UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Skryté Markovovy řetězce – diskrétní případ

Katedra matematické analýzy a aplikací matematiky
Vedoucí diplomové práce: **Mgr. Kamila Fačevicová, Ph.D.**
Vypracovala: **Bc. Aneta Nyklová**
Studijní program: N1103 Aplikovaná matematika
Studijní obor : Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2020

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Aneta Nyklová

Název práce: Skryté Markovovy řetězce – diskrétní případ

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: Mgr. Kamila Fačevicová, Ph.D.

Rok obhajoby práce: 2020

Abstrakt: Cílem diplomové práce je představit diskrétní případ skrytých Markovových řetězců. Úvodní část práce se věnuje pojmům úzce souvisejícími se skrytými Markovovými řetězci. Následuje vysvětlení teorie skrytých Markovových řetězců. Součástí teoretické části jsou příklady, na kterých je problematika ilustrována. V závěrečné kapitole práce se nachází aplikace probraných metod na reálný datový soubor.

Klíčová slova: Markovovy řetězce, evaluace, dekódování, učení, dopředný algoritmus, zpětný algoritmus, Viterbiho algoritmus, Baum-Welch algoritmus

Počet stran: 83

Počet příloh: 2

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Aneta Nyklová

Title: Hidden Markov chains – a discrete case

Type of thesis: Master's

Department: Department of Mathematical Analysis and Applications of Mathematics

Supervisor: Mgr. Kamila Fačevicová, Ph.D.

The year of presentation: 2020

Abstract: The goal of the diploma thesis is to introduce a discrete case of the Hidden Markov chains. The initial part of thesis presents terms which are closely related to Hidden Markov chains. In the next section, the theory of Hidden Markov chains is presented. The theoretical part includes examples on which issues are illustrated. Methods are applied on a real dataset in the final chapter.

Key words: Markov chains, evaluation, decoding, learning, forward algorithm, backward algorithm, Viterbi algorithm, Baum-Welch algorithm

Number of pages: 83

Number of appendices: 2

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením Mgr. Kamily Fačevicové, Ph.D. a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Poděkování

Ráda bych poděkovala vedoucí diplomové práce Mgr. Kamile Fačevicové, Ph.D. za cenné rady, věcné připomínky, vstřícnost a čas strávený při konzultacích.

Úvod

Cílem diplomové práce je představit diskrétní případ skrytých Markovových řetězců. V prvopočátku bylo potřeba dané téma prostudovat a následně zkompletovat. Pro lepší přiblížení problematiky je do celé práce zakomponován související příklad.

V pilotní kapitole je krátce představen pojem náhodný proces. Tento termín je v práci zmíněn, jelikož jeden z typů náhodných procesů je ukryt právě uvnitř skrytých Markovových řetězců.

V následující podkapitole jsou uvedeny Markovovy řetězce. Zde se nenechme zmást. Ačkoliv se podle názvu může jevit, že právě Markovovy řetězce jsou klíčovým termínem této práce, není tomu úplně tak. Vynechané slovíčko "skryté" zde hraje velikou roli. Tohle slovo symbolizuje skryté stavy, které odpovídají odlišné veličině oproti té veličině, která je pozorována. Rozdíl můžeme tedy najít v pozorované veličině, v našem případě je to navíc diskrétně rozdělená náhodná veličina.

Prostřednictvím skrytých Markovových řetězců jsou řešeny tři základní úlohy, které bývají také někdy označovány jako problém. A sice evaluace, dekódování a učení. Tyto úlohy jsou řešeny pomocí různých algoritmů. Úlohy jsou spolu s algoritmy v práci popsány, představeny pomocí příkladů a následně aplikovány na reálných datech v praktické části.

Jak je řečeno výše, v práci se objevuje také praktická část. Na tomto místě je téma demonstrováno na reálných datech, která se týkají biologické oblasti.

Ačkoliv existuje komerční software `Latent Gold` pro práci se skrytými Markovovými řetězci, výpočty v této diplomové práci byly zpracovány v softwaru R [11].

Kapitola 1

Základní informace a pojmy

V této kapitole uvedeme základní teoretické pojmy potřebné pro správné pochopení problematiky skrytých Markovových řetězců.

V první řadě musíme vysvětlit termíny, které se budou v práci vyskytovat a souvisí s hlavním tématem. Úvodní kapitola byla vytvořena pomocí literatury [3] a [6] .

1.1. Náhodný proces

Náhodný proces je pojem, který je úzce spjat s Markovovými řetězci.

Před nadefinováním náhodného procesu si nejprve zavedeme pojem náhodná veličina.

Definice 1.1.1 (Náhodná veličina)**.** *Nechť* $(\Omega, \mathcal{A}, P)$ *je pravděpodobnostní prostor. Reálnou funkci* $X : \Omega \to \mathbf{R}^1$ *nazveme náhodnou veličinou, pokud pro každé* $x \in \mathbf{R}^1$ *platí*

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{A}. \tag{1.1}$$

Uspořádaná trojice $(\Omega, \mathcal{A}, P)$ v definici 1.1.1 se nazývá pravděpodobnostní prostor nebo také Kolmogorovo pravděpodobnostní pole. Množina Ω značí prostor elementárních jevů $\omega \in \Omega$. Neprázdný systém $\mathcal{A}$ nazýváme jevovým polem. Funkce P označuje pravděpodobnost. Pro více informací o této problematice odkážeme zájemce na [3].

Náhodná veličina je součástí náhodného procesu, jenž bývá nazýván rovněž jako stochastický proces. Náhodný proces je označován jako zobecnění náhodné veličiny. Při realizaci jedné náhodné veličiny je výsledkem jedna jediná hodnota (například studentova známka u zkoušky). Výsledek stochastického procesu je však funkce nebo také posloupnost (například posloupnost známek u zkoušek v průběhu jednoho týdne). Náhodný proces je posloupností několika náhodných veličin v čase.

Definice 1.1.2 (Náhodný proces)**.** *Nechť* $(\Omega, \mathcal{A}, P)$ *je pravděpodobnostní prostor a nechť* $T \subset \mathbf{R}$*. Systém náhodných veličin* $\{X_t, t \in T\}$ *definovaných na* $(\Omega, \mathcal{A}, P)$ *se nazývá náhodný proces.*

Symbol $t \in T$ v definici 1.1.2 označuje čas. Dalším symbolem, který budeme v následujícím textu využívat je $\mathbf{S}$, což značí množinu možných realizací náhodných veličin.

Markovský řetězec je určitý typ náhodného procesu. Zde si předvedeme, jaký může být náhodný proces:

1. Náhodný proces s diskrétním časem:

   Pokud T je z množiny celých čísel $\mathbf{Z}$ nebo z množiny přirozených čísel $\mathbf{N}_0$, pak se jedná o proces s diskrétním časem. Náhodný proces s diskrétním časem se dále člení na:

   (a) Náhodný proces s diskrétním časem a diskrétními stavy:

   Tento náhodný proces nastává v případě, kdy kromě diskrétního času jsou také realizace náhodných veličin X_t diskrétní, množinu $\mathbf{S}$ pak máme konečnou nebo nekonečně spočetnou.

   (b) Náhodný proces s diskrétním časem a spojitými stavy:

   Náhodný proces tohoto typu pozorujeme v případě, kdy jsou X_t spojité náhodné veličiny a $\mathbf{S}$ je tedy nespočetná množina (například interval).

2. Náhodný proces se spojitým časem:

   Naopak proces se spojitým časem nastane v případě, kdy $T \in \langle a, b\rangle$, pro $-\infty \leq a < b \leq \infty$. Také tento náhodný proces se dále dělí na:

   (a) Náhodný proces se spojitým časem a diskrétními stavy:

   Tento druh náhodného procesu se objevuje v případě diskrétních realizací náhodných veličin X_t, množina hodnot náhodných veličin $\mathbf{S}$ nabývá konečných nebo nekonečně spočetných hodnot. Navíc, jak je zmíněno výše, T nabývá hodnot z nějakého intervalu.

   (b) Náhodný proces se spojitým časem a spojitými stavy:

   Poslední možností je náhodný proces se spojitým časem a spojitými náhodnými veličinami X_t, které mohou nabýt výhradně hodnot ze spojitého intervalu (množina $\mathbf{S}$ je nespočetná).

1.2. Markovovy řetězce

Před samotným uvedením skrytých Markovových řetězců si krátce představíme klasické Markovovy řetězce.

Markovův řetězec je vlastně určitým druhem náhodného procesu, který jsme si zavedli výše. Markovův řetězec spadá pod první uvedený stochastický proces. Jedná se tedy o náhodný proces s diskrétním časem a diskrétními stavy. Pro objasnění ještě uvedeme vysvětlení terminologie modelu a řetězce. Markovův řetězec je Markovův model s diskrétním časem.

Jelikož se jedná o velice zajímavé téma, bylo by na místě zveřejnit jméno autora. Markovské modely získaly název po svém tvůrci, kterým byl ruský matematik Andrej Markov (14. červen 1856 – 20. červenec 1922).

Poznámka 1.2.1. *V následujícím textu budeme pracovat s množinou stavů, pro niž platí* $\mathbf{S} = \{1, 2, \ldots, N\}$.

Definice 1.2.1 (Markovův řetězec)**.** *Řekneme, že náhodný proces* $\{X_t, t \in \mathbf{N}_0\}$*, pro který* $\mathbf{S}$ *je množina stavů, nazýváme Markovův řetězec s diskrétním časem, pokud*

$$\begin{aligned} P(X_{t+1} =& j|X_t = i, X_{t-1} = i_{t-1}, \cdots, X_1 = i_1, X_0 = i_0) \\ &= P(X_{t+1} = j|X_t = i), \end{aligned} \tag{1.2}$$

pro každé $t \in \{0, 1, \dots\}$ *a každé* $i_0, i_1, \dots, i_{t-1}, i, j \in \mathbf{S}$*, pro které* $P(X_t = i, X_{t-1} = i_{t-1}, \dots, X_1 = i_1, X_0 = i_0) > 0$*.*

Definice 1.2.2 (Homogenní Markovův řetězec)**.** *Pokud pravděpodobnosti přechodu nezávisí na čase* t*, ve kterém nastává přechod, tzn. pokud*

$$p_{ij}(t, t+1) = p_{ij}, \quad i, j \in \mathbf{S}\ \forall t,$$

pak Markovův řetězec s dikréním časem nazveme homogenní.

