \int_{A_j^*} q_j(\mathbf{x}) \, \mathrm{d}\mathbf{x} = R(\mathbb{A}^*) \end{aligned}$$

□

Uvedená věta tak dává optimální klasifikační pravidlo. Pokud existuje t splňující

$$q_t(\mathbf{x}) < q_i(\mathbf{x}), \quad i = 1, \ldots, k, \; i \neq t, \tag{4.4}$$

objekt zařadíme do třídy A_t. Z nerovnosti (4.4) také plyne

$$p_t f_t(\mathbf{x}) > p_i f_i(\mathbf{x}), \quad i = 1, \ldots, k, \ i \neq t. \tag{4.5}$$

V případě, že ve vztahu (4.4) platí rovnost i pro další i, nezáleží na tom, který z těchto indexů vybereme.

Dále pokud předpokládáme, že f_i je hustota pravděpodobnosti n-rozměrného normálního rozdělení se známou střední hodnotou $\boldsymbol{\mu}_i$ a známou regulární varianční maticí $\mathbf{V}_i$, $i = 1, \ldots, k$, lze dojít k následujícím výsledkům. Hustota f_i má tvar

$$f_i(\mathbf{x}) = \frac{1}{\sqrt{2\pi|\mathbf{V}_i|}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^\top \mathbf{V}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) \right\},$$

po úpravě (4.5) dostaneme

$$\ln p_t + \ln f_t(\mathbf{x}) > \ln p_i + \ln f_i(\mathbf{x}), \quad i = 1, \ldots, k, \ i \neq t. \tag{4.6}$$

Jestliže navíc označíme

$$D_i = -\frac{1}{2}|\mathbf{V}_i| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^\top \mathbf{V}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) \ln p_i,$$

platí (4.6) právě tehdy, když

$$D_t > D_i, \quad i = 1, \ldots, k, \ i \neq t.$$

Lze tedy postupovat tak, že pro daný nezařazený objekt spočteme hodnoty D_i, $i = 1, \ldots, k$, objekt pak náleží do té tříty, pro kterou je tato hodnota maximální. Popsaná metoda je známá jako kvadratická diskriminační analýza (D_i je kvadratickou funkcí $\mathbf{x}$).

Jesliže jsou si všechny varianční matice rovny, tedy $\mathbf{V}_1 = \mathbf{V}_2 = \ldots = \mathbf{V}_k = \mathbf{V}$, označme

$$d_i = \boldsymbol{\mu}_i^\top \mathbf{V}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^\top \mathbf{V}^{-1}\boldsymbol{\mu}_i + \ln p_i. \tag{4.7}$$

Protože platí

$$D_i = d_i - \frac{1}{2}\ln|\mathbf{V}| - \frac{1}{2}\mathbf{x}^\top \mathbf{V}^{-1}\mathbf{x},$$

dostáváme pro tento případ ekvivalenci

$$D_t > D_i \Longleftrightarrow d_t > d_i, \quad i = 1, \ldots, k, \ i \neq t.$$

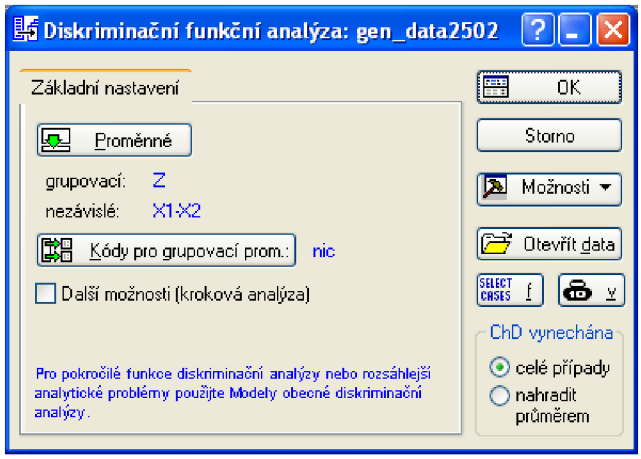
d_i je lineární funkcí $\mathbf{x}$, této metodě tak říkáme lineární diskriminační analýza.

D_i a d_i, $i = 1, \ldots, k$, nazýváme klasifikačními funkcemi.

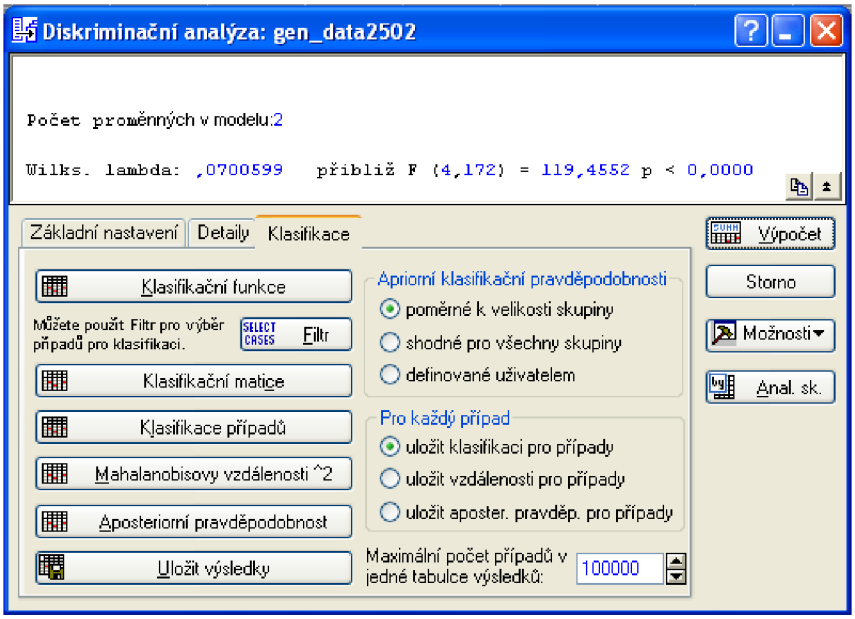
4.2 Výpočet klasifikačních funkcí v programu STATISTICA

Předpokládejme stejné zadání jako v podkapitole 3.3. Klasifikaci ovšem provedeme metodami diskriminační analýzy.

V programu STATISTICA z nabídky vícerozměrných metod vybereme *Dikriminační analýza*. Zvolíme prediktory a proměnnou, která označuje typy jednotlivých objektů, tzv. grupovací proměnnou, viz Obrázek 16. Obdržíme následující výsledky (Obrázek 17). Pokud vybereme *Klasifikační funkce*, dostáváme tabulku s koeficienty jednotlivých klasifikačních funkcí lineární diskriminační analýzy. Zde je označíme $\widehat{d}_i$, $i = 1, 2, 3$, abychom je odlišili od přesných (teoretických) tvarů těchto funkcí, viz dále.



Obrázek 16: Základní nastavení.



Obrázek 17: Výsledky.

Prediktor	$\widehat{d_1}$	$\widehat{d_2}$	$\widehat{d_3}$
X_1	$2,53062$	$-2,72518$	$-0,44768$
X_2	$-0,29828$	$0,22094$	$2,84691$
konstanta	$-3,56275$	$-3,96343$	$-5,46548$

Tedy

$$\widehat{d_1} = 2,53062X_1 - 0,29828X_2 - 3,56275,$$
$$\widehat{d_2} = -2,72518X_1 + 0,22094X_2 - 3,96343,$$
$$\widehat{d_3} = -0,44768X_1 + 2,84691X_2 - 5,46548.$$

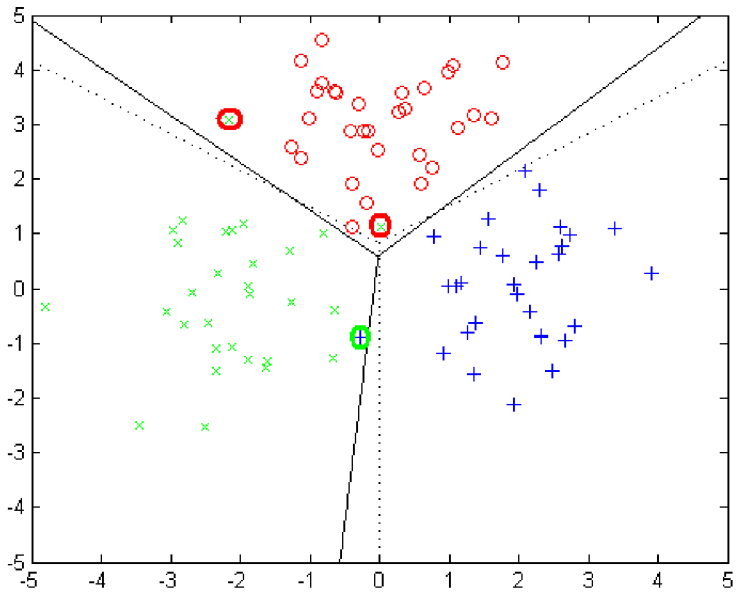
Protože se jedná pouze o ilustrativní příklad, příslušné vektory středních hodnot $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_3$ a varianční matici $\mathbf{V}$ známe. Jednotlivé klasifikační funkce pro kontrolu spočteme

dosazením do vztahu (4.7). Dostáváme

$$d_1 = 2X_1 - 3,0986,$$
$$d_2 = -2X_1 - 3,0986,$$
$$d_3 = 3X_2 - 5,5986.$$

Volba *Klasifikace případů* v nabídce na Obrázku 17 nám ukáže, která pozorování byla chybně klasifikována. Jedná se o pozorování P_{24}, P_{34} a P_{59}. Situaci ilustruje Obrázek 18, chybné zařazení uvedených tří pozorování je opět znázorněno pomocí barev. Plnými čarami jsou vyznačeny body (x_1, x_2), pro které $\widehat{d}_i = \widehat{d}_j$, přerušovanými čarami pak body, pro něž $d_i = d_j$, $i, j = 1, 2, 3$. Byla použita lineární diskriminační analýza, jedná se tedy v obou případech o polopřímky. Sestavíme také tabulku, kde porovnáme Z a $\widehat{Z}$:

	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$
$Z=1$	29	1	0
$Z=2$	0	28	2
$Z=3$	0	0	30



Obrázek 18: Vyhodnocení.

5 Klasifikační stromy

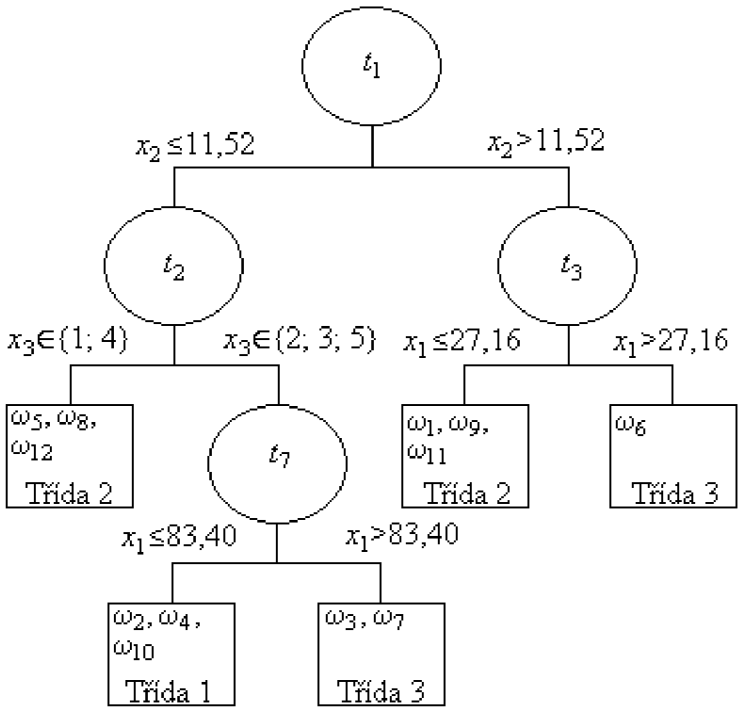
5.1 Ilustrativní příklad

Je dán učební soubor $\mathcal{S} = \{(\mathbf{x}_1^\top, z_1)^\top, (\mathbf{x}_2^\top, z_2)^\top, \ldots, (\mathbf{x}_{12}^\top, z_{12})^\top\}$, $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^3$, $z_i \in \{1, 2, 3\}$, $i = 1, 2, \ldots, 12$. Navíc $x_1 \in (0, 100)$, $x_2 \in \mathbb{R}$ a $x_3 \in \{1, 2, 3, 4, 5\}$. Objekty $\mathcal{S}$ označíme ω_i, $i = 1, 2, \ldots, 12$. Soubor $\mathcal{S}$ je určen tabulkou:

i	X_1	X_2	X_3	Z
1	0, 19	15, 60	1	2
2	12, 96	−8, 32	3	1
3	94, 14	−50, 66	5	3
4	70, 73	−0, 91	2	1
5	64, 73	1, 86	4	2
6	30, 45	101, 02	5	3
7	91, 65	−5, 69	2	3
8	98, 04	−1073, 10	1	2
9	23, 06	61, 17	4	2
10	2, 63	0, 02	5	1
11	15, 62	242, 28	3	2
12	7, 82	10, 24	1	2

Pomocí prediktorů X_1, X_2 a X_3 chceme nalézt klasifikační strom (viz dále), který každému objektu ω přiřadí hodnotu veličiny Z.

Výsledný klasifikační strom pro soubor $\mathcal{S}$ může mít například tvar, který je na Obrázku 19.



Obrázek 19: Klasifikační strom – ilustrativní příklad.