Uvedená podmínka 1.2 v definici 1.2.1 se nazývá Markovova podmínka neboli Markovova vlastnost. Náhodný proces, který disponuje touto vlastností, se někdy také označuje jako náhodný proces "bez paměti". Markovská vlastnost nám totiž říká, že pravděpodobnost přechodu z jednoho stavu X_t do dalšího stavu X_{t+1} není závislá na tom, jak systém dosáhnul stavu X_t.

V předchozím odstavci mluvíme o pravděpodobnosti přechodu, jež je nezbytnou součástí Markovových řetězců.

Definice 1.2.3 (Pravděpodobnost přechodu prvního řádu)**.** *Pravděpodobností přechodu prvního řádu (neboli pravděpodobnost přechodu za jeden krok) nazveme podmíněnou pravděpodobnost, pro kterou platí*

$$p_{ij}(t, t+1) = P(X_{t+1} = j|X_t = i), \quad i, j \in \mathbf{S}, \quad t = 0, 1, \cdots, T-1.$$

- Pravděpodobnosti přechodu se zapisují do matice pravděpodobností přechodu $\mathbf{P}(t, t+1) = (p_{ij}(t, t+1))_{i,j=1}^{N}$, kde N značí počet stavů. V dalším textu budeme pracovat pouze s homogenními Marovovými řetězci a matici pravděpodobností přechodu proto můžeme psát ve zkráceném tvaru $\mathbf{P} = (p_{ij})_{i,j=1}^{N}$.

- Matice pravděpodobností přechodu $\mathbf{P}$ vypadá následovně

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ p_{31} & p_{32} & \cdots & p_{3N} \\ \vdots & \vdots & \vdots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{pmatrix}.$$

- $p_{i,i}$ označuje pravděpodobnost, kdy model zůstává ve stavu i.

  $p_{i,i+1}$ představuje přechodovou pravděpodobnost přechodu ze stavu i do dalšího stavu $i+1$.

- Pro pravděpodobnosti vycházející z totožného stavu i musí platit tato rovnost $\sum_j p_{ij} = 1$ a současně také pravděpodobnosti přechodu musí splňovat $0 \leq p_{ij} \leq 1$.

- Pravděpodobnosti přechodu určují pravděpodobnost změny stavu při současné změně času.

Dalšími pravděpodobnostmi jsou počáteční pravděpodobnosti zapisované do vektoru $\mathbf{p}^{(0)} = (p_1^{(0)}, p_2^{(0)}, \cdots, p_N^{(0)})$.

Definice 1.2.4 (Počáteční rozdělení pravděpodobnosti)**.** *Rozdělení Markovova řetězce v čase $t = 0$ nazveme počátečním rozdělením a značíme*

$$p_i^{(0)} = P(X_0 = i), \quad i \in \mathbf{S}. \tag{1.3}$$

- Počáteční pravděpodobnosti musí splňovat podmínky $\sum_{i \in \mathbf{S}} p_i^{(0)} = 1$ a

  $p_i^{(0)} \geq 0, \forall i \in \mathbf{S}$.

- Počáteční pravděpodobnosti nám říkají, jak pravděpodobný je začátek v jednotlivých stavech.

Kapitola 2

Skryté Markovovy řetězce

Nyní můžeme přejít k samotným Markovovým řetězcům. Při tvorbě této kapitoly bylo využito [2], [8], [9], [10], [12] a [16].

2.1. Úvod do problematiky

Základy skrytých Markovových modelů položil spolu se svými kolegy Leonard Esau Baum. Markovovy modely nachází využití hned v několika sférách života. Jejich uplatnění se objevuje například v různých druzích rozpoznávání – obličeje, písma, podpisů, gest atd. Své místo nalézají také v medicíně nebo v biomedicíne, kde se například analyzují signály EKG či EEG. Svou roli zastávají také v biologii nebo v bioinformatice pro analýzu biologických sekvencí. Mezi další uplatnění patří meteorologie, kdy se zkoumá déšť, směr větru nebo třeba zemětřesení. Další zajímavé aplikace jsou z oblasti techniky, jako je mobilní robotika. V oblasti financí se skryté Markovovy řetězce využívají k detekcím anomálií, jako jsou podvody s kreditními kartami a další. Jedno z nejčastějších využití skrytých Markovových modelů je v rozpoznávání řeči.

Rozdíl oproti klasickým Markovovým řetězcům spočívá v tom, že namísto stavu řetězce sledujeme pozorování, které je realizací diskrétně rozdělené náhodné veličiny. U skrytých Markovových řetězců známe pouze to, co pozorujeme. Zde skryté tedy znamená, že vidíme pouze pozorování, které stavy generují, ale nic jiného. Stavy jsou skryté a odpovídají jiné veličině nežli té veličině, kterou skutečně

pozorujeme. Z tohoto faktu je také odvozen název skrytý. Avšak stavy nejsou méně důležité, naopak hrají velkou roli.

Skryté Markovovy řetězce se dělí na řetězce, kde pozorování mají diskrétní anebo spojité rozdělení pravděpodobnosti. My se zde budeme zabývat prvním typem a sice skrytými Markovovými řetězci s diskrétní povahou pozorování. Zde vidíme, že máme tedy dva typy skrytých Markovových modelů. A sice diskrétní skrytý Markovův model, kde náhodná veličina, kterou pozorujeme nabývá hodnot z diskrétního rozdělení pravděpodobnosti. Druhým typem je spojitý skrytý Markovův model. Zde se náhodná veličina řídí spojitým rozdělením pravděpodobnosti.

V předchozí kapitole jsme si uvedli, že Markovův řetězec disponuje Markovskou vlastností. Stejnou vlastnost má také skrytý Markovův řetězec. Pro připomenutí, rysem Markovské vlastnosti je "bezpamětnost". Tato vlastnost se vyznačuje tím, že pravděpodobnost přechodu z jednoho stavu do druhého nezávisí na tom, jak se systém do prvního stavu dostal. Znamená to, že stav v budoucnu závisí pouze na stavu současném a na žádném jiném z minulosti.

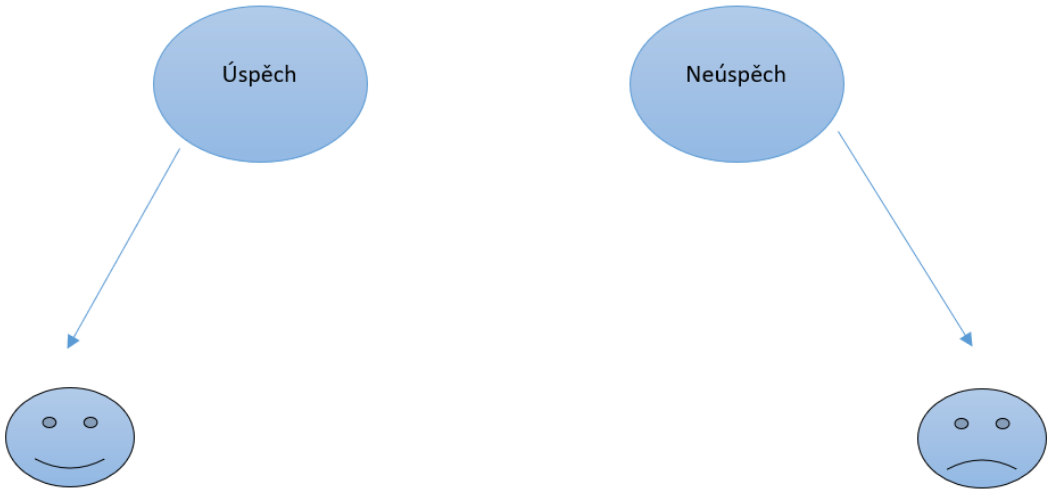
Jelikož se vše lépe chápe na konkrétním příkladu, tak se budeme snažit vše vysvětlit pomocí příkladu. V celé práci budeme postupně rozvíjet jeden příklad.

Příklad 2.1.1. *Skrytý Markovův řetězec budeme jednoduše postupně vysvětlovat na příkladech se studentem, který právě prochází zkouškových obdobím. Nyní si zhruba nastíníme, o čem příklad bude a v dalších příkladech budeme navazovat. Jedná se tedy o příklad s pokračováním.*

Představme si nějakého chytrého studenta, který má v jednom týdnu více zkoušek. Minimálně jednu za den. Tento student každý den po zkoušce volá domů, aby se pochlubil s výsledkem zkoušky. Jeho rodina ho už zná natolik, že dokáže odhadnout výsledek již z jeho nálady. Budeme zde uvažovat pouze dva stavy — úspěch (úspěšný výsledek zkoušky) anebo neúspěch (neúspěšný výsledek zkoušky). Známky neuvažujeme. Navíc máme dvě hodnoty pozorované proměnné buď veselý (student je veselý), nebo smutný (student je smutný).

Pokud je student veselý, rodiče usoudí, že zkouška dopadla úspěšně. Pokud

nemá dobrou náladu, pak nastává druhá možnost a sice, že zkouška nebyla úspěšná.



Obrázek 2.1: Zkouškové období.

2.2. Pojmy

V této chvíli si nadefinujeme pojmy, které potřebujeme:

- $\{X_t\}$ ... Homogenní Markovův řetězec pro $t = 0, 1, \ldots, T$.
- $\mathbf{S} = \{1, 2, \ldots, N\}$ ... Množina skrytých stavů.
- $\mathbf{p}^0 = (p_1^0, p_2^0, \cdots, p_N^0)$ ... Počáteční rozdělení pravděpodobnosti, které vyjadřuje pravděpodobnost, se kterou model začne v daném skrytém stavu.
- $\mathbf{P} = (p_{ij})_{i,j=1}^N$ ... Matice pravděpodobností přechodu mezi skrytými stavy za jeden krok.
- $\mathbf{O} = \{O_0, O_1, \ldots, O_T\}$ ... Posloupnost diskrétních náhodných veličin pozorovaných v čase t, kde $t = \{0, 1, \ldots, T\}$.
- $V = \{v_1, v_2, \ldots, v_M\}$ ... Množina možných realizací náhodné veličiny O_t, kterou pro každé t uvažujeme stejnou.
- t ... Čas, ve kterém sledujeme realizaci náhodné veličiny O_t.
- $b_j(k) = P(O_t = v_k | X_t = j), \quad k = 1, \ldots, M$ ... Rozdělení, kterým se řídí náhodná veličina O_t, v případě, že skrytý systém je ve stavu $j \in \mathrm{S}$ v čase t.
- $\mathbf{B} = (b_j(k))_{j,k=1}^{N,M}$ ... Matice rozdělení pravděpodobností výstupu.
  - Matice pravděpodobností výstupu $\mathbf{B}$ má tvar

$$\mathbf{B} = \begin{pmatrix} b_1(1) & b_1(2) & \cdots & b_1(M) \\ b_2(1) & b_2(2) & \cdots & b_2(M) \\ b_3(1) & b_3(2) & \cdots & b_3(M) \\ \vdots & \vdots & \vdots & \vdots \\ b_N(1) & b_N(2) & \cdots & b_N(M) \end{pmatrix}.$$

  - Pravděpodobnosti $b_j(k)$ označují praděpodobnost, s jakou je generováno pozorování v_k ze stavu j.