5.2 Základní pojmy

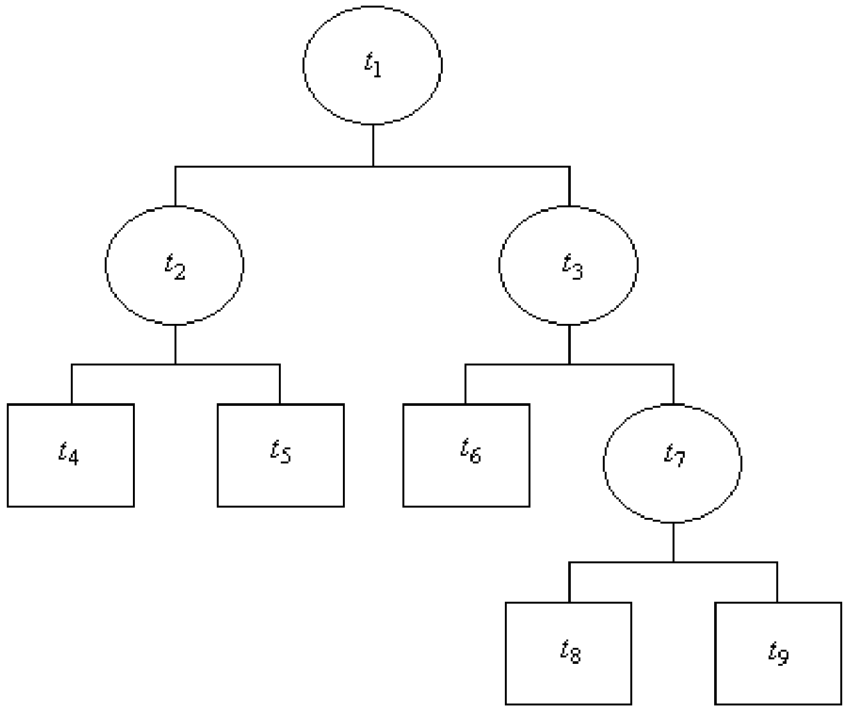
V této kapitole se budeme zabývat dalšími metodami pro hledání vhodného klasifikačního pravidla ve smyslu rozkladu množiny $\mathcal{X}$.

Definice 5.1. Řekneme, že konečná neprázdná množina $T \subset \mathbb{N}$ s dvojicí funkcí $left$, $right$ z T do $T \cup \{0\}$ je binárním stromem, jestliže jsou splněny následující podmínky

1. $\forall t \in T$ je buď $left(t) = right(t) = 0$ nebo $left(t) > t$ a $right(t) > t$,
2. $\forall t \in T$, pokud je t různé od nejmenšího čísla v T, existuje právě jedno $s \in T$, že buď $t = left(s)$ nebo $t = right(s)$.

Například v klasifikačním stromu na Obrázku 20, $t_4 = left(t_2)$ a $t_5 = right(t_2)$.

Uzlem stromu nazýváme každý prvek $t \in T$. Nejmenší prvek T se nazývá kořen stromu a značíme jej $root(T)$, na Obrázku 20, $t_1 = root(T)$. Prvky $t \in T$, pro něž platí, že $left(t) = right(t) = 0$, nazveme listy T, množinu všech listů stromu T označíme $\widetilde{T}$, na Obrázku 20 jsou listy stromu T vyznačeny hranatými obrysy.



Obrázek 20: Příklad klasifikačního stromu, strom T.

Definice 5.2. Nechť $t \in T$, $A(t)$ označíme množinu prvků $\mathcal{X}$, které náleží uzlu t, a třídou $j(t)$ rozumíme množinu prvků $\mathcal{C}$, kterou klasifikujeme daný uzel t.

Například v klasifikačním stromu z předchozí podkapitoly, viz Obrázek 19, je $A(t_2) = = \mathbb{R} \times (-\infty; 11,52\rangle \times \mathbb{R} \subset \mathbb{R}^3$. Pomocí uvedených pojmů můžeme nyní precizovat zavedení klasifikačního stromu.

Definice 5.3. Nechť $\widetilde{T}$ je množina všech listů stromu T. Každému $t \in T$ přiřadíme množinu $A(t) \subset \mathcal{X}$ a třídu $j(t) \subset \mathcal{C}$ takovým způsobem, aby

1. $\{A(t),\ t \in \widetilde{T}\}$ byl disjunktním rozkladem $\mathcal{X}$,

2. $\{A(s),\ s = left(t)$ nebo $s = right(t)\}$ byl disjunktní rozklad $A(t)$ pro všechna $t \notin \widetilde{T}$.

Dostáváme tak klasifikační pravidlo, které nazveme klasifikačním stromem. Pro všechna $\mathbf{x} \in \mathcal{X}$ a pro všechna $t \in \widetilde{T}$ nechť platí

$$\widehat{z} = j(t) \Longleftrightarrow \mathbf{x} \in A(t),$$

kde $\widehat{z}$ je odhadem z příslušné každému $\mathbf{x} \in \mathcal{X}$.

Použití tohoto pravidla můžeme ilustrovat na příkladu v předchozí podkapitole například pro objekt ω_1, $\mathbf{x}_1 = (1, 19; 15, 60; 1)$. Na začátku je objekt ω_1 v uzlu t_1, protože pro ω_1 je $x_2 > 11, 52$, zařadíme jej do uzlu t_3. Dále $x_1 < 27, 16$, ω_1 tak spadá do uzlu, pro který $j = 2$. Pro objekt ω_1 tak obdržíme odhad $\widehat{z} = 2$. Skutečnou hodnotu veličiny Z pro klasifikované objekty zpravidla neznáme.

Definice 5.4. Řekneme, že uzel s je rodič uzlu t (a opačně t je synem s), jestliže pro $t, s \in T$ platí, že $t = left(s)$ nebo $t = right(s)$. Pak *parent* je funkce z T do $T \cup \{0\}$, pro níž platí:

1. $parent(root(T)) = 0$,

2. $\forall t \in T$, $t \neq root(t)$ je rodičem t uzel $parent(t)$.

Pokud lze $s \in T$ zapsat ve tvaru $s = parent(t)$, nebo $s = parent(parent(t))$, případně $s = parent(\ldots(parent(t))\ldots)$. s nazveme předchůdcem t a t následovníkem s.

Řekneme, že T_1 je podstromem stromu T, jestliže trojice $T_1 \subset T$, $T_1 \neq \emptyset$ spolu s funkcemi $left_1$, $right_1$ z T_1 do $T_1 \cup 0$, definovanými

$$left_1(t) = \begin{cases} left(t) & \text{pro } left(t) \in T_1, \\ 0 & \text{jinak,} \end{cases}$$
$$right_1(t) = \begin{cases} right(t) & \text{pro } right(t) \in T_1, \\ 0 & \text{jinak.} \end{cases}$$

splňuje definici binárního stromu. Označme dále T_t, $t \in T$, podstrom stromu T, který má za kořen uzel t a dále obsahuje všechny jeho následovníky. T_t pak nazveme větví stromu T vycházející z uzlu t. Větvím $T_{left(t_1)} = T_L$ a $T_{right(t_1)} = T_R$ říkáme primární větve stromu T.

Příklad 5.5. Definované pojmy můžeme ilustrovat opět pomocí Obrázku 20.

- t_2 je rodičem t_4
- t_4 je synem t_2
- $t_1 = parent(t_3)$
- t_9 je následovníkem t_3, t_3 je předchůdcem t_9
- strom $T_1 = \{t_2, t_4, t_5\}$ je podstromem stromu T
- $T_{t_3} = \{t_3, t_6, t_7, t_8, t_9\}$

5.3 Konstrukce klasifikačních stromů metodou top-down

Při tomto postupu máme na začátku celý učební soubor a postupně jej dělíme. V této části se budeme zabývat především následujícími třemi otázkami. Jak zvolit štěpící pravidlo v každém uzlu, zda je možné daný uzel prohlásit za list a dělení ukončit a jakým způsobem přiřadit třídu $j(t)$ danému listu.

Vycházíme z učebního souboru $\mathcal{S}$ o N objektech. Nejprve se soustřeďme na výběr štěpícího pravidla. Zaveďme označení $p(t)$ pro pravděpodobnost, že daný objekt bude při klasifikaci zařazen do uzlu t, tedy $p(t) = P(\mathbf{X} \in A(t)),\ t \in T$.

Definice 5.6. Řekneme, že s je štěpení uzlu $t \in T$, jestliže přiřadí tomuto uzlu hodnoty $t_L = left(t)$ a $t_R = right(t)$ tak, že

$$p_L(t) = \frac{p(t_L)}{p(t)}, \qquad p_R(t) = \frac{p(t_R)}{p(t)}. \tag{5.8}$$

Pravděpodobnosti $p_L(t)$ a $p_R(t)$ odhadujeme:

$$\widehat{p}_L(t) = \frac{N(t_L)}{N(t)}, \qquad \widehat{p}_R(t) = \frac{N(t_R)}{N(t)},$$

$N(t),\ t \in T$, značí počet objektů z $\mathcal{S}$, pro které $\mathbf{x} \in A(t)$.

Nyní je třeba posoudit kvalitu takovéhoto štěpení, k tomuto účelu definujeme další pojmy.

Definice 5.7. Funkci φ definovanou na množině uspořádaných k-tic $(p_1, \ldots, p_k) \in \mathbb{R}^k$ nazveme funkcí nečistoty, pokud splňuje následující podmínky.

1. $p_i \geq 0,\ i = 1, \ldots, k$,
2. $\sum_{i=0}^{k} p_i = 1$,
3. φ nabývá maxima pouze v bodě $(\frac{1}{k}, \ldots, \frac{1}{k})$,
4. φ nabývá minima pouze v bodech $(1, 0, \ldots, 0), (0, 1, \ldots, 0), \ldots, (0, 0, \ldots, 1)$,
5. φ je ryze konkávní.

Příklad 5.8. Nechť $k = 2$. Pak můžeme zvolit následující funkci nečistoty,

$$\varphi(p_1, p_2) = p_1 p_2.$$

Protože dle definice $p_1 + p_2 = 1,\ p_i \geq 0,\ i = 1, 2$, dostáváme

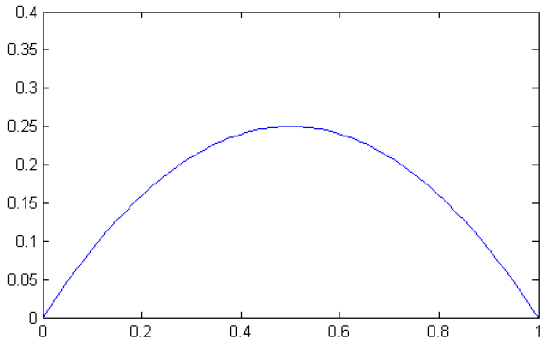
$$\varphi(p_1, p_2) = p_1 p_2 = p_1(1 - p_1) = p_1 - p_1^2,$$

graf si můžeme prohlédnout na Obrázku 21.

Označme $p(i|t),\ t \in T,\ i = 1, \ldots, k$ pravděpodobnost, že objekt, který byl zařazen do třídy $A(t)$, je typu i.

Definice 5.9. Řekneme, že $i(t)$ je míra nečistoty uzlu $t \in T$, jestliže se jedná o funkci nečistoty φ definovanou na množině uspořádaných k-tic $(p(1|t), \ldots, p(k|t))$, tedy

$$i(t) = \varphi\big(p(1|t), \ldots, p(k|t)\big).$$



Obrázek 21: Funkce nečistoty $\varphi(p_1, p_2) = p_1 p_2 = p_1(1 - p_1) = p_1 - p_1^2$.

Kvalitu štěpení s uzlu t tak můžeme posuzovat pomocí poklesu nečistoty $\triangle i(s,t)$. Ten lze definovat takto:

$$\triangle i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R),$$

kde p_L a p_R mají stejný význam jako v Definici 5.6. Pak platí

Věta 5.10. *Jestliže s je libovolné štěpení, pak $\triangle i(s,t) \geq 0$, speciálně*

$$\triangle i(s,t) = 0 \iff p(i|t_L) = p(i|t_R) = p(i|t), \forall i = 1, \ldots, k.$$

Důkaz.

Viz [7].

□

Při výběru rozkladu množiny $A(t)$ se přirozeně snažíme volit to štěpení, které má maximální pokles nečistoty $\triangle i(s,t)$. Množinu všech štěpení nejčastěji představuje množina rozkladů uzlu $A(t)$ v závislosti na hodnotách prediktorů. Kvalitu výsledného klasifikačního stromu můžeme hodnotit pomocí míry nečistoty jeho listů.

Definice 5.11. Nechť T je klasifikační strom a nechť

$$I(T) = \sum_{t \in \widetilde{T}} i(t) p(t),$$

pak $I(T)$ nazveme nečistotou stromu T.

Nyní se zaměřme na otázku, jak každému listu $t \in \widetilde{T}$ přiřadit $j(t)$, $j(t) \in \{1, \ldots, k\}$. Každé volbě $j(t)$ přiřadíme chybu klasifikace $r(t)$, pro kterou nechť platí

$$r(t) = \sum_{j \neq j(t)} p(j|t) = 1 - p(j(t)|t).$$

Lze tedy postupovat tak, že z množiny $\{1, \ldots, k\}$ vybereme takové $j(t)$, pro které je chyba klasifikace $r(t)$ minimální. Využít můžeme také ekvivalentní podmínku. Danému listu přiřadíme takové j, pro které

$$p(j|t) = \max_i p(i|t), \ i = 1, \ldots, k.$$

V případě, že je splněna podmínka pro více takovýchto j, zvolíme libovolně jedno z nich. Celková chyba klasifikace stromem T má pak tvar,

$$R(T) = \sum_{t \in \widetilde{T}} r(t)p(t).$$

Popsaný postup platí pro případ, kdy je chyba špatné klasifikace stejná pro všechna $i \neq j$.

Zbývá rozhodnout, kdy ukončit štěpení a prohlásit uzel za list. Tento problém má při konstrukci klasifikačního stromu velký význam. Formulujeme počáteční pravidlo

$$\max_{s \in S} \triangle i(s,t) < \beta, \tag{5.9}$$

kde S představuje množinu všech štěpení a β zvolené číslo. Pokud je nerovnost splněna, uzel prohlásíme za list a štěpení ukončíme. Toto pravidlo se ovšem ukazuje jako nevyhovující. Není totiž znám postup, jak najít uspokojivou dolní hranici poklesu nečistoty β. Navíc hodnota výrazu (5.9) nedává žádnou informaci o poklesu míry nečistoty v dalších krocích. Hledání optimálního pravidla pro ukončení štěpení tak představuje značný nedostatek, který však odstraňuje následující metoda.

5.4 Konstrukce klasifikačních stromů metodou growing-pruning

Zde budeme postupovat tak, že nejprve zkonstruujeme rozsáhlý strom T_{max} postupným štěpením uzlů, dokud jednotlivé listy nebudou obsahovat pouze předem daný počet prvků nebo prvky pouze jedné třídy. Pak pomocí prořezávání (pruning) sestavíme posloupnost podstromů stromu T_{max}. Nakonec pomocí odhadu $R(T)$ vybereme nejvhodnější.