  - Také zde platí pro pravděpodobnosti pocházející ze stejného stavu $\sum_k b_j(k) = 1, \forall j$ a zároveň pro pravděpodobnosti výstupu platí $0 \leq b_j(k) \leq 1, \forall j, k$.

- $\lambda = (\mathbf{P}, \mathbf{B}, \mathbf{p}^{(0)})$ ... Skrytý Markovův model.

  - Vidíme, že k nadefinování skrytého Markovového modelu potřebujeme dvě matice a jeden vektor – matice pravděpodobností přechodu, matice pravděpodobností výstupu a vektor počátečních pravděpodobností.

2.3. Skrytý Markovův řetězec – diskrétní případ

Nyní se podíváme, jak skrytý Markovův řetězec pracuje:

1. V prvním kroku se z počátečního rozdělení $p^{(0)}$ generuje úvodní skrytý stav $X_0 = i$ a čas t je roven nule.

2. V kroce druhém se vygeneruje pozorování O_t na základě matice pravděpodobností výstupu $\mathbf{B}$, tedy rozdělení pravděpodobnosti výstupu přiměřené skrytému stavu $X_t = i$.

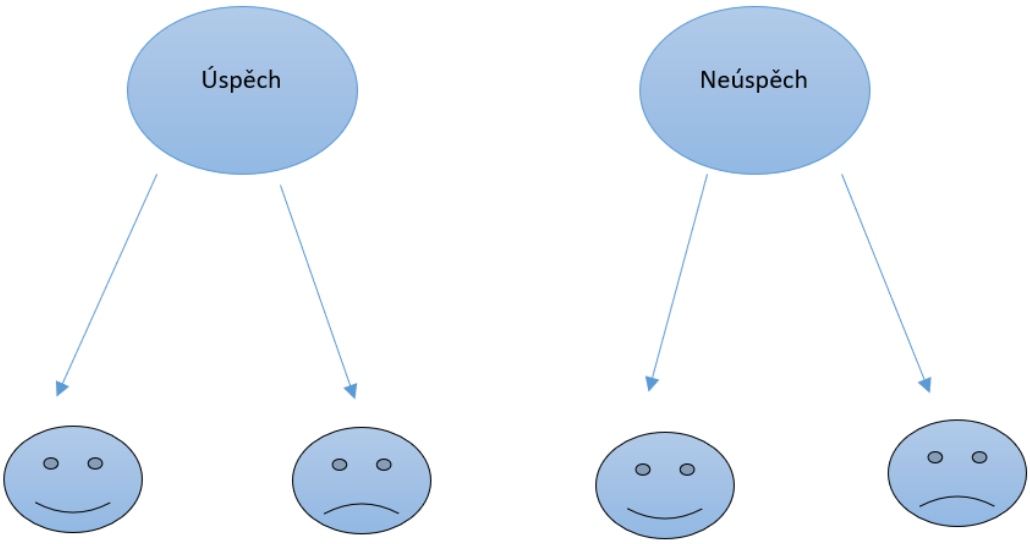
3. V další fázi se podle matice pravděpodobností přechodu $\mathbf{P}$ generuje nový skrytý stav $X_{t+1} = j$ a čas t se položí rovno $t + 1$.

4. Do té doby, kdy bude platit nerovnost $t < T$ se druhý, třetí a čtvrtý krok opakuje.

5. V konečné fázi procesu získáme dvě posloupnosti, posloupnost skrytých stavů $\mathbf{X} = \{X_0, X_1, \ldots, X_T\}$ a posloupnost pozorování $\mathbf{O} = \{O_0, O_1, \ldots, O_T\}$.

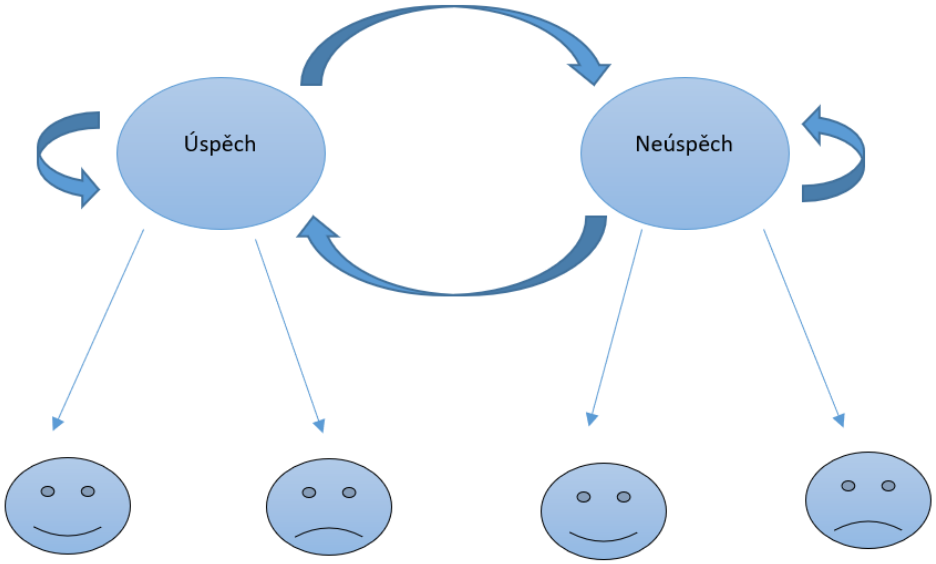
Příklad 2.3.1. *Navážeme na předchozí příklad a trošku ho "zkomplikujeme". Nyní budeme předpokládat, že student je převážně veselý, když zkoušku zvládne úspěšně, ale mohlo ho znepokojit něco jiného. Na druhou stranu student je převážně*

smutný, když zkouška nedopadne dobře, ale mohla se mu stát nějaká veselá událost. To znamená, že student může být smutný, navzdory tomu, že zkouška dopadla dobře. Stejně tak může být veselý v případě, že zkouška nebyla úspěšná (Obrázek 2.2).



Obrázek 2.2: Skryté stavy a pozorování.

Nyní přidáme informace, za jakých okolností může být student veselý či smutný. Tyto hodnoty budou zastupovat výstupní pravděpodobnosti. Navíc také doplníme hodnoty, které nám budou říkat, jaká je pravděpodobnost úspěchu nebo neúspěchu následující den. Tyto hodnoty budou reprezentovat pravděpodobnosti přechodu (Obrázek 2.3). Abychom měli všechny potřebné informace, je potřeba také zavést počáteční rozdělení pravděpodobností. V našem případě jsou to pravděpodobnosti určující, že student u první zkoušky uspěje nebo naopak neuspěje.



Obrázek 2.3: Skrytý Markovův řetězec.

To, k čemu se zde snažíme přirovnat skrytý Markovův řetězec je fakt, že rodiče neví výsledek, vědí pouze studentovu náladu. Tedy výsledek zkoušky je skrytý stav a studentova nálada je to dané pozorování (Obrázek 2.3).

○

2.4. Základní úlohy

Za pomoci skrytých Markovových řetězců se řeší tři základní úlohy. Tyto úlohy se nazývají evaluace, dekódování a učení. V následující části práce se na tyto úlohy a jejich algoritmy podíváme podrobněji.

- Evaluace

  Evaluace znamená, s jakou pravděpodobností je posloupnost pozorování $\mathbf{O}$ známým modelem λ vygenerována.

- Dekódování

  Při dekódování hledáme nejpravděpodobnější posloupnost skrytých stavů $\mathbf{X} = \{X_0, X_1, \ldots, X_T\}$, která odpovídá dané posloupnosti pozorování $\mathbf{O}$ a modelu λ.

- Učení

  Učení bychom mohli vysvětlit jako hledání optimálních hodnot $\mathbf{p}^{(0)}, \mathbf{P}, \mathbf{B}$ na základě posloupnosti pozorování $\mathbf{O}$.

Každá úloha, která je také někdy nazývána slovem problém, je řešitelná za pomocí určitého algoritmu.

2.4.1. Evaluace

Prvním z problémů je evaluace, jinými slovy vyhodnocení pravděpodobností. Ptáme se na otázku, s jakou pravděpodobností model generuje pozorování, pokud máme tuto posloupnost pozorování $\mathbf{O} = \{O_0, O_1, \ldots, O_T\}$ zadanou a také známe model λ. Nejjednodušším způsobem výpočtu pravděpodobnosti posloupnosti pozorování $\mathbf{O}$ je vypočítání všech posloupností skrytých stavů délky T, kde T je počet pozorování.

Posloupností stavů rozumíme

$$\mathbf{X} = (X_0, X_1, \ldots, X_T), \tag{2.1}$$

kde X_0 značí úvodní stav. Pravděpodobnost posloupnosti pozorování $\mathbf{O}$ je při dané posloupnosti stavů $\mathbf{X}$ a platnosti modelu λ

$$P(\mathbf{O}|\mathbf{X}, \lambda) = \prod_{t=0}^{T} P(O_t|X_t, \lambda), \tag{2.2}$$

kde se předpokládá statistická nezávislost pozorování. Nyní máme

$$P(\mathbf{O}|\mathbf{X}, \lambda) = b_{X_0}(O_0) \cdot b_{X_1}(O_1) \cdot \cdots \cdot b_{X_T}(O_T). \tag{2.3}$$

Pravděpodobnost posloupnosti skrytých stavů $\mathbf{X}$ můžeme zapsat následovně

$$P(\mathbf{X}|\lambda) = p^{(0)}_{X_0} p_{X_0X_1} p_{X_1X_2} \cdot \cdots \cdot p_{X_{T-1}X_T}. \tag{2.4}$$

Pravděpodobnost, že $\mathbf{O}$ nastane společně s $\mathbf{X}$, což znamená spojení $\mathbf{O}$ a $\mathbf{X}$, je

$$P(\mathbf{O}, \mathbf{X}|\lambda) = P(\mathbf{O}|\mathbf{X}, \lambda) P(\mathbf{X}|\lambda). \tag{2.5}$$

Za platnosti modelu λ obdržíme výslednou pravděpodobnost pozorování $\mathbf{O}$, kde sčítáme přes všechny posloupnosti skrytých stavů $\mathbf{X}$ a následně každé $X_0, \ldots, X_T$ pravděpodobnosti $P(\mathbf{O}, \mathbf{X}|\lambda)$ z předchozího vztahu 2.5

$$\begin{aligned} P(\mathbf{O}|\lambda) &= \sum_{\mathbf{X}} P(\mathbf{O}|\mathbf{X}, \lambda) P(\mathbf{X}|\lambda) \\ &= \sum_{X_0, X_1, \ldots, X_T} p^{(0)}_{X_0} b_{X_0}(O_0) p_{X_0X_1} b_{X_1}(O_1) \cdot \cdots \cdot p_{X_{T-1}X_T} b_{X_T}(O_T). \end{aligned} \tag{2.6}$$

Vysvětlení této pravděpodobnosti je následující. Na počátku, kdy čas t je roven nule, jsme ve stavu X_0 s pravděpodobností $p^{(0)}_{X_0}$ a s pravděpodobností $b_{X_0}(O_0)$ generujeme O_0. Posuneme se z času 0 do času 1, přejdeme s pravděpodobností $p_{X_0X_1}$ ze stavu X_0 do stavu X_1 a s pravděpodobností $b_{X_1}(O_1)$ vygenerujeme další pozorování O_1. V tomto postupu pokračujeme do času T a s pravděpodobností $p_{X_{T-1}X_T}$ přecházíme ze stavu X_{T-1} do posledního stavu X_T a s pravděpodobností $b_{X_T}(O_T)$ generujeme pozorování O_T.

Tento výpočet je velice náročný již pro malé N a T. Proto se tento problém řeší s pomocí některého z následujících algoritmů.

1. Dopředný (Forward) algoritmus
2. Zpětný (Backward) algoritmus

1. **Dopředný algoritmus**

   Algoritmus se skládá ze tří kroků, ve kterých se objevuje nová proměnná $\alpha_t(i)$. Tato proměnná určuje pravděpodobnost, že posloupnost, kterou pozorujeme je posloupnost $\{O_0, O_1, \ldots, O_t\}$. Navíc, že je proces ve stavu $X_t = i$ v čase t. Uvažujeme všechny možnosti, jelikož stavy $X_0, X_1, \ldots, X_{t-1}$ nemáme stanoveny.

   Nová proměnná má tvar

   $$\alpha_t(i) = P(O_0, O_1, \ldots, O_t, X_t = i|\lambda). \tag{2.7}$$

   Kroky dopředného algorimu jsou následující:

   (a) Inicializace

   $$\alpha_0(i) = p_i^{(0)} b_i(O_0), \quad i = 1, 2, \ldots, N. \tag{2.8}$$

   Stanovíme, s jakou pravděpodobností proces započne ve stavu i, a že v tomto stavu se vygeneruje realizace O_0, na základě počátečního rozdělení pravděpodobnosti $p_i^{(0)}$. Po provedení tohoto kroku pro veškeré stavy množiny $\mathbf{S}$, získáme úvodní hodnoty $\alpha_0(i)$. Pro každý stav $j = 1, 2, \ldots, N$ a pro každý časový okamžik $t = 0, 1, \ldots, T-1$ bude postačující vzít v úvahu pouze pravděpodobnosti přechodu ze všech stavů i do stavu j a vygenerování náhodné veličiny O_{t+1} tímto stavem.

Následuje další krok, kterým je indukce.

(b) Indukce

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) p_{ij}\right] b_j(O_{t+1}), \qquad j = 1, 2, \ldots, N, \quad t = 0, 1, \ldots, T-1. \tag{2.9}$$

Jedná se o nejdůležitější krok v dopředném algoritmu. V této fázi bereme v potaz veškeré možnosti, jak jsme se mohli v čase $t+1$ do stavu j dostat při generování náhodné veličiny O_{t+1}.

$\alpha_t(i)$ je pravděpodobnost společné události, kde jsou pozorována pozorování $O_0, O_1, \ldots, O_t$ a i je stavem v čase t. Potom výstup $\alpha_t(i)p_{ij}$ je pravděpodobností společné události, kde pozorujeme $O_0, O_1, \ldots, O_t$ a stavu j je dosaženo v čase $t+1$ přes stav i v čase t. Součet tohoto výstupu přes všech N možných stavů, i $(i = 1, 2, \ldots, N)$ v čase t, vede k pravděpodobnoti stavu j v čase $t+1$ se všemi předchozími částečnými pozorování. Pokud známe stav j, můžeme vidět, že $\alpha_{t+1}(j)$ je získána spočítáním pro pozorování O_{t+1} ve stavu j.

(c) Ukončení

$$P(\mathbf{O}|\lambda) = \sum_{i=1}^{N} \alpha_T(i). \tag{2.10}$$

Uvedená pravděpodobnost je stanovena pomocí ukončovacího kroku jako součet hodnot $\alpha_T(i)$.

Touto sumou obdržíme požadovanou pravděpodobnost $P(\mathbf{O}|\lambda)$ prostřednictvím součtu konečných dopředných proměnných $\alpha_T(i)$.

Podívejme se nyní na dopředný algoritmus pohledem příkladu (pokračování předchozího příkladu). Jako výpočetní nástroj byl použit software R [11].

Příklad 2.4.1. *Jak jsme si řekli v předchozím příkladu – uvažujeme studenta a jeho rodiče. Při úloze evaluace chceme vědět, s jakou pravděpodobností je vygenerována daná posloupnost při známém modelu λ. Úvodními informacemi z úvodního modelu jsou počáteční pravděpodobnosti, tedy počáteční pravděpodobnost úspěchu či neúspěchu studenta u zkoušky. Dalšími informacemi jsou pravděpodobnosti přechodu, těmi rozumíme s jakou pravěpodobností může student "přecházet" z úspěšných zkoušek na neúspěšné a naopak, plus také z úspěšné na další úspěšnou nebo z neúspěšné na neúspěšnou. Informace, které dále potřebujeme vědět jsou pravděpodobnosti výstupu, kterými jsou myšleny pravděpodobnosti výstupních pozorování "veselý" či "smutný student".*

Další informace, kterou k tomu příkladu potřebujeme vědět, je posloupnost nálady studenta za uplynulý týden. Po telefonátech studenta rodiče usoudí, že v pondělí byl veselý, v úterý byl také veselý a ve středu znovu veselý, ale ve čtrvtek byl smutný a v pátek byl také smutný. Nálady studenta v jednotlivých dnech tedy tvoří naši posloupnost pozorování $\mathbf{O}$. Abychom zjistili, jak je výsledek závislý na daných parametrech, tedy pravděpodobnost $\mathbf{O}$, vyzkoušíme si více variant.

- *Počáteční rozdělení* $\mathbf{p}^{(0)}$

$$\boldsymbol{p}^{(0)} = (p_1^{(0)}; p_2^{(0)}).$$

- *Matice pravděpodobností přechodu* $\boldsymbol{P}$

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}.$$

- *Matice pravděpodobností výstupu* $\boldsymbol{B}$

$$\mathbf{B} = \begin{pmatrix} b_1(1) & b_1(2) \\ b_2(1) & b_2(2) \end{pmatrix}.$$

.

- Scénář 1

  V první variantě volíme parametry následovně:

  – Počáteční rozdělení pravděpodobností – $\mathbf{p}^{(0)}$

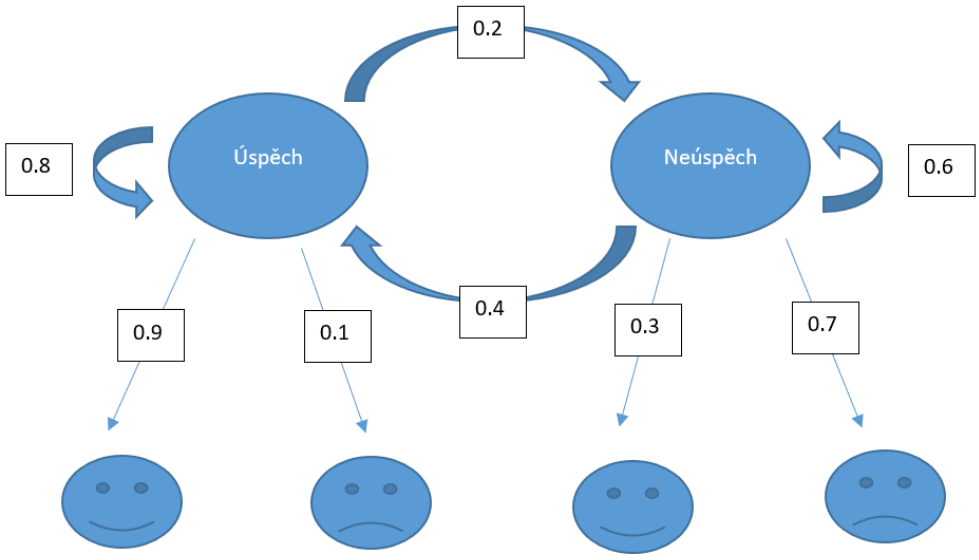
$$\mathbf{p}^{(0)} = \begin{array}{c} \begin{array}{cc} \text{Úspěch} & \text{Neúspěch} \end{array} \\ \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \end{array}.$$

  – Matice pravděpodobností přechodu $\mathbf{P}$

$$\mathbf{P} = \begin{array}{c} \\ \text{Úspěch} \\ \text{Neúspěch} \end{array} \begin{array}{c} \begin{array}{cc} \text{Úspěch} & \text{Neúspěch} \end{array} \\ \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{pmatrix} \end{array}.$$

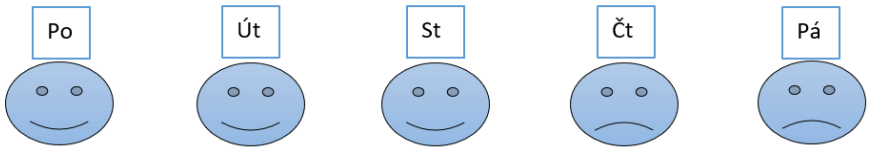
  – Matice pravděpodobností výstupu – $\mathbf{B}$

$$\mathbf{B} = \begin{array}{c} \\ \text{Úspěch} \\ \text{Neúspěch} \end{array} \begin{array}{c} \begin{array}{cc} \text{Veselý} & \text{Smutný} \end{array} \\ \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix} \end{array}.$$



Obrázek 2.4: Skrytý Markovův řetězec.

– Posloupnost pozorování



Obrázek 2.5: Studentova nálada v průběhu týdne.

Při takto zadaných parametrech můžeme přistoupit k jednotlivým krokům algoritmu.