Definice 5.12. Řekneme, že podstrom T_1 stromu T se nazývá prořezaný podstrom T, jestliže $root(T_1) = root(T)$. Volíme označení $T_1 \preceq T$. Prořezáním stromu pak rozumíme postup, který ze stromu T vytvoří prořezaný podstrom $T_1 \preceq T$ odříznutím některých větví (odříznutím větve T_t máme na mysli, že jsou odstraněni všichni následovníci t, t se tak stává listem). Strom, jenž vznikne z T odřezáním větve T_t značíme $T - T_t$.

Začněme tedy konstrukcí stromu T_{max}. Opět vyjdeme z učebního souboru $\mathcal{S}$. Množina všech štěpení S je v tomto případě generována standardizovanou množinou otázek.

Definice 5.13. Nechť má učební soubor $\mathcal{S}$ standardní strukturu. $\mathcal{Q}$ nazveme standardizovanou množinou otázek, pokud jsou splněny následující podmínky.

1. Každé štěpení $s \in S$ závisí na hodnotě pouze jednoho prediktoru.
2. $\mathcal{Q}$ obsahuje všechny otázky typu $\{\text{Je } x_i < c?\},\ c \in \mathbb{R}$ pro každou spojitou náhodnou veličinu X_i.
3. $\mathcal{Q}$ obsahuje všechny otázky typu $\{\text{Je } x_i \in H?\}, H \subset \{h_1, \ldots, h_m\}$ pro každou nominální náhodnou veličinu X_i, která nabývá pouze hodnot z $\{h_1, \ldots, h_m\}$.

Připomeňme, že štěpení s uzlu t posuzujeme podle poklesu nečistoty

$$\triangle i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R).$$

Existuje řada variant pro volbu nečistoty $i(t)$ uzlu t, třeba pomocí Giniho indexu, kdy

$$i(t) = \sum_{i \neq j} p(i|t)p(j|t) = 1 - \sum_{i=1}^{k} p^2(i|t),$$

nebo například

$$i(t) = -\sum_{i=1}^{k} p(i|t) \log \big[p(i|t)\big].$$

Připomeňme, že pravděpodobnosti p_L a p_R jsou dány vztahy (5.8). Nejlepším štěpením pak nazveme to, které maximalizuje pokles nečistoty. Nalezneme tak nejlepší štěpení $s_i^*(t)$ podle jednotlivých prediktorů $x_i,\ i = 1, \ldots, n$, pak vybereme to $s_{max}^*(t)$ to, pro které

$$\triangle i(s_i^*(t), t) = \max_i \triangle i(s_i(t), t).$$

Postupným prováděním těchto štěpení dostaneme T_{max}.

Zaměřme se nyní na prořezání stromu T_{max}. Každému prořezanému stromu $T \preceq T_{max}$ přiřadíme jeho složitost $|\widetilde{T}|$ (počet jeho listů).

Definice 5.14. Pro nezáporné reálné číslo α zavedeme penalizovanou míru celkové chyby klasifikace stromem:

$$R_\alpha(T) = R(T) + \alpha|\widetilde{T}|.$$

Parametr α nazveme parametrem složitosti.

Pro každou hodnotu α najdeme podstrom $T(\alpha) \preceq T_{max}$, pro který je penalizovaná míra $R_\alpha(T)$.

Definice 5.15. Řekneme, že $T(\alpha) \preceq T_{max}$ je minimální optimálně prořezaný strom T_{max}, jestliže pro jeho parametr složitosti α splňuje následující podmínky.

1. $R_\alpha(T(\alpha)) = \min_{T(\alpha) \preceq T_{max}} R_\alpha(T)$,
2. $R_\alpha(T) = R_\alpha(T(\alpha)) \Longrightarrow T(\alpha) \preceq T$.

Věta 5.16. *Nechť pro daný parametr složitosti α existuje pro každý strom T právě jeden podstrom $T(\alpha) \preceq T$. Dále předpokládejme, že T je netriviální, $root(T) = t_1$ a $T_L = T_{left(t_1)},\ T_R = T_{right(t_1)}$ značí primární větve T. Pak*

$$R_\alpha(T(\alpha)) = \min\big\{R_\alpha(t_1), R_\alpha(T_L(\alpha)) + R_\alpha(T_R(\alpha))\big\},$$

navíc

$$R_\alpha(t_1) \le R_\alpha(T_L(\alpha)) + R_\alpha(T_R(\alpha)) \Longrightarrow T(\alpha) = \{t_1\},$$
$$R_\alpha(t_1) > R_\alpha(T_L(\alpha)) + R_\alpha(T_R(\alpha)) \Longrightarrow T(\alpha) = \{t_1\} \cup T_L(\alpha) \cup T_R(\alpha).$$

Důkaz.

Viz [3].

□

Uvedená věta ukazuje, že minimální optimálně prořezaný strom je svou definicí jednoznačně určen. Rostoucí α zvyšuje penále za složitost stromu, což vede ke stejnému nebo menšímu optimálně prořezanémnu podstromu. Platí následující věta.

Věta 5.17. *Nechť α_1, α_2 jsou parametry složitosti, potom*

$$\alpha_2 \ge \alpha_1 \Longrightarrow T(\alpha_2) \preceq T(\alpha_1).$$

Pokud je navíc $\alpha_2 > \alpha_1$ *a* $T(\alpha_2) \prec T(\alpha_1)$, *platí*

$$\alpha_1 < \frac{R(T(\alpha_2)) - R(T(\alpha_1))}{|\widetilde{T}(\alpha_1)| - |\widetilde{T}(\alpha_2)|} \leq \alpha_2.$$

Důkaz.

Viz [3].

□

Prořezávání začneme podstromem $T_1 = T(0) \preceq T_{max}$, pro který platí $R(T_1) = R(T_{max})$. T_1 získáme následujícím způsobem. Předpokládejme uzly t_L a t_R, které jsou listy T_{max}, vzniklé štěpením uzlu t. Platí

$$R(t) \geq R(t_L) + R(t_R).$$

Pokud nastane rovnost, odřízneme uzly t_L a t_R a t prohlásíme za list. Postup opakujeme, dokud je to možné, dostaneme $T_1 \preceq T_{max}$. Pro každý uzel $t \in T_1 - \widetilde{T}_1$ platí, že $R(t) > R(T_t)$.

Označme $\{t\}$ podvětev T_t obsahující pouze uzel t. Jestliže pro příslušné penalizované míry platí

$$R_\alpha(T_t) < R_\alpha(\{t\}), \tag{5.10}$$

ponecháme větev T_t součástí stromu T_1. Dále hledáme takovou hodnotu parametru složitosti α, pro kterou nastane rovnost. Řešíme tak nerovnici (5.10). Uvedené penalizované míry mají tvar

$$R_\alpha(\{t\}) = R(t) + \alpha,$$
$$R_\alpha(T_t) = R(T_t) + \alpha|\widetilde{T}_t|.$$

Dostáváme

$$\alpha < \frac{R(t) - R(T_t)}{|\widetilde{T}_t| - 1}.$$

Označíme $g_1(t)$ funkci, která každému $t \in T_1$ přiřadí

$$g_1(t) = \begin{cases} \infty & \text{pro } t \in \widetilde{T}_1, \\ \frac{R(t) - R(T_t)}{|\widetilde{T}_t| - 1} & \text{jinak.} \end{cases}$$

Dále nalezneme $\bar{t}_1 \in T_1$, pro které

$$g_1(\bar{t}_1) = \min_{t \in T_1} g_1(t),$$

pak položíme $\alpha_2 = g_1(\bar{t}_1)$. V uzlu $\bar{t}_1$ už je výhodnější použít ke klasifikaci pouze $\{\bar{t}_1\}$ namísto větve $T_{\bar{t}_1}$. Odříznutím větve $T_{\bar{t}_1}$ dostáváme podstrom $T_2 \preceq T_1$,

$$T_2 = T_1 - T_{\bar{t}_1}.$$

Postup opakujeme. Jestliže v i-tém kroku dojde k situaci, že minima $g_i(t)$ je dosaženo ve více různých uzlech, odřízneme všechny jim odpovídající větve. Obdržíme tak vnořenou posloupnost podstromů

$$T_1 \succ T_2 \succ \ldots \succ \{t_1\}$$

a odpovídající posloupnost parametrů složitosti

$$\alpha_1 = 0 < \alpha_i < \alpha_{i+1},\ i \geq 1.$$

Věta 5.18. *Nechť* $\{\alpha_i\}$ *je posloupnost získaná uvedeným postupem. Pak pro* $i \geq 1$ *a* $\alpha_i < \alpha < \alpha_{i+1}$ *platí*

$$T(\alpha) = T(\alpha_i) = T_i.$$

Poslední částí konstrukce klasifikačního stromu metodou growing-pruning je výběr nejlepšího prořezaného podstromu. Postupovat budeme tak, že určíme odhad celkové chyby klasifikace stromem $R(T)$, označíme jej $\widehat{R}(T)$. Vybereme pak ten podstrom, pro který bude tento odhad minimální.

$Q(i,j)$ označíme pravděpodobnost, že objekt třídy j je chybně klasifikován do třídy i, tedy

$$Q(i,j) = P(\widehat{Z} = i | Z = j).$$

Připomeňme, že $R(j)$ značí chybu špatné klasifikace objektu třídy j, platí

$$R(j) = \sum_{i=1}^{k} C(i,j)Q(i,j),$$

kde $C(i,j)$ je chyba špatné klasifikace objektu třídy j třídou i.

Nyní uvedeme dvě používané metody pro stanovení odhadu chyby klasifikace stromem $R(T)$.

Testový odhad $\widehat{R}_{ts}(T)$

Z učebního souboru $\mathcal{S}$ náhodně vybereme N' pozorování, nazveme je testovým souborem $\mathcal{S}_2$, zbývající pozorování budou tvořit učební soubor $\mathcal{S}_1$. Strom T_{max} a posloupnost jeho prořezaných podstromů $T_1 \succ T_2 \succ \ldots \succ \{t_1\}$ vytvoříme na základě $\mathcal{S}_1$. Objekty z testového souboru $\mathcal{S}_2$ oklasifikujeme stromy $T_1, T_2, \ldots, \{t_1\}$. N'_j označíme počet objektů $\mathcal{S}_2$, které patří do j-té třídy. Dále nechť pro libovolný prořezaný podstrom T N'_{ij} představuje počet objektů z $\mathcal{S}_2$ j-té třídy, které tento strom zařadil do třídy i-té. Můžeme tak sestavit následující odhady,

$$\widehat{Q}_{ts}(i,j) = \begin{cases} \frac{N'_{ij}}{N'_j} & \text{pro} N'_j \neq 0, \\ 0 & \text{pro} N'_j = 0, \end{cases}$$

$$\widehat{R}_{ts}(j) = \sum_{i=1}^{k} C(i,j)\widehat{Q}_{ts}(i,j),$$

$$\widehat{R}_{ts}(T) = \sum_{j=1}^{k} \widehat{R}_{ts}(j)p_j.$$

Přičemž pravděpodobnosti p_j můžeme odhadnout pomocí objektů testového souboru $\mathcal{S}_2$. Jestliže tedy zvolíme

$$p_j = \frac{N'_j}{N'},$$

pak

$$\widehat{R}_{ts}(T) = \frac{1}{N'} \sum_{i,j=1}^{k} C(i,j) N'_{ij}.$$

Odhad $\widehat{R}_{ts}(T)$ tak představuje chyby klasifikace všech objektů testového souboru stromem T.

Odhad křížovým ověřováním $\widehat{R}_{cv}(T)$

Nechť V značí násobnost křížového ověřování. Učební soubor $\mathcal{S}$ je rozdělen na V přibližně stejně velkých podsouborů $\mathcal{S}_v,\ v = 1, \ldots, V$. Zavedeme V nových učebních souborů $\mathcal{S}^{(v)} = \mathcal{S} - \mathcal{S}_v,\ v = 1, \ldots, V$ takových, že každý z nich obsahuje $\frac{V-1}{V} N$ objektů. Pomocí $\mathcal{S}$ a všech $\mathcal{S}^{(v)}$ zkonstruujeme T_{max} a příslušné $T^{(v)}_{max}$. Předpokládejme, že pro každý parametr složitosti α, představují $T(\alpha)$ a $T^{(v)}(\alpha),\ v = 1, \ldots, V$, odpovídající nejmenší optimálně prořezané podstromy. Podsoubor $\mathcal{S}_v$ lze použít jako testovací soubor pro strom $T^{(v)}(\alpha)$.

Nyní pro každé $v = 1, \ldots, V$ oklasifikujeme $\mathcal{S}_v$ stromem $T^{(v)}_{max}$. Označíme pro pevně daná v, i, j $N^{(v)}_{ij}$ počet objektů z $\mathcal{S}_v$ j-té třídy, které byly $T^{(v)}_{max}$ chybně klasifikovány jako prvky i-té třídy. Nechť

$$N_{ij} = \sum_{v=1}^{V} N^{(v)}_{ij}.$$

Tato metoda vychází z toho, že pro velká v by měly mít stromy $T(\alpha)$ a $T^{(v)}(\alpha)$ stejnou klasifikační přesnost a chybu. Proto můžeme odhadnout pravděpodobnost, že $T(\alpha)$ zařadí objekt j-té třídy do i-té následujícím způsobem,

$$\widehat{Q}_{cv}(i,j) = \frac{N_{ij}}{N_j}.$$

Nyní můžeme určit i ostatní odhady, postupně dostáváme

$$\widehat{R}_{cv}(j) = \sum_{i=1}^{k} C(i,j) \widehat{Q}_{cv}(i,j),$$

$$\widehat{R}_{cv}(T(\alpha)) = \sum_{j=1}^{k} \widehat{R}_{cv}(j) p_j.$$

Jestliže pravděpodobnosti $p_j,\ j = 1, 2, \ldots, k$, odhadneme podobně jako v předchozím případě, tedy

$$p_j = \frac{N_j}{N},$$

obdržíme po úpravě

$$\widehat{R}_{cv}(T(\alpha)) = \frac{1}{N} \sum_{i,j=1}^{k} C(i,j) N_{ij}.$$

Věta 5.18 nám umožňuje položit

$$\alpha'_i = \sqrt{\alpha_k \alpha_{k+1}}$$

a

$$\widehat{R}_{cv}(T_i) = \widehat{R}_{cv}(T(\alpha'_i)).$$

Pomocí uvedených odhadů vybereme z posloupnosti $T_1, T_2, \ldots, \{t\}$ podstrom T_{i_0},který je dán jako

$$\widehat{R}(T_{i_0}) = \min_i \widehat{R}(T_i).$$

Odhad $R(T)$ křížovým ověřováním je výpočtově náročnější, bývá tak využíván pro soubory menšího rozsahu.