Inicializace:

V naší posloupnosti pozorování je na prvním místě, tedy v pondělí, usměvaný smajlík. To nám značí, že rodiče vyhodnotili, že student měl v pondělí dobrou náladu.

$\alpha_0(1) = p_1^{(0)} b_1(1) = 0.5 \cdot 0.9 = 0.45$

$\alpha_0(1)$ je násobek počáteční pravděpodobnosti pro "úspěch"a výstupní pravděpodobnosti "veselé nálady", když zkouška dopadla úspěšně.

$\alpha_0(2) = p_1^{(0)} b_2(1) = 0.5 \cdot 0.3 = 0.15$

$\alpha_0(2)$ je pak násobek počáteční pravděpodobnosti pro "neúspěch"a výstupní pravděpodobnosti "veselé nálady", když zkouška dopadla neúspěšně.

Indukce:

Nyní pokračujeme druhým dnem, což je úterý, kdy byl student také veselý. Jak již víme dobrou náladu mohl mít v případě, že zkouška byla úspěšná, ale také když byla zkouška neúspěšná.

Tomu, že druhá zkouška dopadla dobře mohly předcházet dvě situace:

(a) První zkouška byla úspěšná s pravděpodobností $\alpha_0(1) \cdot p_{11}$.

(b) První zkouška byla neúspěšná s pravděpodobností $\alpha_0(2) \cdot p_{21}$.

Potom:

$\alpha_1(1) = [\alpha_0(1)\cdot p_{11}+\alpha_0(2)\cdot p_{21}]\cdot b_1(1) = [0.45\cdot 0.8+0.15\cdot 0.4]\cdot 0.9 = 0.378$

Také neúspěšnému výsledku druhého dne mohly předcházet dvě situace:

(a) První zkouška byla úspěšná s pravděpodobností $\alpha_0(1) \cdot p_{12}$.

(b) První zkouška byla neúspěšná s pravděpodobností $\alpha_0(2) \cdot p_{22}$.

Potom:

$\alpha_1(2) = [\alpha_0(1)\cdot p_{12}+\alpha_0(2)\cdot p_{22}]\cdot b_2(1) = [0.45\cdot 0.2+0.15\cdot 0.6]\cdot 0.3 = 0.54$

V podobném smyslu můžeme pokračovat dalšími dny – středa, čtvrtek a pátek.

Středa:

$\alpha_2(1) = [\alpha_1(1) \cdot p_{11} + \alpha_1(2) \cdot p_{21}] \cdot b_1(1) = [0.378 \cdot 0.8 + 0.54 \cdot 0.4] \cdot 0.9 = 0.2916$

$\alpha_2(2) = [\alpha_1(1) \cdot p_{12} + \alpha_1(2) \cdot p_{22}] \cdot b_2(1) = [0.378 \cdot 0.2 + 0.54 \cdot 0.6] \cdot 0.3 = 0.0324$

Čtvrtek:

$\alpha_3(1) = [\alpha_2(1)\cdot p_{11}+\alpha_2(2)\cdot p_{21}]\cdot b_1(2) = [0.2916\cdot 0.8+0.0324\cdot 0.4]\cdot 0.1 = 0.024624$

$\alpha_3(2) = [\alpha_2(1)\cdot p_{12}+\alpha_2(2)\cdot p_{22}]\cdot b_2(2) = [0.2916\cdot 0.2+0.0324\cdot 0.6]\cdot 0.7 = 0.054432$

Pátek:

$\alpha_4(1) = [\alpha_3(1) \cdot p_{11} + \alpha_3(2) \cdot p_{21}] \cdot b_1(2) = [0.024624 \cdot 0.8 + 0.054432 \cdot 0.4] \cdot 0.1 = 0.0041472$

$\alpha_4(2) = [\alpha_3(1) \cdot p_{12} + \alpha_3(2) \cdot p_{22}] \cdot b_2(2) = [0.024624 \cdot 0.2 + 0.054432 \cdot 0.6] \cdot 0.7 = 0.0263088$

Ukončení

Na závěr spočítáme výslednou pravděpodobnost modelu:

$P(\mathbf{O}|\lambda) = \alpha_4(1) + \alpha_4(2) = 0.0041472 + 0.0263088 \doteq 0.030456$

Tímto jsme zjistili, že pravděpodobnost posloupnosti studentovy nálady je přibližně 0.030456 za platnosti modelu $\lambda(\mathbf{p^{(0)}}, \mathbf{P}, \mathbf{B})$.

Příklad byl spočítán pomocí softwaru R, kde uvažujeme dále i více scénářů.

Zde si uvedeme pouze, jak se měnily výsledky na základě pohybu hodnot parametrů $\mathbf{p^{(0)}}$, $\mathbf{P}$ a $\mathbf{B}$. V následujícím textu je popsáno, jak se měnily parametry a možnosti jsou vyobrazeny pomocí grafů. Při prvotním zadání byl výsledek přibližně 0.030456, který je v grafech vyobrazen pomocí černé čárkované přímky.

- Scénář 2

  V dalším scénáři jsme zafixovali všechny hodnoty a hýbali jsme pouze s pravděpodobností přechodu p_{11} (pravděpodobnost přechodu, kdy po úspěšné zkoušce druhý den následuje také úspěšná zkouška) a tedy i s pravděpodobností p_{12} (pravděpodobnost přechodu, kdy po úspěšné zkoušce následuje neúspěšná zkouška). (Obrázek 2.6)

  Zjistili jsme, že pravděpodobnost posloupnosti pozorování $\mathbf{O}$ je nejvyšší při hodnotě pravděpodobnosti p_{11} v rozmezí 0.4 až 0.6. Přesněji při hodnotě $p_{11} = 0.55$ dosahuje pravděpodobnost $P(\mathbf{O}|\lambda)$ svého maxima.

  Při úvodním zadání ($p_{11} = 0.8$) byla pravděpodobnost $P(\mathbf{O}|\lambda) = 0.030456$, této hodnotě je rovna také, pokud p_{11} je v rozmezí 0.13 až 0.14. Nad touto úrovní je $P(\mathbf{O}|\lambda)$, pokud hodnota p_{11} je mezi 0.15 až 0.97. $P(\mathbf{O}|\lambda)$ je nižší než při úvodním zadání v případě, kdy pravděpodobnost p_{11} je rovna buď hodnotě v rozmezí od 0 po 0.12, nebo v rozmezí od 0.81 po 1.

- Scénář 3

  Nyní budeme měnit hodnotu pravděpodobnoti p_{21} (pravděpodobnost přechodu z neúspěšné zkoušky na úspěšnou) a zároveň tedy také p_{22} (pravděpodobnost dvou po sobě následujících neúspěšných zkoušek). Všechny ostatní pravděpodobnosti necháme na původních hodnotách. (Obrázek 2.7)

  S rostoucí hodnotou p_{21} pravděpodobnost pozorování **O** klesá. Pravděpodobnost $P(\mathbf{O}|\lambda)$ je nejvyšší při nulové pravděpodobnosti p_{21}.

  Pokud se podíváme na graf scénáře 3, vidíme, že úvodní zadání $p_{21} = 0.4$ rozděluje výsledky $P(\mathbf{O}|\lambda)$ téměř na půl. Při zadání první poloviny hodnot p_{21} (0 až 0.39) je pravděpodobnost $P(\mathbf{O}|\lambda)$ vyšší. Naopak pokud je p_{21} v rozmezí 0.41 až 1, pak je výsledek $P(\mathbf{O}|\lambda)$ nižší než pokud $p_{21} = 0.4$.

- Scénář 4

  Následují scénář bude založen na pohybu pravděpodobnosti $b_1(1)$ (pravděpodobnost výstupu veselé nálady při úspěšné zkoušce) a současně pravděpodobnosti $b_1(2)$ (pravděpodobnost výstupu smutné nálady při úspěšné zkoušce). (Obrázek 2.8)

  S rostoucí hodnotou pravděpodobnosti $b_1(1)$ roste také pravděpodobnost posloupnosti pozorování **O**. Pravděpodobnost **O** roste až do doby, kdy pravděpodobnost $b_1(1)$ dosáhne hodnoty 0.8. Na této hodnotě dosahuje pravděpodobnost $P(\mathbf{O}|\lambda)$ svého vrcholu, poté začne mírně klesat.

  Stejného výsledku pravděpodobnosti $P(\mathbf{O}|\lambda)$ jako při počátečním zadání sledované hodnoty $b_1(1) = 0.9$ dosáhneme, pokud je $b_1(1)$ v rozmezí od 0.65 až 0.70.

- Scénář 5

  V pátém scénáři ponecháme všechny pravděpodobnosti, pouze budeme hýbat s pravděpodobností $b_2(1)$ (pravděpodobnost výstupu veselé nálady při neúspěšné zkoušce) a současně s pravděpodobností $b_2(2)$ (pravděpodobnost výstupu smutné nálady při neúspěšné zkoušce). (Obrázek 2.9)

  Pravděpodobnost posloupnosti pozorování $\mathbf{O}$ je tím větší, čím je menší pravděpodobnost $b_2(1)$. Pravděpodobnost $P(\mathbf{O}|\lambda)$ je nejvyšší při nulové hodnotě pravděpodobnosti $b_2(1)$.

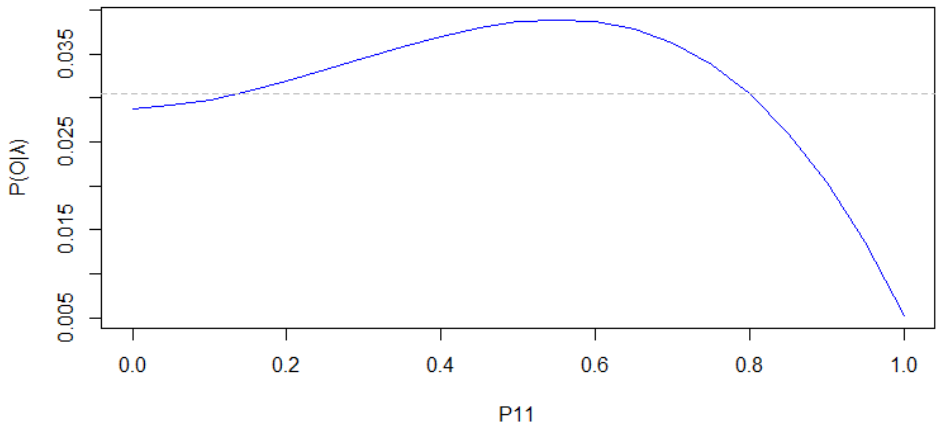
  I tento scénář porovnáme s výsledkem z úvodního zadání sledované hodnoty $b_2(1)$. Vyšší pravděpodobnosti $P(\mathbf{O}|\lambda)$ je dosaženo pouze v případě, kdy $b_2(1)$ je rovna hodnotě od 0 po 0.29. Naopak v případě, kdy $b_2(1)$ je mezi 0.31 a 1, pak je pravděpodobnost posloupnosti pozorování $P(\mathbf{O}|\lambda)$ nižší než při prvotním zadání $b_2(1)$.

V následující tabulce 2.1 uvedeme nejvyšší hodnoty pravděpodobností $P(\mathbf{O}|\lambda)$, které jsme obdrželi z jednotlivých scénářů.

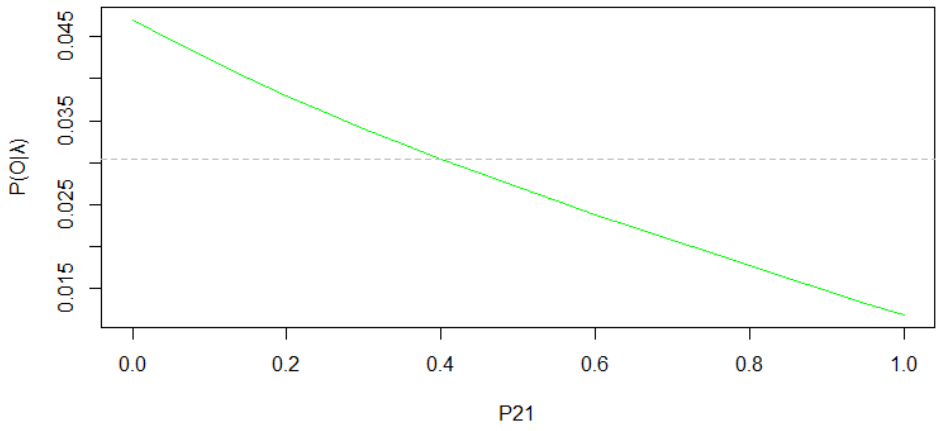
Tabulka 2.1: Srovnání scénářů.

Scénáře	$P(\mathbf{O}\|\lambda)$
Scénář 1	0.030456
Scénář 2	0.038915
Scénář 3	0.047077
Scénář 4	0.031300
Scénář 5	0.035085

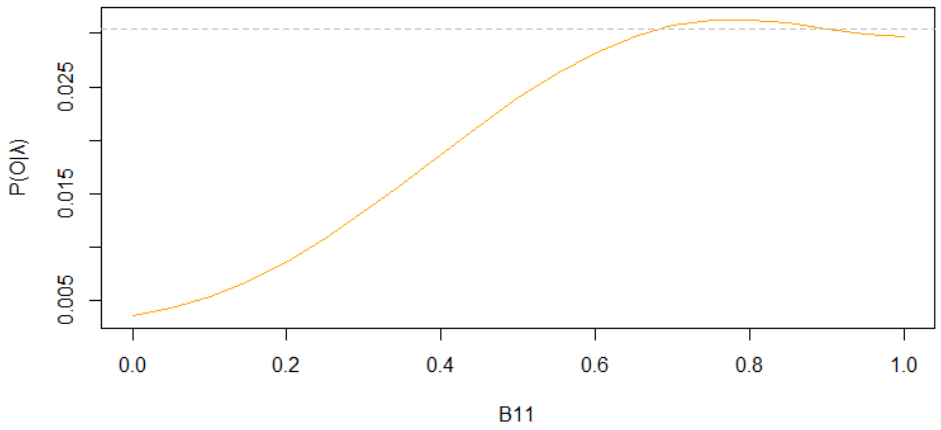
Celý příklad je dostupný k nahlédnutí v příloze.



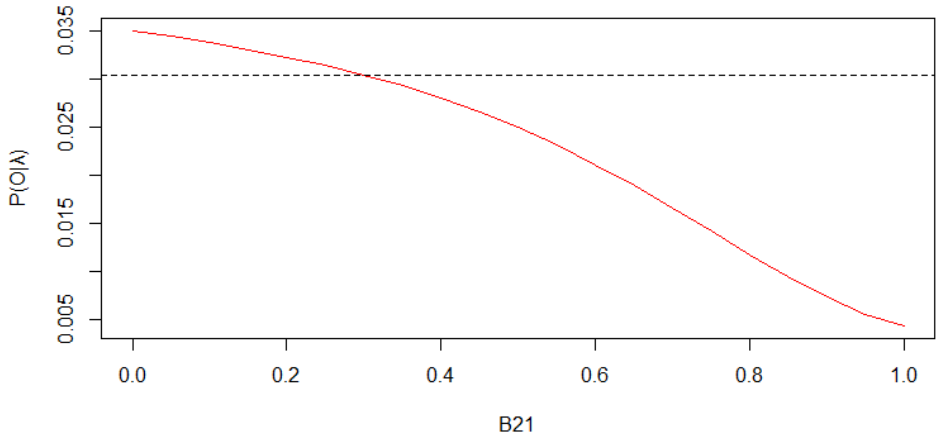
Obrázek 2.6: Scénář 2: Pravděpodobnost dvou po sobě následujících úspěšných zkoušek.



Obrázek 2.7: Scénář 3: Pravděpodobnost úspěšné zkoušky následující po neúspěšné zkoušce.



Obrázek 2.8: Scénář 4: Pravděpodobnost veselé nálady v den úspěšné zkoušky.



Obrázek 2.9: Scénář 5: Pravděpodobnost veselé nálady v den neúspěšné zkoušky.

2. **Zpětný algoritmus**

Jak již název napovídá, jedná se podobný algorimus jako je dopředný, avšak se zpětným krokem. Definujeme si zde veličinu

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, \ldots, O_T | X_t = i, \lambda). \tag{2.11}$$

Tato veličina nám říká pravděpodobnost posloupnosti pozorování $\mathbf{O} = \{O_{t+1}, O_{t+2}, \ldots, O_T\}$ od času $t+1$ po T, pokud jsme v čase t ve stavu i a platí daný model λ.

Stejně jako dopředný algorimus, tak i zpětný se skládá ze tří kroků:

(a) Inicializace

$$\beta_T(i) = 1, \quad 1 \leq i \leq N \tag{2.12}$$

V prvním kroce položíme $\beta_T(i)$ rovno 1 pro všechna i.

(b) Indukce

$$\begin{aligned} \beta_t(i) = &\sum_{j=1}^{N} p_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \\ &t = T-1, T-2, \ldots, 0, \quad 1 \leq i \leq N. \end{aligned} \tag{2.13}$$

Indukce představuje zásadní krok také ve zpětném algoritmu.

Při předpokladu, že jsme ve stavu i v čase t , kdy je zohledněna posloupnost pozorování $\{O_{t+1}, \ldots, O_T\}$, bereme v potaz veškeré posloupnosti skrytých stavů $\{X_{t+1}, X_{t+2}, \ldots, X_T\}$, kterými měl skrytý řetězec možnost projít. Následně chování od času $t+1$ popisujeme prostřednictvím $\beta_{t+1}(j)$. Člen p_{ij} bere v potaz pravděpodobnosti přechodu ze stavu i do stavu $j,\ j = 1, ..., N$. Prvek $b_j(O_{t+1})$ zohledňuje pravděpodobnost výstupu pozorování O_{t+1} ve stavech $j,\ j,\ j = 1, ..., N$.

Ten samý postup postupně aplikujeme ve všech stavech i a v každém časovém okamžiku t.

(c) Ukončení

$$P(\mathbf{O}|\lambda) = \sum_{i=1}^{N} p_i^{(0)} b_i(O_0)\beta_0(i). \tag{2.14}$$

Pomocí počátečního rozdělení a pravděpodobnosti vygenerování pozorování O_0 jsou rozděleny váhy pro sumu hodnot $\beta_0(i)$ při ukončovacím procesu. Ukončovací krok je tedy proveden pomocí vážené sumy veškerých $\beta_0(i)$.

Podobně jako tomu bylo u dopředného algoritmu, také zpětný algoritmus nastíníme pomocí příkladu se studentem.

Příklad 2.4.2. *Zadání příkladu bude totožné se zadáním příkladu 2.4.1, rozdíl bude v použitém algoritmu a sice zde využijeme zpětný algorimus.*

- *Počáteční rozdělení pravděpodobností* $\boldsymbol{p}^{(0)}$

$$\mathbf{p}^{(0)} = \begin{array}{c} \\ \end{array} \begin{array}{cc} \textit{Úspěch} & \textit{Neúspěch} \\ \end{array} \begin{pmatrix} 0.5 & 0.5 \end{pmatrix}$$

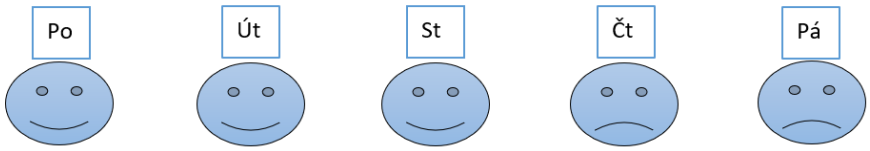
- *Matice pravděpodobností přechodu* $\boldsymbol{P}$

$$\mathbf{P} = \begin{array}{c} \\ \textit{Úspěch} \\ \textit{Neúspěch} \end{array} \begin{array}{cc} \textit{Úspěch} & \textit{Neúspěch} \\ \end{array} \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{pmatrix}$$

- *Matice pravděpodobností výstupu* $\boldsymbol{B}$

$$\mathbf{B} = \begin{array}{c} \\ \textit{Úspěch} \\ \textit{Neúspěch} \end{array} \begin{array}{cc} \textit{Veselý} & \textit{Smutný} \\ \end{array} \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix}$$

- *Posloupnost pozorování*



Obrázek 2.10: Studentova nálada v průběhu týdne.

Inicializace

$\beta_4(1) = 1$

$\beta_4(2) = 1$

Indukce

V naší posloupnosti máme, že poslední den – pátek byl student smutný. Den před tím – ve čtvrtek, mohl student zkoušku zvládnout úspěšně nebo neúspěšně. Po situaci, kdy zkouška byla ve čtvrtek úspěšná mohou v pátek následovat dva scénáře:

(a) Páteční zkouška dopadla úspěšně s pravděpodobností $p_{11} \cdot b_1(2) \cdot \beta_4(1)$.

(b) Páteční zkouška nedopadla úspěšně s pravděpodobností $p_{12} \cdot b_2(2) \cdot \beta_4(2)$.