Odhady $\widehat{R}(T_i)$ jsou funkcemi počtu listů $|\widetilde{T}_i|$. Pozice $\min_i \widehat{R}(T_i)$ je velmi nestabilní, malé změny rozložení dat nebo parametru složitosti α způsobí při křížovém ověřování velké změny $|\widetilde{T}_i|$. Proto je třeba pravidlo pro výběr nejlepšího prořezaného podstromu ještě upravit tak, abychom zmenšili uvedenou nestabilitu a vybrali co nejjednodušší podstrom ve smyslu $|\widetilde{T}_i|$, jehož přesnost zůstane srovnatelná s $\min_i \widehat{R}(T_i)$.

Definice 5.19. Nechť T_{i_0} je strom splňující

$$\widehat{R}(T_{i_0}) = \min_i \widehat{R}(T_i)$$

a SE směrodatná odchylka odhadu $\widehat{R}(T_i)$. Jestliže pravidlo pro výběr nejlepšího prořezaného podstromu vybere podstrom T_{i_1}, pro který

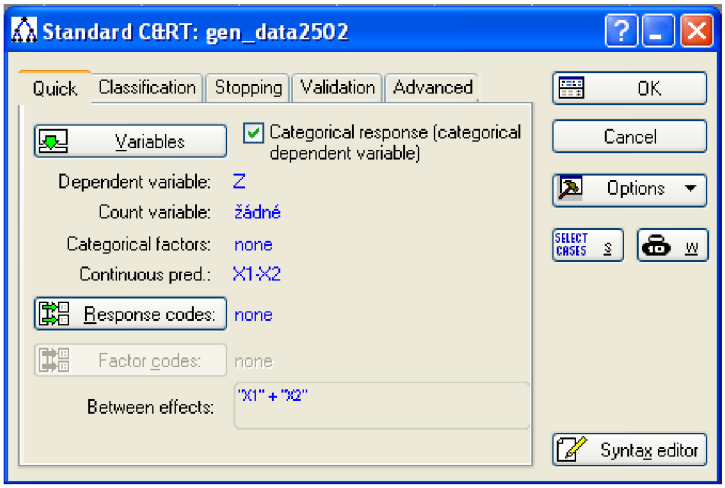
$$\widehat{R}(T_{i_1}) \leq \widehat{R}(T_{i_0}) + SE\big(\widehat{R}(T_{i_0})\big),$$

nazveme jej 1 SE pravidlem.

5.5 Výpočet klasifikačního stromu v programu STATISTICA

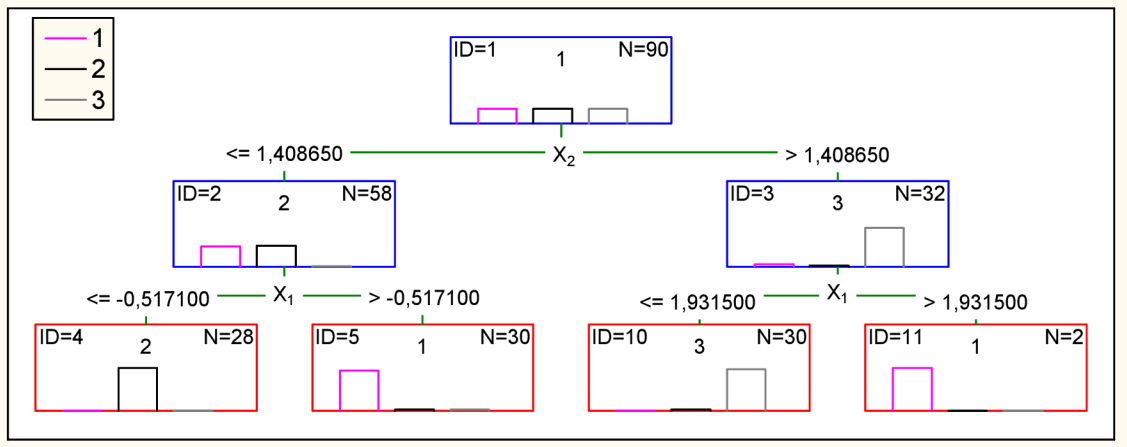
Zabývejme se nyní stejnou úlohou jako v podkapitole 3.3. Zde ji budeme řešit pomocí klasifikačního stromu konstruovaného v počítačovém programu.

V programu STATISTICA vybereme *General Classification/Regression Tree Models*. Pokud dále zvolíme *CART*, objeví se před námi nabídka pro volbu parametrů konstrukce klasifikačního stromu, viz Obrázek 22.



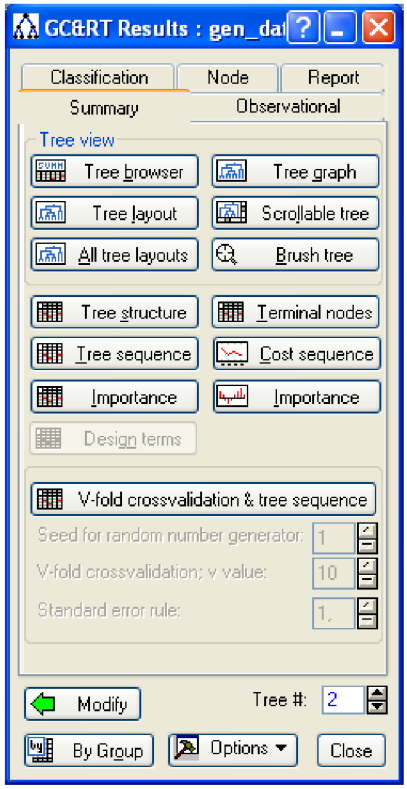
Obrázek 22: Úvodní nabídka.

Zde volíme prediktory a také skupiny, do kterých jednotlivá pozorování patří (*Dependent variable*). V záložce *Classification* určíme chyby špatných klasifikací, zadáme *Equal* (tedy že chyby jsou stejné pro každé pozorování), dále tvar míry nečistoty uzlu – *Gini measure* (podle Giniho indexu). Na tomto místě ještě zbývá zadat apriorní pravděpodobnosti, vybereme *Estimated*, tedy odhadnuté. V záložce *Stopping* zvolíme *Prune of misclassification error* – prořezání na základě chyby špatné klasifikace klasifikačním pravidlem. V další záložce, *Validation*, zadáme, zda si přejeme odhad $\widehat{R}(T)$ stanovit pomocí křížového ověřování *V-fold cross-validation* nebo pomocí testového souboru *Test sample*, volíme křížové ověřování včetně jeho parametrů.

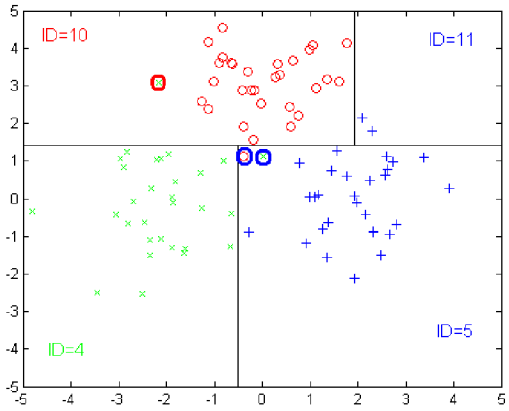


Obrázek 23: Klasifikační strom.

Výsledky obdržíme v podobě na Obrázku 24. Klasifikační strom si můžeme prohléd-



Obrázek 24: Výsledky.



Obrázek 25: Vyhodnocení.

nout pomocí volby *Tree graph*, viz Obrázek 23.

Čistotu jednotlivých listů můžeme blíže posoudit pomocí volby *Terminal nodes*, v záložce *Nodes* si můžeme nechat vypsat pozorování, která padla do jednotlivých uzlů. Zjistíme tak, že špatně byla klasifikována pozorování P_{34}, P_{59} a P_{87}, situaci ilustruje Obrázek 25 (čarami jsou oddělena pozorování příslušející do jednotlivých listů stromu, popis koresponduje v Obrázkem 23).

Hodnoty Z a $\widehat{Z}$ můžeme opět porovnat v tabulce:

	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$
$Z=1$	30	0	0
$Z=2$	1	28	1
$Z=3$	1	0	29

6 Aplikace klasifikačních metod na reálná data

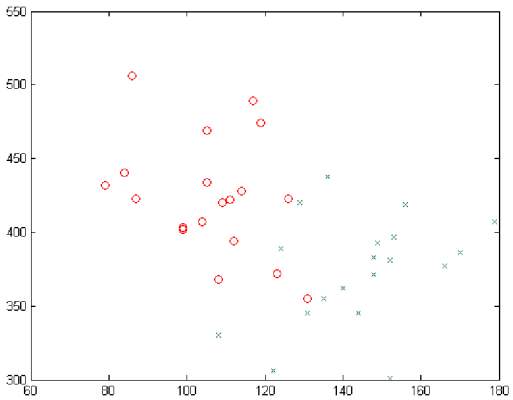
6.1 Srovnání metod na výběru malého rozsahu

Je dán datový soubor rozsahu $N = 70$, který byl převzat z [5] (viz příloha P1 – soubory data_sal.sta, data_sal.xls). Byly pozorovány průměry kroužků na šupinách lososů po prvním roce ve sladké a v mořské vodě. Každé pozorování tak představuje charakteristiky jednoho lososa. V souboru jsou zastoupeny dvě skupiny lososů, polovina jich pochází z Aljašky a polovina z Kanady. Místo původu lososa lze rozlišit podle velikosti zmíněných kroužků na jeho šupinách. K dispozici tedy máme následující prediktory:

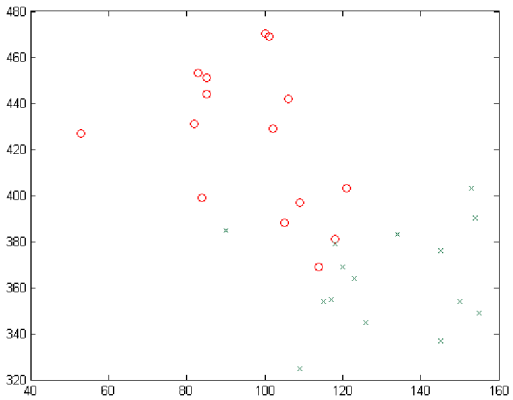
X_1 ... průměr kroužků po prvním roce ve sladké vodě,
X_2 ... průměr kroužků po prvním roce v mořské vodě.

Hodnoty jsou udány v setinách palců. Veličina Z značí místo původu lososa, volíme $Z = 1$, pokud pochází z Aljašky a $Z = 2$, pokud je z Kanady.

Soubor byl rozdělen na dva podsoubory. Na učební soubor $\mathcal{S}_1 = \{\omega_1, \omega_2, \ldots, \omega_{40}\}$ o rozsahu $N_1 = 40$, který obsahuje 20 pozorování, pro něž $Z = 1$ a rovněž 20 pozorování, pro která $Z = 2$, dále na $\mathcal{S}_2$ o rozsahu $N_2 = 30$, také $\mathcal{S}_2 = \{\omega'_1, \omega'_2, \ldots, \omega'_{30}\}$ obsahuje stejný počet pozorování obou typů.



Obrázek 26: Učební soubor $\mathcal{S}_1$.



Obrázek 27: Soubor $\mathcal{S}_2$.

Úkolem je sestavit pomocí učebního souboru $\mathcal{S}_1$ klasifikační pravidla, klasifikovat pozorování ze souboru $\mathcal{S}_2$ a porovnat výsledky.

6.1.1 Klasifikace pomocí shlukové analýzy

Budeme postupovat tak, že nejprve provedeme shlukování učebního souboru. Ukončíme jej $(N_1 - 2)$-hým rozkladem, který obsahuje právě dva shluky – $\mathbb{A}_{38} = \{A_1, A_2\}$. Pro každé pozorování $\omega' \in \mathcal{S}_2$ spočteme jeho vzdálenost od shluků A_1 a A_2. Hodnotu prediktoru X_i, $i = 1, 2$, příslušejícího objektu ω budeme značit $X_i(\omega)$. Vzdálenost objektu ω' od shluku $A = \{\omega_1, \omega_2, \ldots, \omega_m\}$, kterou označíme $\delta_1(A, \omega')$, definujeme takto:

$$\delta_1(A, \omega') = \sqrt{\left(\overline{X}_1(A) - X_1(\omega')\right)^2 + \left(\overline{X}_2(A) - X_2(\omega')\right)^2}, \tag{6.11}$$

kde

$$\overline{X}_i(A) = \frac{1}{m}\sum_{j=1}^{m} X_i(\omega_j), \quad i = 1, 2,$$

jsou aritmetické průměry veličin X_1 a X_2 pro objekty ω patřící do shluku A.

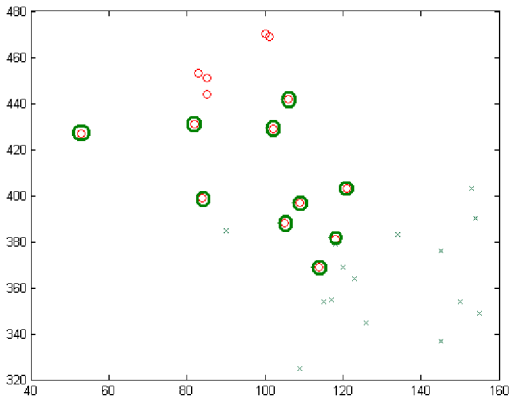
Jedná se tedy o eukleidovskou vzdálenost objektu ω' od „středu“ shluku A. Pro každé klasifikované pozorování určíme shluk, ke kterému má toto pozorování nejblíže (jeho vzdálenost k tomuto shluku je minimální).

Pro shluky A_1 a A_2 určíme počty jejich objektů, pro které $Z = 1$ a $Z = 2$. $n_j(A_i)$ označíme počet objektů shluku A_i, pro něž $Z = j$. Odhad veličiny Z přiřadíme každému objektu $\omega' \in \mathcal{S}_2$ podle následujícího pravidla. Pokud má objekt ω' minimální vzdálenost ke shluku A_i a $n_1(A_i) > n_2(A_i)$, pak $\widehat{Z}(\omega') = 1$, v případě, že $n_1(A_i) < n_2(A_i)$, pak $\widehat{Z}(\omega') = 2$, jestliže nastane rovnost, je lhostejné, jaký odhad zvolíme.

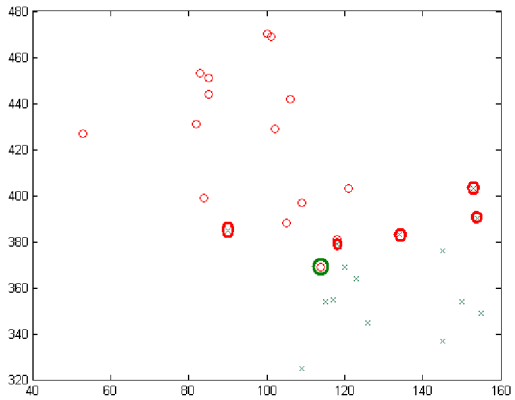
Nyní můžeme klasifikaci soubor $\mathcal{S}_2$ uvedeným postupem, shlukování učebního souboru $\mathcal{S}_1$ provedeme metodou nejbližšího souseda. Pro jednotlivá pozorování ω' obdržíme $\widehat{Z}$, porovnáme je se Z v tabulce:

	$\widehat{Z} = 1$	$\widehat{Z} = 2$
$Z = 1$	5	10
$Z = 2$	0	15

Správně bylo klasifikováno 20 případů, což činí přibližně $66,7\%$. Výsledky ilustruje Obrázek 28, špatně klasifikovaná pozorování jsou vyznačena barevně. Postup zopakujeme,



Obrázek 28: Metoda nejbližšího souseda, vzdálenost δ_1.



Obrázek 29: Metoda nejvzdálenějšího souseda, vzdálenost δ_1.

ale učební soubor budeme nyní shlukovat metodou nejvzdálenějšího souseda. Dostáváme tyto výsledky:

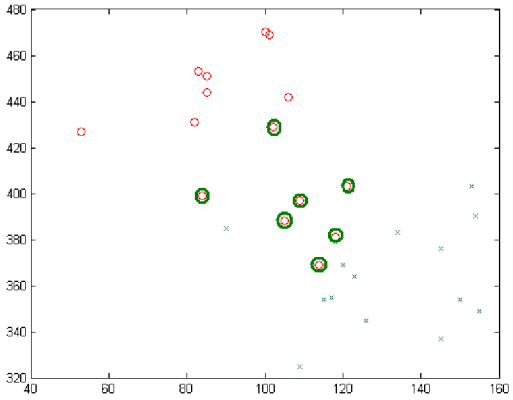
	$\widehat{Z} = 1$	$\widehat{Z} = 2$
$Z = 1$	14	1
$Z = 2$	5	10

Správně bylo klasifikováno 24 případů, což činí 80%, viz Obrázek 29.

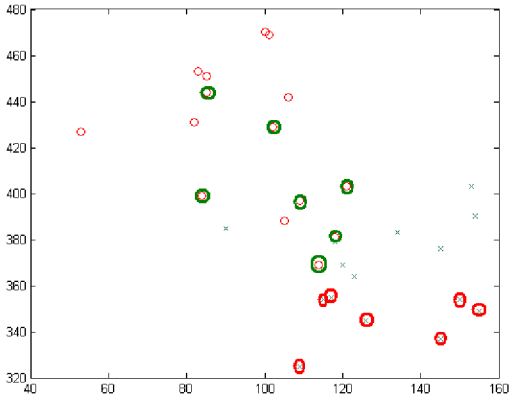
Klasifikaci provedeme ještě jednou, učební soubor $\mathcal{S}_2$ budeme tentokrát shlukovat metodou průměrné nepodobnosti, održíme

	$\widehat{Z}=1$	$\widehat{Z}=2$
$Z=1$	8	7
$Z=2$	0	15

Správně bylo klasifikováno 23 případů, což činí přibližně $76,7\%$, viz Obrázek 30.



Obrázek 30: Metoda průměrné nepodobnosti, vzdálenost δ_1.



Obrázek 31: Metoda nejbližšího souseda, vzdálenost δ_2.

Vidíme, že nejlépe dopadla klasifikace, kdy jsme volili shlukování učebního souboru metodou nejvzdálenějšího souseda. V ostatních případech byla špatná klasifikace jednostranná, pozorováním se $Z=1$ byla přiřazena $\widehat{Z}=2$.

Vyzkoušejme nyní změnit způsob, jak přiřadit shluku $A=\{\omega_1,\omega_2,\ldots,\omega_k\}$ a objektu ω' jejich vzdálenost, označme ji tentokrát $\delta_2(A,\omega')$. Můžeme ji spočítat jako průměr eukleidovských vzdáleností objektu ω' s jednotlivými objekty shluku A, tedy

$$\delta_2(A,\omega')=\frac{1}{m}\sum_{i=1}^{m}\sqrt{\big(X_1(\omega_i)-X_1(\omega')\big)^2+\big(X_2(\omega_i)-X_2(\omega')\big)^2}. \tag{6.12}$$

Opět provedeme klasifikaci pomocí shlukové analýzy, budeme měnit volbu metody shlukování učebního souboru a jako vzdálenost objektu od shluku použijeme právě definovanou vzdálenost δ_2.

Pro shlukování učebního souboru $\mathcal{S}_1$ metodou nejbližšího souseda nyní dostáváme

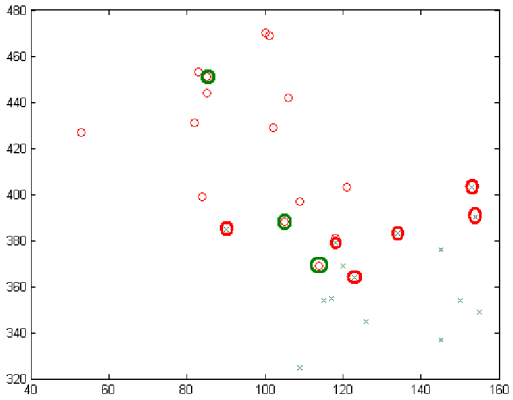
	$\widehat{Z}=1$	$\widehat{Z}=2$
$Z=1$	8	7
$Z=2$	7	8

Správně bylo klasifikováno 16 případů, což činí přibližně $53,3\%$, viz Obrázek 31.

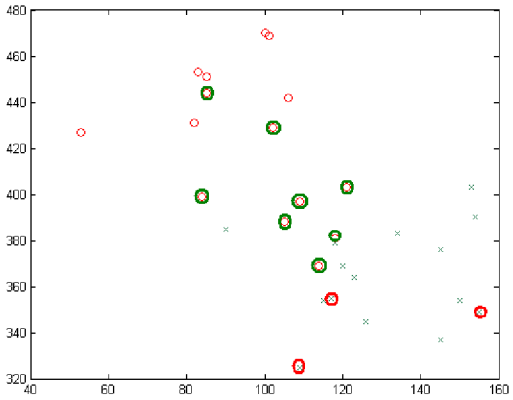
Shlukování $\mathcal{S}_1$ metodou nejvzdálenějšího souseda vede k výsledku

	$\widehat{Z}=1$	$\widehat{Z}=2$
$Z=1$	12	3
$Z=2$	6	9

Správně bylo klasifikováno 21 případů, což činí 70%, viz Obrázek 32. A konečně při shlu-



Obrázek 32: Metoda nejvzdálenějšího souseda, vzdálenost δ_2.



Obrázek 33: Metoda průměrné nepodobnosti, vzdálenost δ_2.

kování učebního souboru metodou průměrné nepodobnosti obdržíme

	$\widehat{Z}=1$	$\widehat{Z}=2$
$Z=1$	7	8
$Z=2$	3	12

Správně bylo klasifikováno 19 případů, což činí přibližně $63,3\%$, viz Obrázek 33.

Náročnější výpočet vzdálenosti objektu od shluku δ_2 nevede k lepším výsledkům. I v tomto případě dopadla nejlépe klasifikace, kdy byl učební soubor shlukován metodou nejvzdálenějšího souseda.

6.1.2 Klasifikace diskriminační analýzou

Zde využijeme funkci *classify* systému MATLAB. Tato funkce pomocí učebního souboru klasifikuje další objekty podle jejich prediktorů metodami diskriminační analýzy. Zadáme tedy učební soubor $\mathcal{S}_1$ a prediktory příslušné jednotlivým objektům klasifikovaného souboru $\mathcal{S}_2$. Výstupem funkce *classify* jsou odhady $\widehat{Z}$ pro daná pozorování z $\mathcal{S}_2$. Volbou dalšího parametru určíme, zda použijeme lineární nebo kvadratickou diskriminaci. Nejprve zvolíme lineární, obdržíme tyto výsledky:

	$\widehat{Z}=1$	$\widehat{Z}=2$
$Z=1$	15	0
$Z=2$	6	9

Správně bylo klasifikováno 24 objektů, což činí 80%.

Použitá funkce systému MATLAB nenabízí mezi výsledky klasifikační funkce d_1 a d_2. Danému objektu ω' je přiřazen odhad $\widehat{Z} = j$, pokud $d_j > d_i$, $i, j = 1, 2$. Této nerovnosti je ekvivalentní nerovnost

$$(X_1(\omega'), X_2(\omega'))\mathbf{L} + K > 0. \tag{6.13}$$

Vektor $\mathbf{L}$ a konstantu K lze vyčíslit. Nerovnosti $d_1 > d_2$ tak přísluší

$$\mathbf{L} = \begin{pmatrix} -0,1455 \\ 0,0301 \end{pmatrix}, \qquad K = 6,4047.$$

Pro nerovnost $d_2 > d_1$ odbdržíme

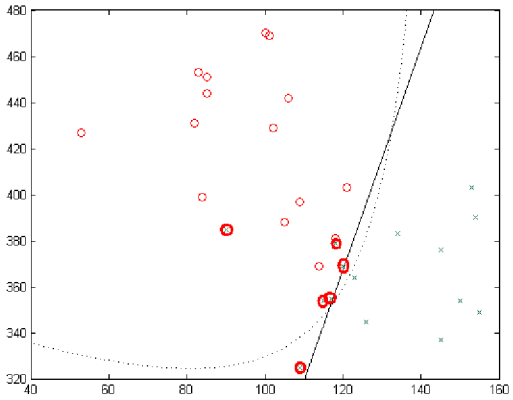
$$\mathbf{L} = \begin{pmatrix} 0,1455 \\ -0,0301 \end{pmatrix}, \qquad K = -6,4047.$$

Pokud tyto vektory a koeficienty postupně dosadíme do vztahu (6.13) a nerovnost nahradíme rovností, dostaneme tutéž rovnici přímky, společně s výskedky klasifikace je znázorněna na Obrázku 34. Čárkovaně je nakreslena křivka pro případ kvadratické diskriminace, viz dále.

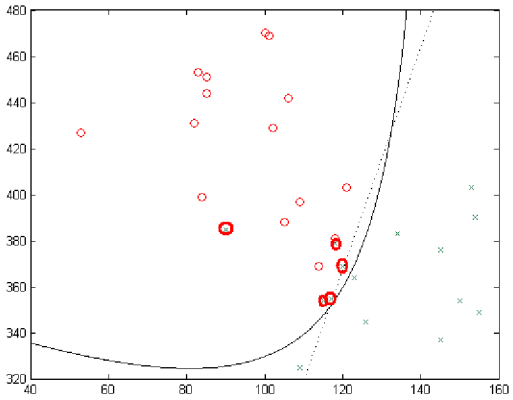
Nyní klasifikujme soubor $\mathcal{S}_2$ kvadratickou diskriminací. Dostáváme

	$\widehat{Z} = 1$	$\widehat{Z} = 2$
$Z = 1$	15	0
$Z = 2$	5	10

Správně bylo klasifikováno 25 objektů, což činí přibližně $83,3\%$.



Obrázek 34: Výsledky, lineární diskriminace.



Obrázek 35: Výsledky, kvadratická diskriminace.

V tomto případě je klasifikovanému objektu přiřazen odhad $\widehat{Z} = j$, pokud $D_j > > D_i$, $i, j = 1, 2$. Ekvivalentní nerovnost je

$$(X_1(\omega'), X_2(\omega'))^\top \mathbf{Q}\,(X_1(\omega'), X_2(\omega')) + (X_1(\omega'), X_2(\omega'))\mathbf{L} + K > 0. \tag{6.14}$$

Pro případ $D_1 > D_2$ dostáváme

$$\mathbf{Q} = \begin{pmatrix} -0,0008787 & -0,0005890 \\ -0,0005890 & -0,0000118 \end{pmatrix}, \quad \mathbf{L} = \begin{pmatrix} 0,5243 \\ 0,1876 \end{pmatrix}, \quad K = -65,3701.$$

A pro $D_2 > D_1$

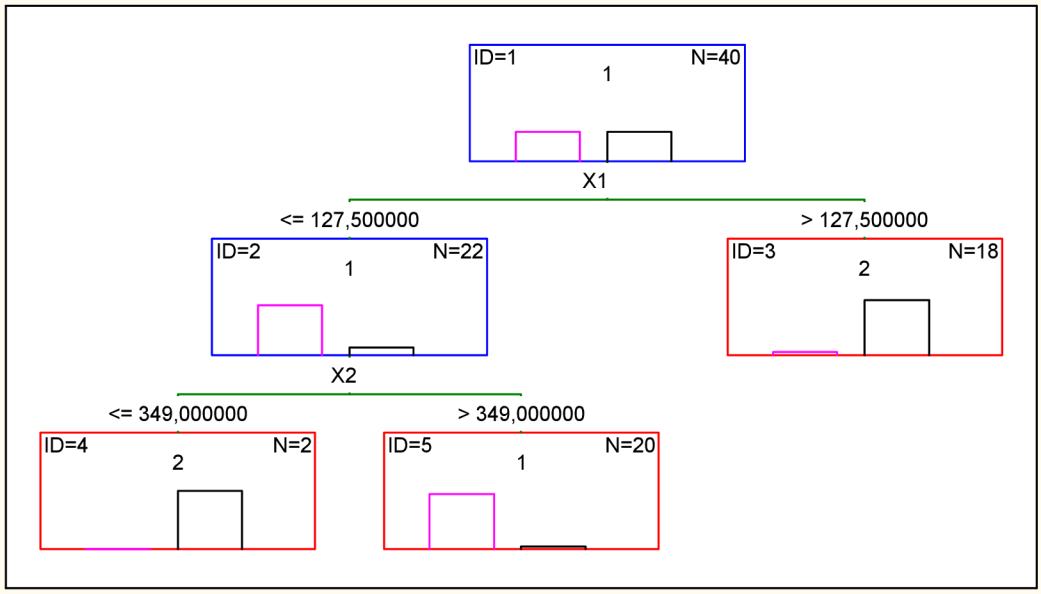
$$\mathbf{Q} = \begin{pmatrix} 0,0008787 & 0,0005890 \\ 0,0005890 & 0,0000118 \end{pmatrix}, \quad \mathbf{L} = \begin{pmatrix} -0,5243 \\ -0,1876 \end{pmatrix}, \quad K = 65,3701.$$

Dosazením vypočtených paramerů $\mathbf{Q}$, $\mathbf{L}$ a K do (6.14), pokud navíc opět nahradíme nerovnost rovností, získáme rovnici paraboly. Ta je zobrazena spolu s výsledky klasifikace pomocí diskriminační analýzy na Obrázku 35, čárkovaně je vykreslena přímka pro případ lineární diskriminace.