Potom:

$$\beta_3(1) = p_{11} \cdot b_1(2) \cdot \beta_4(1) + p_{12} \cdot b_2(2) \cdot \beta_4(2) = 0.8 \cdot 0.1 \cdot 1 + 0.2 \cdot 0.7 \cdot 1 = 0.22$$

Podobně po scénáři, ve kterém čtvrteční zkouška nebyla úspěšná, mohou v pátek nastat dvě situace:

(a) Páteční zkouška dopadla úspěšně s pravděpodobností $p_{21} \cdot b_1(2) \cdot \beta_4(1)$.

(b) Páteční zkouška nedopadla úspěšně s pravděpodobností $p_{22} \cdot b_2(2) \cdot \beta_4(2)$.

Potom:

$$\beta_3(2) = p_{21} \cdot b_1(2) \cdot \beta_4(1) + p_{22} \cdot b_2(2) \cdot \beta_4(2) = 0.4 \cdot 0.1 \cdot 1 + 0.6 \cdot 0.7 \cdot 1 = 0.46$$

Obdobné výpočty budeme aplikovat také na další dny.

$$\beta_2(1) = p_{11} \cdot b_1(2) \cdot \beta_3(1) + p_{12} \cdot b_2(2) \cdot \beta_3(2) = 0.8 \cdot 0.22 \cdot 1 + 0.2 \cdot 0.7 \cdot 0.46 = 0.082$$
$$\beta_2(2) = p_{21} \cdot b_1(2) \cdot \beta_3(1) + p_{22} \cdot b_2(2) \cdot \beta_3(2) = 0.4 \cdot 0.1 \cdot 0.22 + 0.6 \cdot 0.7 \cdot 0.46 = 0.202$$

$$\beta_1(1) = p_{11} \cdot b_1(1) \cdot \beta_2(1) + p_{12} \cdot b_2(1) \cdot \beta_2(2) = 0.8 \cdot 0.9 \cdot 0.082 + 0.2 \cdot 0.3 \cdot 0.202 =$$
$$= 0.07116$$
$$\beta_1(2) = p_{21} \cdot b_1(1) \cdot \beta_2(1) + p_{22} \cdot b_2(1) \cdot \beta_2(2) = 0.4 \cdot 0.9 \cdot 0.082 + 0.6 \cdot 0.3 \cdot 0.202 =$$
$$= 0.06588$$

$$\beta_0(1) = p_{11} \cdot b_1(1) \cdot \beta_1(1) + p_{12} \cdot b_2(1) \cdot \beta_1(2) = 0.8 \cdot 0.9 \cdot 0.07116 + 0.2 \cdot 0.3 \cdot 0.06588 = = 0.055188$$
$$\beta_0(2) = p_{21} \cdot b_1(1) \cdot \beta_1(1) + p_{22} \cdot b_2(1) \cdot \beta_1(2) = 0.4 \cdot 0.9 \cdot 0.07116 + 0.6 \cdot 0.3 \cdot 0.06588 = = 0.037476$$

Ukončení

$$P(O|\lambda) = p_1^{(0)} \cdot b_1(1) \cdot \beta_0(1) + p_2^{(0)} \cdot b_2(1) \cdot \beta_0(2) =$$
$$= 0.5 \cdot 0.9 \cdot 0.055188 + 0.5 \cdot 0.3 \cdot 0.037476 \doteq 0.030456$$

Na základě námi zadaného modelu $\lambda(\mathbf{p^{(0)}}, \mathbf{P}, \mathbf{B})$, je pravděpodobnost posloupnosti studentovy nálady cca 0.030456.

Celý příklad naleznete v příloze. K výpočtu byl využit software R [11].

∘

2.4.2. Dekódování

V druhé úloze dekódujeme, snažíme se na základě posloupnosti **O** najít skrytou část modelu. Hledáme posloupnost skrytých stavů, při které projdeme skrytými stavy s největší pravděpodobností za předpokladu, že známe **O** a λ. Zde je potřeba zavést optimalizační kritérium, jelikož ne vždy lze nalézt "správnou"posloupnost stavů.

Tento problém můžeme ilustrovat na pokračování předchozího příkladu.

Příklad 2.4.3. *Dekódování v našem příkladě je to, co řeší rodiče. Když má student v týdnu, kdy má zkoušky, postupně různou náladu, tak rodiče dekódují výsledky zkoušek. V pondělí je veselý, v úterý a ve středu je také veselý a zbytek týdne (čtvrtek, pátek) je smutný. Rodiče na základě těchto informací usoudí, že zkoušky byly od pondělí do středy úspěšné a ve čtvrtek a pátek neúspěšné.*

∘

Úloha dekódování se dá vyřešit pomocí Viterbiho algoritmu.

Viterbiho algoritmus

V algoritmech pro evaluaci jsme zaváděli nové proměnné, i zde zavedeme nové proměnné δ a ψ. První zmíněná proměnná δ slouží k nalezení nejlepší posloupnosti skrytých stavů $\mathbf{X} = (X_1, X_2, \ldots, X_t)$ při dané posloupnosti pozorování $\mathbf{O} = (O_1, O_2, \ldots, O_t)$. Při pozorování posloupnosti **O** značí δ největší pravděpodobnost, při které lze ve stavu i v čase t určit posloupnost skrytých stavů **X**.

$$\delta_t(i) = \max_{X_0,\ldots,X_{t-1}} P(X_0 \cdots X_{t-1}, X_t = i, O_0, \ldots, O_t|\lambda), \tag{2.15}$$

Druhá uvedená proměnná ψ představuje pomocnou proměnnou, pomocí které zpětně hledáme skryté stavy.

Viterbiho algorimus je rozdělen do čtyř kroků:

1. Inicializace

$$\delta_0(i) = p_i^{(0)} b_i(O_0), \quad 1 \leq i \leq N, \tag{2.16}$$

$$\psi_0(i) = 0, \quad \forall i. \tag{2.17}$$

2. Rekurze

$$\delta_t(j) = \max_{1 \leq i \leq N} [\delta_{t-1}(i) p_{ij}] b_j(O_t), \quad 1 \leq t \leq T, \quad 1 \leq j \leq N, \tag{2.18}$$

$$\psi_t(j) = \arg \max_{1 \leq i \leq N} [\delta_{t-1}(i) p_{ij}], \quad 1 \leq t \leq T, \quad 1 \leq j \leq N. \tag{2.19}$$

3. Ukončení

$$P^* = \max_{1 \leq i \leq N} [\delta_T(i)], \tag{2.20}$$

$$X_T^* = \arg \max_{1 \leq i \leq N} [\delta_T(i)]. \tag{2.21}$$

4. Zpětné hledání cesty

$$X_t^* = \psi_{t+1}(X_{t+1}^*), \quad t = T-1, T-2, \ldots, 1. \tag{2.22}$$

   Proměnná ψ ve čtvrté fázi stanoví posloupnost skrytých stavů X^*, která je nejpravděpodobnější.

Z uvedeného postupu Viterbiho algoritmu můžeme pozorovat určitou anologii s dopředným algoritmem, jenž jsme si uvedli u úlohy evaluace. Rozdíl, kterého si můžeme na první pohled všimnout, je v operaci maximum (u dopředného algoritmu byl využíván součet). Maximum se zde užívá z důvodu, že Viterbiho algoritmus hledá jednu jedinou, avšak nejlepší cestu.

Příklad 2.4.4. *Také u Viterbiho algoritmu budeme pokračovat v našem příkladu. Zadání je stále stejné. Hledáme nejpravděpodobnější průběh zkouškového týdne.*

- *Počáteční rozdělení pravděpodobností* – $\boldsymbol{p}^{(0)}$

$$\mathbf{p}^{(0)} = \begin{array}{c} \begin{array}{cc} \textit{Úspěch} & \textit{Neúspěch} \end{array} \\ \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \end{array}$$

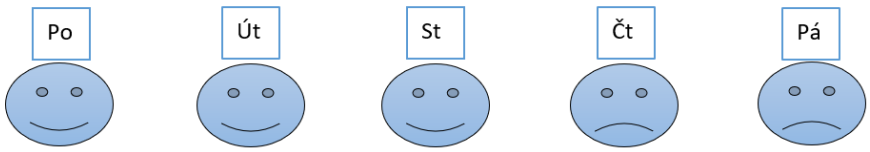
- *Matice pravděpodobností přechodu* – $\boldsymbol{P}$

$$\mathbf{P} = \begin{array}{c} \\ \textit{Úspěch} \\ \textit{Neúspěch} \end{array} \begin{array}{c} \begin{array}{cc} \textit{Úspěch} & \textit{Neúspěch} \end{array} \\ \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{pmatrix} \end{array}$$

- *Matice pravděpodobností výstupu* – $\boldsymbol{B}$

$$\mathbf{B} = \begin{array}{c} \\ \textit{Úspěch} \\ \textit{Neúspěch} \end{array} \begin{array}{c} \begin{array}{cc} \textit{Veselý} & \textit{Smutný} \end{array} \\ \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix} \end{array}$$

- *Posloupnost pozorování*



Obrázek 2.11: Studentova nálada v průběhu týdne.

Inicializace

V tomto kroku se nacházíme v pondělí.

$\gamma_0(1) = p_1^{(0)} \cdot b_1(1) = 0.5 \cdot 0.9 = 0.45$

$\gamma_0(2) = p_2^{(0)} \cdot b_2(1) = 0.5 \cdot 0.3 = 0.15$

$\psi_0(1) = 0$

$\psi_0(2) = 0$

Rekurze

V druhém kroku se ocitáme v úterý. V úterý je student podle naší posloupnosti veselý. Dobrou náladu mohl mít ve dvou případech buď zkoušku udělal úspěšně, nebo neúspěšně (avšak stále byl veselý).

Pokud zkoušku zvládl úspěšně, pak tomu v pondělí předcházel jeden ze dvou scénářů:

1. Student zkoušku v pondělí zvládl úspěšně s pravděpodobností $\gamma_0(1) \cdot p_{11}$.
2. Student zkoušku v pondělí nezvládl úspěšně s pravděpodobností $\gamma_0(2) \cdot p_{21}$.

Potom:
$\gamma_1(1) = \max(\gamma_0(1) \cdot p_{11}; \gamma_0(2) \cdot p_{21}) \cdot b_1(1) =$
$= \max(0.45 \cdot 0.8; 0.15 \cdot 0.4) \cdot 0.9 = \max(0.324; 0.054) = 0.324$
$\psi_1(1) = \arg\max(\gamma_0(1) \cdot p_{11}; \gamma_0(2) \cdot p_{21}) = \arg\max(0.36; 0.06) = 1$
Kde symbol 1 značí úspěch.