Při porovnání výsledků vidíme, že lepší dává klasifikace kvadratickou diskriminační analýzou, rozdíl je ovšem minimální. Při malém rozsahu výběru dat pro výpočet, může být tento rozdíl způsoben konkrétní volbou datových souborů $\mathcal{S}_1$ a $\mathcal{S}_2$.

6.1.3 Klasifikace klasifikačním stromem

Klasifikační strom zkonsruujeme v programu STATISTICA identickým postupem jako v podkapitole 5.5, jeho tvar je na Obrázku 36. Pomocí něj přiřadíme pozorováním ze



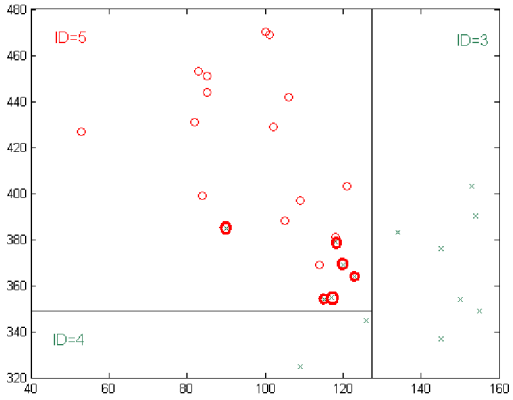
Obrázek 36: Klasifikační strom.

souboru $\mathcal{S}_2$ odhady $\widehat{Z}$. Porovnání Z a $\widehat{Z}$ pro jednotlivé klasifikované objetky byla opět uspořádána do tabulky.

	$\widehat{Z}=1$	$\widehat{Z}=2$
$Z=1$	15	0
$Z=2$	6	9

Správně bylo klasifikováno 24 objektů, což činí 80%. Výsledkům této klasifikace odpovídá Obrázek 37, je také vyznačeno, do kterého uzlu byla daná pozorování zařazena.

Porovnejme nyní výsledky pro použité klasifikační metody. Nejvyšší úspěšnosti bylo docíleno při klasifikaci kvadratickou diskriminační analýzou, přibližně 83, 3%. Podobného



Obrázek 37: Výsledky, klasifikační strom.

poměru dobře zařazených objektů bylo dosaženo také klasifikací klasifikačním stromem, lineární diskriminační analýzou, ale také při vhodné volbě parametrů pro klasifikaci za užití shlukové analýzy. Na této volbě záleží, protože pokud bylo zvoleno shlukování učebního souboru metodou nejbližšího souseda a navíc vzdálenost objektu a shluku δ_2, klesla úspěšnost klasifikace až na přibližně $53,3\%$.

6.2 Aplikace metod na reálný datový soubor většího rozsahu

Zabývejme se nyní úlohou, kdy je úkolem předpovídat znečištění ovzduší na základě předpovědi počasí.

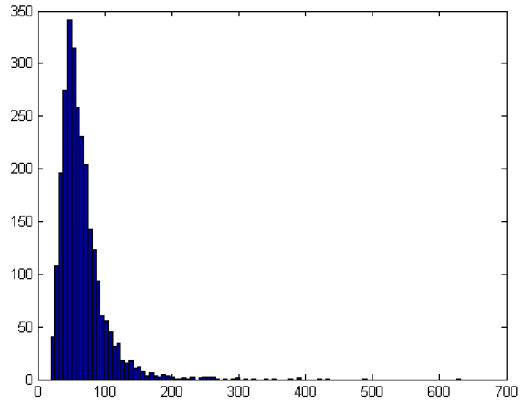
Prachový aerosol (směs malých pevných nebo kapalných částeček v plynu) obsahuje shluky molekul, ledové krystalky, drobný hmyz, atd. Nejčastěji jsou analyzovány částečky menší než 10 μm (označovány PM_{10}), 5 μm (PM_5) a 2,5 μm ($\mathrm{PM}_{2,5}$). Hodnota znečištění PM_{10} má jednotku $[\mu\mathrm{g}\cdot\mathrm{m}^{-3}]$, je to tedy hmotnost částeček menších než 10 μm v metru krychlovém vzduchu.

Je dán soubor $\mathcal{S}$ rozsahu $N = 2703$, který vznikl v monitorovací stanici Zvonařka v Brně (viz příloha P1 – soubory data_poc.sta, data_poc.xls). Měření probíhala v jednotlivých dnech od ledna 1998 do prosince 2005. Pro i-tý den, $i = 1, 2, \ldots, 2703$, máme k dispozici hodnoty těchto veličin:

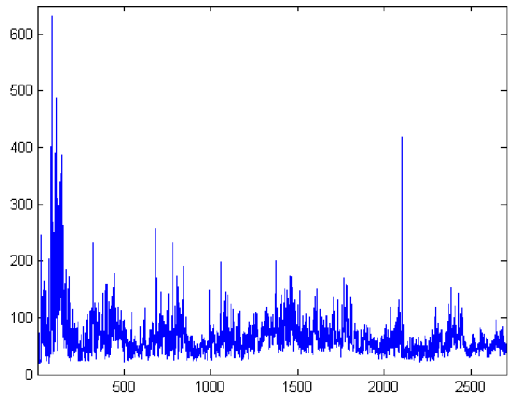
X_{1i} ... teplota vzduchu [°C],
X_{2i} ... rychlost větru $[\mathrm{m}\cdot\mathrm{s}^{-1}]$,
X_{3i} ... směr větru [°],
X_{4i} ... relativní vlhkost vzduchu [%],
X_{5i} ... absolutní vlhkost vzduchu $[\mathrm{g}\cdot\mathrm{m}^{-3}]$,
$\mathcal{Z}_i$ ... hodnota znečištění ovzduší PM_{10} $[\mu\mathrm{g}\cdot\mathrm{m}^{-3}]$.

Pomocí hodnot $X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}$ a $\mathcal{Z}_{i-1}$ chceme predikovat hodnotu $\mathcal{Z}_i$. Situaci lze interpretovat tak, že pomocí meteorologické předpovědi na příští den (hodnoty $X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}$) a dnešního znečištění – $\mathcal{Z}_{i-1}$ predikujeme zítřejší hodnotu znečištění $\mathcal{Z}_i$. Protože hodnota $\mathcal{Z}_{i-1}$ má také význam prediktoru, označíme ji pro zjednodušení dalšího zápisu X_{6i}.

Na Obrázku 38 je histogram veličiny $\mathcal{Z}$ (pro 100 tříd). Základní charakteristiky $\mathcal{Z}$ jsou $\widetilde{\mathcal{Z}} = 58,0450$ a $\overline{\mathcal{Z}} = 67,2070$. Na Obrázku 39 jsou vykresleny hodnoty $\mathcal{Z}$ v čase.



Obrázek 38: Histogram pro $\mathcal{Z}$.



Obrázek 39: Průběh $\mathcal{Z}$ v čase.

Nejprve objekty souboru $\mathcal{S}$ rozdělíme s přihlédnutím k histogramu do k skupin podle hodnot veličiny $\mathcal{Z}$. Příslušnost do jedné z těchto skupin budeme charakterizovat nově zavedenou veličinou Z (která bude nabývat hodnot z množiny $\{1, 2, \ldots, k\}$). Volíme $k = 5$. Pro každý objekt $\omega_i \in \mathcal{S}$ tedy nechť

$$\begin{aligned}
Z_i = 1 &\iff \mathcal{Z}_i \leq 35,79,\\
Z_i = 2 &\iff 35,79 < \mathcal{Z}_i \leq 45,07,\\
Z_i = 3 &\iff 45,07 < \mathcal{Z}_i \leq 77,11,\\
Z_i = 4 &\iff 77,11 < \mathcal{Z}_i \leq 103,89,\\
Z_i = 5 &\iff 103,89 < \mathcal{Z}_i.
\end{aligned}$$

Hodnoty 35, 79; 45, 07; 77, 11 a 103, 89 jsou postupně dolní decil, dolní kvartil, horní kvartil a horní decil veličiny $\mathcal{Z}$.

Soubor $\mathcal{S}$ náhodně rozdělíme na dva podsoubory, na soubor $\mathcal{S}_1$ o rozsahu N_1, jeho objekty označíme opět ω a $\mathcal{S}_2$ o rozsahu N_2, objekty označíme ω', přičemž $N_1 \approx 2N_2$. $\mathcal{S}_1$ bude představovat učební soubor, pomocí něj sestavíme klasifikační pravidla. Těmi budeme klasifikovat objekty $\omega' \in \mathcal{S}_2$ tak, že jim přiřadíme odhad $\widehat{Z}$ a tím i přibližnou hodnotu znečištění PM_{10}. Protože známe hodnoty $\mathcal{Z}$ pro tyto objekty, budeme moci úspěšnost jednotlivých klasifikačních metod porovnat.

Poznámka. Indexy i, kterými byla značena příslušnost veličin $X_1, \ldots, X_6$ a $\mathcal{Z}, Z$ k jednotlivým dnům, budeme dále, pokud to bude možné, vynechávat. Soubor $\mathcal{S}$ byl rozdělen náhodně, v souborech $\mathcal{S}_1$ a $\mathcal{S}_2$ tak na pořadí objektů ω, resp. ω', nezáleží. Hodnoty uvedených veličin příslušných jednotlivým objektům budeme značit jako v předchozí podkapitole. Například $X_2(\omega)$ je hodnota veličiny X_2 pro objekt ω.

6.2.1 Klasifikace pomocí metod shlukové analýzy

Postupovat budeme podobně jako v podkapitole 6.1.1. Provedeme shlukování souboru $\mathcal{S}_1$, které ukončíme tak, aby měl výsledný rozklad $\mathbb{A}$ předem daný počet shluků, volíme 7.

Každému z těchto shluků přiřadíme jednu hodnotu $\overline{\mathcal{Z}}(A)$ znečištění PM_{10}. Shluku $A = \{\omega_1, \omega_2, \ldots, \omega_m\}$ ji přiřadíme takto,

$$\overline{\mathcal{Z}}(A) = \exp\left(\frac{1}{m}\sum_{i=1}^{m} \ln \mathcal{Z}(\omega_i)\right).$$

Hodnoty znečištění $\mathcal{Z}$ tedy nejprve zlogaritmujeme, spočteme z nich aritmetický průměr a ten pak odlogaritmujeme. Činíme tak proto, abychom snížili vliv velkých rozdílů jednotlivých hodnot $\mathcal{Z}$ na $\overline{\mathcal{Z}}$.

Pro každý objekt $\omega' \in \mathcal{S}_2$ spočteme jeho vzdálenost s jednotlivými shluky rozkladu $\mathbb{A}$. Vztah (6.11) pro výpočet vzdálenosti $\delta_1(A, \omega')$ objektu ω' od shluku $A = \{\omega_1, \omega_2, \ldots, \omega_m\}$ upravíme pro případ, kdy máme šest prediktorů, na tvar

$$\delta_1(A, \omega') = \left[\sum_{j=1}^{6} \left(\frac{\overline{X}_j(A) - X_j(\omega')}{s_j}\right)^2\right]^{\frac{1}{2}}, \tag{6.15}$$

kde s_j je výběrová směrodatná odchylka hodnot $X_j(\omega_i)$, $i = 1, 2, \ldots, N_1$ (tedy pro všechna $\omega \in \mathcal{S}_1$). Rozdíl $\overline{X}_j(A) - X_j(\omega')$ dělíme výběrovou směrodatnou odchylkou s_j, protože veličiny $X_1, \ldots, X_6$ mají různé jednotky. s_j je určena vztahy

$$s_j = \left[\frac{1}{N_1 - 1}\sum_{i=1}^{N_1} \left(X_j(\omega_i) - \overline{X}_j\right)^2\right]^{\frac{1}{2}}, \quad \overline{X}_j = \frac{1}{N_1}\sum_{i=1}^{N_1} X_j(\omega_i), \quad j = 1, 2, \ldots, 6. \tag{6.16}$$

Dále určíme, pro který shluk je tato vzdálenost minimální. Podle toho mu přiřadíme odhad $\widehat{Z}$. Pokud má tedy objekt ω' nejblíže ke shluku A, $\widehat{Z}(\omega')$ mu přiřadíme podle tohoto pravidla,

$$\begin{aligned}
\widehat{Z}(\omega') = 1 &\iff \overline{\mathcal{Z}}(A) \leq 35,79,\\
\widehat{Z}(\omega') = 2 &\iff 35,79 < \overline{\mathcal{Z}}(A) \leq 45,07,\\
\widehat{Z}(\omega') = 3 &\iff 45,07 < \overline{\mathcal{Z}}(A) \leq 77,11,\\
\widehat{Z}(\omega') = 4 &\iff 77,11 < \overline{\mathcal{Z}}(A) \leq 103,89,\\
\widehat{Z}(\omega') = 5 &\iff 103,89 < \overline{\mathcal{Z}}(A).
\end{aligned}$$

Teď již můžeme přistoupit ke klasifikaci souboru $\mathcal{S}_2$. Shlukování souboru $\mathcal{S}_1$ provedeme metodou nejbližšího souseda. Výsledky klasifikační procedury, odhady $\widehat{Z}$, porovnáme s hodnotami Z pro jednotlivé objekty ω' v tabulce.

	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$	$\widehat{Z}=4$	$\widehat{Z}=5$
$Z=1$	0	0	77	0	0
$Z=2$	0	0	112	0	0
$Z=3$	0	0	408	0	2
$Z=4$	0	0	114	0	0
$Z=5$	0	0	65	0	12

Z celkového počtu 790 klasifikovaných objektů jich bylo správně zařazeno 420, což činí přibližně $53,2\%$. Všimněme si, že například $\widehat{Z} \neq 4$ pro žádný objekt $\omega' \in \mathcal{S}_2$. Toto je jistě

nežádoucí, má to svůj původ už ve shlukování souboru $\mathcal{S}_1$, žádnému ze shluků rozkladu $\mathbb{A}$ totiž nebyla přiřazená hodnota $\overline{\mathcal{Z}} \in (77,11;103,89\rangle$. Uvedený jev, který se bude opakovat i v dalších výpočtech v této podkapitole. Této situaci lze předejít použitím vyššího počtu shluků při klasifikaci. Nyní provedeme shlukování (metodou nejbližšího souseda) učebního souboru $\mathcal{S}_1$, které ukončíme tak, aby měl výsledný rozklad 50 shluků. Dostaneme tyto výsledky:

	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$	$\widehat{Z}=4$	$\widehat{Z}=5$
$Z=1$	4	8	67	1	1
$Z=2$	3	9	92	3	2
$Z=3$	33	8	281	55	41
$Z=4$	4	1	68	18	19
$Z=5$	1	0	23	13	42

Zde bylo klasifikováno 797 objektů, z toho správně 354, což činí přibližně $44,4\%$. Použití vyššího počtu shluků nevede k lepšímu výsledku, v dalších výpočtech tak budeme provádět klasifikace pomocí rozkladu souboru $\mathcal{S}_1$, který bude obsahovat opět 7 shluků.

Pokud postup zopakujeme s tím, že $\mathcal{S}_1$ budeme shlukovat metodou nejvzdálenějšího souseda, obdržíme následující tabulku:

	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$	$\widehat{Z}=4$	$\widehat{Z}=5$
$Z=1$	0	0	74	0	0
$Z=2$	0	0	125	2	0
$Z=3$	0	0	422	16	2
$Z=4$	0	0	86	33	1
$Z=5$	0	0	35	18	10

Klasifikováno bylo 824 objektů, správně 465, přibližně $56,4\%$.

Nyní zvolme zvolme pro shlukování souboru $\mathcal{S}_1$ metodu průměrné nepodobnosti. Dostáváme tuto tabulku:

	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$	$\widehat{Z}=4$	$\widehat{Z}=5$
$Z=1$	0	0	74	2	0
$Z=2$	0	0	108	4	8
$Z=3$	0	0	358	22	23
$Z=4$	0	0	101	6	19
$Z=5$	0	0	38	17	18

Klasifikováno bylo 798 objektů, správně 382, což činí přibližně $47,9\%$.

Stejně jako v podkapitole 6.1.1 klasifikaci provedeme i za použití jiného způsobu pro výpočet vzdálenosti objektu ω' od shluku $A = \{\omega_1, \omega_2, \ldots, \omega_m\}$, upravený vztah (6.12) pro $\delta_2(A, \omega')$ má nyní, kdy pracujeme s více prediktory, tvar

$$\delta_2(A,\omega') = \frac{1}{m}\sum_{i=1}^{m}\left[\sum_{j=1}^{6}\left(\frac{X_j(\omega_i)-X_j(\omega')}{s_j}\right)^2\right]^{\frac{1}{2}}, \tag{6.17}$$

kde s_j je opět dána vztahy (6.16).

Pro shlukování souboru $\mathcal{S}_1$ metodou nejbližšího souseda teď obdržíme tabulku:

	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$	$\widehat{Z}=4$	$\widehat{Z}=5$
$Z=1$	0	0	33	0	44
$Z=2$	0	0	66	0	57
$Z=3$	0	0	230	0	175
$Z=4$	0	0	59	0	66
$Z=5$	0	0	28	0	61

Bylo klasifikováno 819 objektů, z toho správně 291, což činí přibližně $35,5\%$.

Pokud budeme $\mathcal{S}_1$ shlukovat metodou nejvzdálenějšího souseda, dostaneme

	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$	$\widehat{Z}=4$	$\widehat{Z}=5$
$Z=1$	0	0	74	0	20
$Z=2$	0	0	81	1	21
$Z=3$	0	0	323	15	86
$Z=4$	0	0	73	10	42
$Z=5$	0	0	38	17	29

Klasifikováno bylo 830 objektů, správně zařazeno bylo 362, což činí přibližně $43,6\%$.

Shlukování souboru $\mathcal{S}_1$ metodou průměrné nepodobnosti vede k výsledku:

	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$	$\widehat{Z}=4$	$\widehat{Z}=5$
$Z=1$	0	0	46	1	37
$Z=2$	0	0	73	0	64
$Z=3$	0	0	201	8	190
$Z=4$	0	0	50	12	47
$Z=5$	0	0	32	8	49

Zde bylo klasifikováno 818 objektů, z toho správně 262, což činí přibližně $32,0\%$.

Vidíme tedy, že počítání vzdálenosti objektu od shluku podle vztahu (6.17) vede k horším výsledkům. Různý způsob výpočtu vzdálenosti objektu od shluku ovlivňuje úspěšnost metody více než použité shlukování učebního souboru $\mathcal{S}_1$. Nejlepšího výsledku bylo dosaženo použitím shlukování $\mathcal{S}_1$ metodou nejvzdálenějšího souseda a výpočtem vzdálenosti objektu od shluku podle vztahu (6.15) – $56,4\%$.

6.2.2 Klasifikace diskriminační analýzou

I v této podkapitole budeme postupovat velmi podobně jako v 6.1.2. Použijeme funkci MATLABu *classify*, vstupem jsou prediktory z učebního soubor $\mathcal{S}_1$, vektor veličin Z příslušných každému $\omega \in \mathcal{S}_1$ a prediktory z $\mathcal{S}_2$. Výstupem je pak vektor odhadů $\widehat{Z}$ příslušných každému $\omega' \in \mathcal{S}_2$. Pokud zvolíme lineární diskriminaci, dostaneme porovnáním veličin Z a $\widehat{Z}$ pro objekty souboru $\mathcal{S}_2$ tuto tabulku:

	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$	$\widehat{Z}=4$	$\widehat{Z}=5$
$Z=1$	32	39	5	5	4
$Z=2$	28	45	16	12	22
$Z=3$	64	71	68	73	79
$Z=4$	11	13	17	29	31
$Z=5$	21	9	3	24	37

Klasifikováno bylo 758 objektů, z toho správně 211, což činí přibližně $27,8\%$.

Klasifikace diskriminační analýzou vede k výsledku:

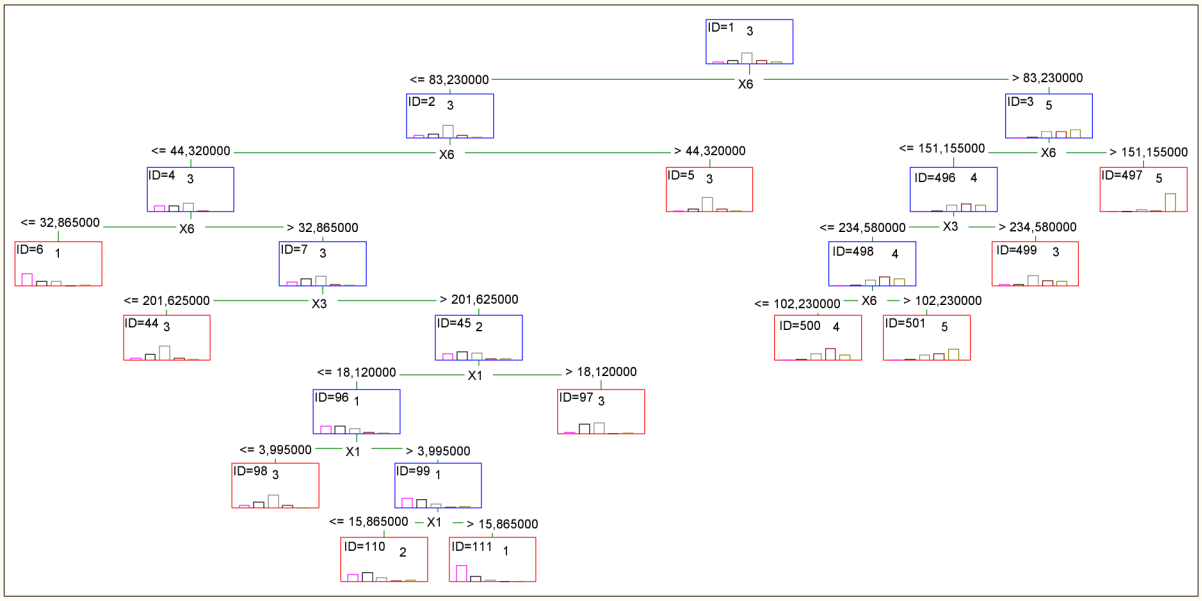
	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$	$\widehat{Z}=4$	$\widehat{Z}=5$
$Z=1$	62	13	3	0	0
$Z=2$	53	42	15	6	2
$Z=3$	79	134	115	73	16
$Z=4$	8	9	31	48	12
$Z=5$	6	4	13	22	39

Zde bylo klasifikováno 806 objektů, z toho správně 306, což činí přibližně $38,0\%$.

Vyšší procentuální úspěšnost má i v tomto případě klasifikace pomocí kvadratické diskriminace.

6.2.3 Klasifikace klasifikačním stromem

Vyjdeme z učebního souboru $\mathcal{S}_1$. Klasifikační strom zkonstruujeme v programu STATISTIKA stejným postupem jako v podkapitole 5.5. Připomeňme, že postupně volíme jednotlivé proměnné, chyby špatné klasifikace – *Equal*, tvar míry nečistoty uzlu – *Gini measure*, apriorní pravděpodobnosti – *Estimated*, způsob prořezávání – *Prune of misclassification error* a metodu stanovení $R(T)$ – *V-fold cross-validation*. Z nabídky výsledných stromů pak vybíráme ten, který je na Obrázku 40. Klasifikace jednotlivých objektů $\omega' \in \mathcal{S}_2$ tímto



Obrázek 40: Klasifikační strom.

stromem vedla k následujícím výsledkům:

	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$	$\widehat{Z}=4$	$\widehat{Z}=5$
$Z=1$	43	17	32	0	0
$Z=2$	19	12	102	1	2
$Z=3$	17	6	366	22	14
$Z=4$	4	1	92	24	21
$Z=5$	2	1	34	10	40

Klasifikováno bylo 882 objektů, z toho bylo správně zařazeno 485, což činí přibližně $55,0\%$.

Výsledky porovnáme. Nejvyšší úspěšnosti bylo dosaženo klasifikací pomocí shlukování, když byl učební soubor $\mathcal{S}_1$ shlukován metodou nejvzdálenějšího souseda, vzdálenost objektu od shluku byla počítána podle (6.15), úspěšnost klasifikace zde činila přibližně $56,4\%$. Ke srovnatelným výsledkům vedla klasifikace klasifikačním stromem a také ostatní klasifikace za použití shlukové analýzy, pokud byla vzdálenost objektu od shluku počítána podle (6.15). Úspěšnosti okolo 40% bylo dosahováno klasifikací pomocí shlukování, když byla vzdálenost objektu od shluku počítána podle (6.17). Nejméně vhodná se pro tuto úlohu jeví klasifikace diskriminační analýzou.

V předchozím textu jsme předpovídali znečištění ovzduší tak, že jsme klasifikovali objekty (dny) $k = 5$ klasifikačními třídami. Úlohu nyní zjednodušme, zvolme $k = 3$. Každému objektu $\omega_i \in \mathcal{S}$ přiřadíme hodnotu nově zaváděné veličiny Z:

$$\begin{aligned} Z_i = 1 &\iff \mathcal{Z}_i \leq 45,07, \\ Z_i = 2 &\iff 45,07 < \mathcal{Z}_i \leq 77,11, \\ Z_i = 3 &\iff 77,11 < \mathcal{Z}_i. \end{aligned}$$

Připomeňme, že $45,07$ je dolní a $77,11$ horní kvartil veličiny $\mathcal{Z}$.

Úspěšnost predikování znečištění ovzduší, pokud volíme $k = 3$, otestujeme nejprve pro klasifikaci pomocí shlukování, učební soubor budeme shlukovat metodou nejvzdálenějšího souseda, vzdálenost objektu od shluku budeme počítat podle (6.15). Dále provedeme tuto klasifikaci kvadratickou diskriminační analýzou a klasifikačním stromem.

Postup je identický, liší se jen způsob přiřazení odhadu $\widehat{Z}$ objektu ω' z klasifikovaného souboru $\mathcal{S}_2$. Pokud má tedy objekt ω' nejblíže ke shluku A, pak

$$\begin{aligned} \widehat{Z}(\omega') = 1 &\iff \overline{\mathcal{Z}}(A) \leq 45,07, \\ \widehat{Z}(\omega') = 2 &\iff 45,07 < \overline{\mathcal{Z}}(A) \leq 77,11, \\ \widehat{Z}(\omega') = 3 &\iff 77,11 < \overline{\mathcal{Z}}(A). \end{aligned}$$

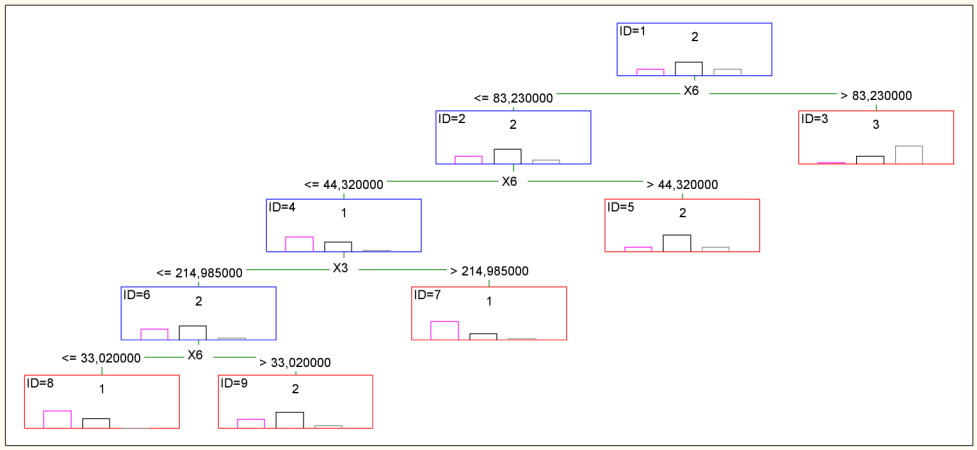
Dostaneme tyto výsledky:

	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$
$Z=1$	0	207	11
$Z=2$	0	369	44
$Z=3$	0	110	82

Klasifikováno bylo 823 objektů, správně 451, což činí přibližně $54,8\%$.

Pro klasifikaci kvadratickou diskriminací obdržíme následující tabulku pro porovnání Z a $\widehat{Z}$ pro klasifikované objekty:

	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$
$Z=1$	156	34	9
$Z=2$	166	198	44
$Z=3$	19	102	109



Obrázek 41: Klasifikační strom.

Zde bylo klasifikováno 837 objektů, správně bylo zařazeno 463, což je přibližně $55,3\%$.

Klasifikační strom zkontruovaný stejným postupem jako v předchozím textu je nyní na Obrázku 41. Pokud ním budeme klasifikovat objekty souboru $\mathcal{S}_2$, dostaneme následující tabulku:

	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$
$Z=1$	99	120	9
$Z=2$	34	341	50
$Z=3$	8	109	112

Klasifikováno bylo 882 objektů, z toho správně 552. Poměr správně zařazených objektů tak činí přibližně $62,6\%$.

V jistém smyslu zjednodušení úlohy přineslo lepší výsledky. Nejvíce se úspěšnost klasifikace zvýšila u kvadratické diskriminace, nejvyšší hodnoty poměru správně zařazených objektů však dosáhla klasifikace klasifikačním stromem – přibližně $62,6\%$. Uvedená klasifikace pomocí shlukování dosáhla srovnatelného výsledku jako při volbě $k=5$.

6.2.4 Modelování předpovídání znečištění ovzduší pomocí klasifikace

V předchozí části této kapitoly jsme testovali úspěšnost použití klasifikačních metod pro predikci znečištění ovzduší, datový soubor $\mathcal{S}$ byl náhodně rozdělen na $\mathcal{S}_1$ a $\mathcal{S}_2$. Pomocí učebního souboru $\mathcal{S}_1$ jsme pak klasifikovali objekty souboru $\mathcal{S}_2$.

Zde zvolíme jiný postup, pokusíme se modelovat reálnou situaci předpovídání znečištění. Učební soubor $\mathcal{S}_1$ utvoří prvních N_1 pozorování základního datového souboru $\mathcal{S}$. Pomocí tohoto učebního souboru budeme klasifikovat (N_1+1)-ní objekt z $\mathcal{S}$ (použijeme opět $k=5$ klasifikačních tříd). Můžeme to interpretovat tak, že máme k dispozici údaje o počasí a znečištění z uplynulých N_1 dnů. V našem značení to jsou hodnoty $X_{1i}, X_{2i}, \ldots, X_{6i}$, $i=1,2,\ldots,N_1$. Pro následující den známe předpověď počasí a dnešní hodnotu znečištění, to představují prediktory $X_{1N_1+1}, X_{2N_1+1}, \ldots, X_{6N_1+1}$, pomocí nich budeme tento (N_1+1)-ní objekt klasifikovat a obdržíme odhad $\widehat{Z}_{N_1+1}$, tím i přibližnou hodnotu znečištění ovzduší PM_{10} na zítřejší den $\mathcal{Z}_{N_1+1}$. Výsledek vyhodnotíme (skutečné hodnoty znečištění známe). V dalším kroku utvoří učební soubor $\mathcal{S}_1$ prvních N_1+1 pozorování

z $\mathcal{S}$, postup opakuje. Postupně tak stanovíme a vyhodnotíme odhady $\widehat{Z}$ pro všech $N - N_1$ pozorování souboru $\mathcal{S}$.

Budeme se zabývat pouze vybranými metodami s ohledem na výsledky v předchozím textu a také na výpočetní nároky. Byly vybrány klasifikace lineární a diskriminační analýzou a klasifikace pomocí shlukování metodou nejvzdálenějšího souseda (vzdálenost objektu a shluku byla počítána podle (6.15)). Volíme $N_1 = 1000$.

Užitím lineární diskriminace v uvedeném postupu dojdeme k tomuto konečnému výsledku:

	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$	$\widehat{Z}=4$	$\widehat{Z}=5$
$Z=1$	46	47	30	1	0
$Z=2$	76	99	65	30	0
$Z=3$	110	191	343	277	24
$Z=4$	9	11	86	109	32
$Z=5$	0	1	16	58	42

Postupně bylo klasifikováno 1703 objektů (dáno volbou $N_2 = 1000$), z toho 639 správně, což činí přibližně $37,5\%$.

Klasifikace kvadratickou diskriminací vede k výsledku:

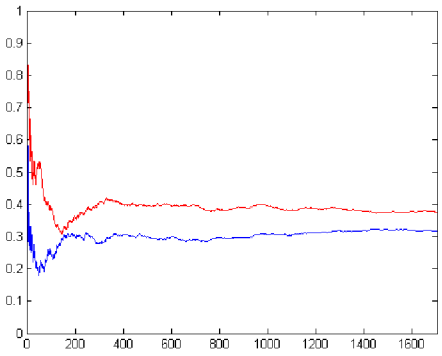
	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$	$\widehat{Z}=4$	$\widehat{Z}=5$
$Z=1$	95	23	5	1	0
$Z=2$	134	81	31	22	2
$Z=3$	223	263	220	218	21
$Z=4$	12	29	74	113	19
$Z=5$	3	4	17	62	31

Z klasifikovaných objektů bylo 540 zařazeno správně, což činí přibližně $31,7\%$.

Poměr správně zařazených objektů byl počítán po vyhodnocení klasifikace každého dalšího objektu, vznikla tak posloupnost $N - N_1$ hodnot. Pro ilustraci, pokud například byl první klasifikovaný objekt ($(N_1 + 1)$-ní objekt $\mathcal{S}$) zařazen správně, je poměr správně klasifikovaných a všech klasifikovaných objektů 1. Jestliže byl dále $(N_1 + 2)$-hý objekt zařazen špatně, činil poměr správně zařazených a všech klasifikovaných objektů $\frac{1}{2}$, pokud byl $(N_1 + 3)$-tí objekt klasifikován správně, činil uvedený poměr $\frac{2}{3}$. Takto bylo postupně stanoveno všech $N - N_1 = 1703$ hodnot. Tyto posloupnosti jsou znázorněny na Obrázku 42 (modrou barvou pro klasifikaci kvadratickou diskriminací, červeně pak pro lineární dikriminaci).

Pro klasifikaci pomocí shlukování metodou nejvzdálenějšího souseda (vzdálenost objektu a shluku byla počítána podle (6.15)) zvolíme $N_1 = 2200$ (kvůli výpočetním nárokům). Obdržíme tuto tabulku pro porovnání Z a $\widehat{Z}$ pro klasifikované:

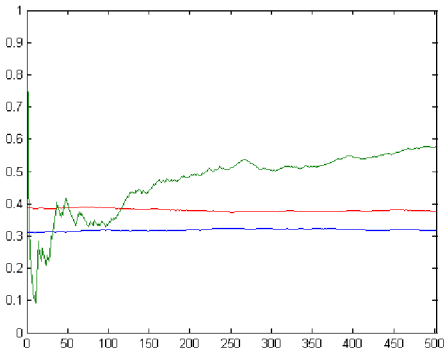
	$\widehat{Z}=1$	$\widehat{Z}=2$	$\widehat{Z}=3$	$\widehat{Z}=4$	$\widehat{Z}=5$
$Z=1$	0	0	50	0	0
$Z=2$	0	0	111	0	0
$Z=3$	0	0	287	3	2
$Z=4$	0	0	36	2	1
$Z=5$	0	0	9	1	1



Obrázek 42: Posloupnosti poměrů správně zařazených objektů, lineární a kvadratická diskriminace.

Klasifikováno bylo 503 objektů, z toho správně 290, což činí přibližně $57,7\%$.

Na Obrázku 43 je znázorněna posloupnost poměrů správně zařazených objektů pro klasifikaci pomocí shlukování zelenou barvou. Pro porovnání je zde zobrazeno i posledních 503 členů posloupností pro kvadratickou (modře) a lineární diskriminaci (červeně).



Obrázek 43: Posloupnosti poměrů správně zařazených objektů pro použité metody.

Poměr správně zařazených objektů s velikostí učebního souboru $\mathcal{S}_1$ rostl. Celkově jsme obdrželi podobné výsledky jako v předchozí podkapitole.

7 Závěr

Část této diplomové práce, která se zabývá teorií, obsahuje popis vybraných klasifikačních metod. Jsou zde uvedeny hlavní výsledky, kterých se pak využívá, text je doplněn ilustrativními příklady. U metod shlukové analýzy byla věnována pozornost volbě koeficientu nepodobnosti shluků, zmíněny byly tři nejznámější způsoby (metody nejbližšího a nejvzdálenějšího souseda, dále metoda průměrné nepodobnosti). Z teorie diskriminační analýzy byly zmíněny metody lineární a kvadratické diskriminace, uvedeny jsou také vztahy, které lze použít pro analytický výpočet. V dalším textu, který byl věnován klasifikačním stromům, uvádíme dvě základní metody pro konstrukci klasifikačních stromů, metodu top-down (shora dolů) a metodu growing-pruning (růst-prořezávání). V závěrečné části každé z kapitol bylo užití metod demonstrováno klasifikací simulovaných dat pomocí počítačového programu STATISTIKA.

V praktické části je popsána aplikace metod na reálná data. Metody byly nejprve porovnávány pro výběr malého rozsahu. Tato data byla převzata z [5], nejlepšího výsledku bylo docíleno při klasifikaci diskriminační analýzou a klasifikačním stromem. Následuje použití těchto metod pro klasifikaci pro větší datový soubor. Úkolem bylo předpovídat znečištění ovzduší na základě meteorologické predikce počasí na následující den. Metody zde byly testovány s různými volbami parametrů. Dobrého výsledku se však dosáhnout nepodařilo. Na základě provedených výpočtů tak lze usoudit, že modely predikce znečištění lze uspokojivěji řešit pomocí regresních metod (viz [8]).

Reference

[1] ANDĚL, J.: *Matematická statistika.* Praha: STLN, 1978.

[2] ANTOCH, J.: *Klasifikace a regresní stromy.* Sborník ROBUST, 1988.

[3] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., STONE, C. J.: *Classification and Regression Trees.* The Wadsworth Statistics/ Probability Series, Belmont, California: Wadsworth, 1984. 358 p. ISBN 978-0412048418.

[4] FORBELSKÁ, M.: *Parametrická a neparametrická diskriminační analýza.* [Disertační práce.] Ostrava: Ostravská univerzita v Ostravě, Přírodovědecká fakulta, 2003.

[5] JOHNSON, R. A., WICHERN, D. W.: *Applied Multivariate Statistical Analysis.* Prentice Hall, 1992. 800 p. ISBN 978-0131877153.

[6] LUKASOVÁ, A., ŠARMANOVÁ, J.: *Metody shlukové analýzy.* Praha: STLN, 1985. 212 p.

[7] SEDLAČÍK, M.: *Využití ROC křivek při konstrukci klasifikačních a regresních stromů.* [Disertační práce.] Brno: Masarykova univerzita v Brně, 2006.

[8] HRDLIČKOVÁ, Z., et al.:*Identification of factors affecting air pollution by dust aerosol PM_{10} in Brno City, Czech Republic.* Atmospheric Environment (2008). doi: 10.1016/j.atmosenv.2008.08.017.

Seznam zkratek a symbolů

ω	objekt
Ω	množina objektů
$\mathbb{R}$	množina reálných čísel
$\mathbb{N}$	množina přirozených čísel
$X_1, X_2, \ldots$	prediktory
$\mathbf{X}$	vektor prediktorů
$\mathcal{X}$	obor hodnot vektoru $\mathbf{X}$
Z	veličina určující typ objektu
$\widehat{Z}$	odhad veličiny Z
$\mathcal{C}$	obor hodnot veličiny Z
$\mathcal{S}, \mathcal{S}_1, \mathcal{S}_2, \ldots$	soubory dat
$(\Omega, \mathfrak{A}, P)$	pravděpodobnostní prostor
$A, A_1, A_2, \ldots, B, \ldots$	shluky; podmnožiny $\mathcal{X}$
$\mathbb{A}, \mathbb{A}_1, \mathbb{A}_2, \ldots$	rozklady množiny $\mathcal{X}$
$\mathcal{A}$	množina všech $\mathbb{A}$
d	koeficient nepodobnosti objektů
D	koeficient nepodobnosti shluků
D_{nn}	koeficient nepodobnosti shluků definovaný metodou nejbližšího souseda
D_{fn}	koeficient nepodobnosti shluků definovaný metodou nejvzdálenějšího souseda
D_{av}	koeficient nepodobnosti shluků definovaný metodou průměrné nepodobnosti
$N_2(\boldsymbol{\mu}, \mathbf{V})$	dvojrozměrné normální rozdělení pravděpodobnosti s vektorem středních hodnot $\boldsymbol{\mu}$ a varianční maticí $\mathbf{V}$
R	chyba špatné klasifikace
$d_1, d_2, \ldots$	klasifikační funkce pro lineární diskriminaci
$D_1, D_2, \ldots$	klasifikační funkce pro kvadratickou diskriminaci
$T, T_1, T_2, \ldots$	klasifikační stromy
$t, t_1, t_2, \ldots$	uzly klasifikačního stromu
$\widetilde{T}$	množina listů klasifikačního stromu T

Seznam příloh

P1 DVD obsahující použité statistické soubory (ve formátech tabulek programů STATISTICA a MS EXCEL):
data_poc.sta, data_poc.xls, data_poc_S1.sta, data_poc_S1.xls (použity v podkapitole 6.2), data_sal.sta, data_sal.xls, data_sal_S1.sta, data_sal_S1.xls, data_sal_S2.sta, data_sal_S2.xls (podkapitola 6.1), data_sim.sta, data_sim.xls (podkapitoly 3.3, 4.2, 5.5),
dále spočtené klasifikační stromy v programu STATISTICA (ve formátu projektu programu STATISTICA):
sim_strom.spf (podkapitola 5.5), sal_strom.spf (podkapitola 6.1), poc_strom_3.spf, poc_strom_5.spf (podkapitola 6.2)
a elektronickou verzi diplomové práce ve formátu pdf