Jak jsme ale řekli, také mohl být veselý navzdory tomu, že zkoušku nezvládl úspěšně a tomu mohly předcházet dvě situace:

1. Student zkoušku v pondělí zvládl úspěšně s pravděpodobností $\gamma_0(1) \cdot p_{12}$.
2. Student zkoušku v pondělí nezvládl úspěšně s pravděpodobností $\gamma_0(2) \cdot p_{22}$.

$\gamma_1(2) = \max(\gamma_0(1) \cdot p_{12}; \gamma_0(2) \cdot p_{22}) \cdot b_2(1) =$
$= \max(0.45 \cdot 0.2; 0.15 \cdot 0.6) \cdot 0.3 = \max(0.027; 0.027) = 0.027$
$\psi_1(2) = \arg\max(\gamma_0(1) \cdot p_{12}; \gamma_0(2) \cdot p_{22}) = \arg\max(0.09; 0.09) = \{1, 2\}$
V tomto scénáři vidíme, že obě situace jsou stejně pravděpodobné.

Potom v dalších dnech postupujeme následovně:

Středa:

$\gamma_2(1) = \max(\gamma_1(1) \cdot p_{11}; \gamma_1(2) \cdot p_{21}) \cdot b_1(1) =$

$= \max(0.324 \cdot 0.8; 0.027 \cdot 0.4) \cdot 0.9 = \max(0.23328; 0.00972) = 0.23328$

$\psi_2(1) = \arg\max(\gamma_1(1) \cdot p_{11}; \gamma_1(2) \cdot p_{21}) = \arg\max(0.2592; 0.0108) = 1$

$\gamma_2(2) = \max(\gamma_1(1) \cdot p_{12}; \gamma_1(2) \cdot p_{22}) \cdot b_2(1) =$

$= \max(0.324 \cdot 0.2; 0.027 \cdot 0.6) \cdot 0.3 = \max(0.01944; 0.00486) = 0.01944$

$\psi_2(2) = \arg\max(\gamma_1(1) \cdot p_{12}; \gamma_1(2) \cdot p_{22}) = \arg\max(0.0648; 0.0162) = 1$

Čtvrtek:

$\gamma_3(1) = \max(\gamma_2(1) \cdot p_{11}; \gamma_2(2) \cdot p_{21}) \cdot b_1(1) =$

$= \max(0.23328 \cdot 0.8; 0.01944 \cdot 0.4) \cdot 0.1 = \max(0.0186624; 0.0007776) = 0.0186624$

$\psi_3(1) = \arg\max(\gamma_2(1) \cdot p_{11}; \gamma_2(2) \cdot p_{21}) = \arg\max(0.186624; 0.007776) = 1$

$\gamma_3(2) = \max(\gamma_2(1) \cdot p_{12}; \gamma_2(2) \cdot p_{22}) \cdot b_2(1) =$

$= \max(0.23328 \cdot 0.2; 0.01944 \cdot 0.6) \cdot 0.7 = \max(0.0326592; 0.0081648) = 0.03265924$

$\psi_3(2) = \arg\max(\gamma_2(1) \cdot p_{12}; \gamma_2(2) \cdot p_{22}) = \arg\max(0.046656; 0.011664) = 1$

Pátek:

$\gamma_4(1) = \max(\gamma_3(1) \cdot p_{11}; \gamma_3(2) \cdot p_{21}) \cdot b_1(1) =$

$= \max(0.0186624 \cdot 0.8; 0.0326592 \cdot 0.4) \cdot 0.1 = \max(0.001492992; 0.001306368) =$

$= 0.001492992$

$\psi_4(1) = \arg\max(\gamma_3(1) \cdot p_{11}; \gamma_3(2) \cdot p_{21}) = \arg\max(0.01492992; 0.01306368) = 1$

$\gamma_4(2) = \max(\gamma_3(1) \cdot p_{12}; \gamma_3(2) \cdot p_{22}) \cdot b_2(1) =$

$= \max(0.0186624 \cdot 0.2; 0.0326592 \cdot 0.6) \cdot 0.7 = \max(0.002612736; 0.01371686) =$

$= 0.01371686$

$\psi_4(2) = \arg\max(\gamma_3(1) \cdot p_{12}; \gamma_3(2) \cdot p_{22}) = \arg\max(0.00373248; 0.01959552) = 2$

Ukončení

$P^* = \max\{\gamma_4(1); \gamma_4(2)\} = 0.013716864$

$X_4^* = \arg\max\{\gamma_4(1); \gamma_4(2)\} = 2$

Zpětné hledání cesty

$X_3^* = \psi_4(2) = 2$

$X_2^* = \psi_3(2) = 1$

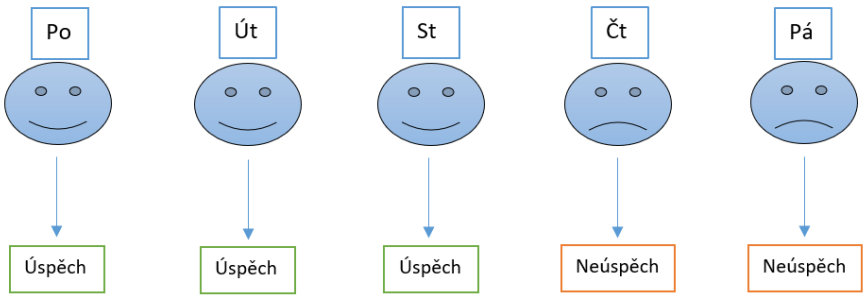
$X_1^* = \psi_2(1) = 1$

$X_0^* = \psi_1(1) = 1$

Tím pádem výsledná posloupnost skrytých stavů vypadá následovně:

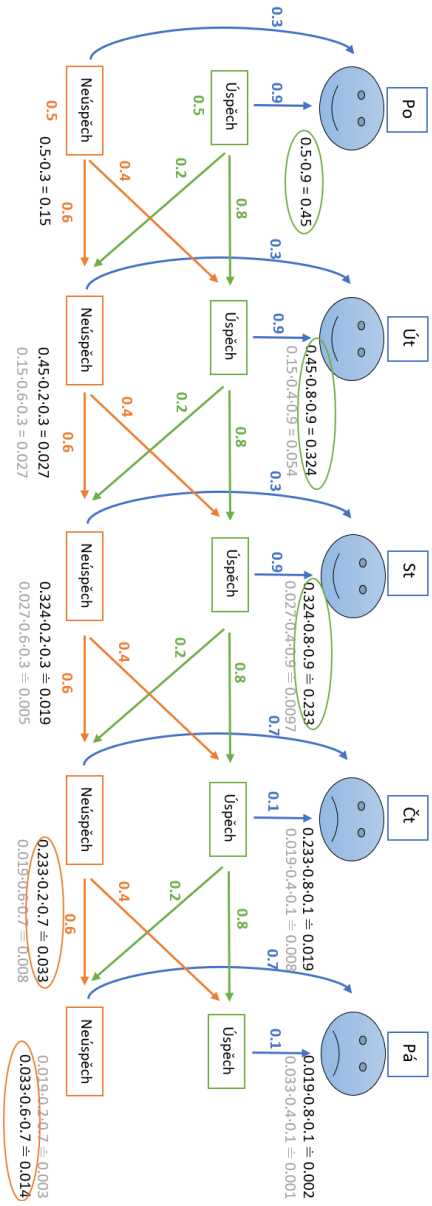
$\mathbf{X}^* = \{1, 1, 1, 2, 2\}$

Jinými slovy: "Na základě vývoje studentovy nálady, je nejpravděpodobnější, že student složil v pondělí zkoušku úspěšně, v úterý úspěšně, ve středu také úspěšně, ve čtvrtek neúspěšně a v pátek také neúspěšně".



Obrázek 2.12: Posloupnost skrytých stavů v průběhu týdne.

Pro ilustraci si pomocí obrázku znázorníme celý postup Viterbiho algoritmu.



Obrázek 2.13: Postup Viterbiho metody.

Vysvětlení obrázku 2.13:

- Symboly usměvavých a zamračených "smajlíků" odpovídají posloupnosti pozorování, tedy veselé či smutné náladě studenta v jednotlivých dnech.
- Zelené a oranžové štítky s nápisem *Úspěch* či *Neúspěch* symbolizují skryté stavy – výsledky studentovy zkoušky.
- Modré šipky nesou informaci výstupních pravděpodobností $b_j(k)$ veselé či smutné nálady studenta při úspěchu nebo neúspěchu.
- Zelené šipky znázorňují pravděpodobnosti přechodu p_{ij} z úspěchu na úspěch či neúspěch.
- Oranžové šipky naopak znázorňují pravděpodobnosti přechodu p_{ij} z neúspěchu na úspěch či neúspěch.
- První část, tedy část týkající se pondělí, odpovídá kroku inicializace. Na základě tohoto kroku je vybrán pravděpodobnější skrytý stav *Úspěch* (pravděpodobnost zakroužkována pomocí zelené barvy).
- V každém dalším dnu je na výběr z obou možností skrytých stavů (*Úspěch* či *Neúspěch*), kdy na základě kroku rekurze je vybrán pravděpodobnější skrytý stav.
  - U každého dne jsou čtyři výpočty pravděpodobností. Tedy u každé možnosti *Úspěchu* nebo *Neúspěchu* jsou uvedeny dva výpočty pravděpodobností. V prvním výpočtu se předpokládá, že předcházející zkouška byla úspěšná. U druhého předpokládáme, že zkouška z předchozího dne byla neúspěšná.
  - Dále vždy jedna z možností těchto pravděpodobností je vyznačena černým písmem a druhá šedým. Pravděpodobnost psaná černým písmem má vyšší hodnotu, nežli pravděpodobnost s šedým písmem.

– U jednotlivých dnů nám pak zbydou dvě pravděpodobnosti (černé písmo). Z těchto pravděpodobností je vybrána ta větší a je zakroužkována buď oranžově, nebo zeleně. Zelená barva značí větší pravděpodobnost *úspěchu* v daný den a oranžová barva naopak označuje *neúspěch* jako více pravděpodobný.

– Podle zakroužkouvaných pravděpodobností pak následně vidíme nejpravděpodobnější posloupnost skrytých stavů.

  Pondělí = úspěch → úterý = úspěch → středa = úspěch → čtvrtek = neúspěch → pátek = neúspěch.

∘

2.4.3. Učení

Pod pojmem učení se schovává proces odhadu parametrů. Pokud známe pozorování, pak se snažíme parametry modelu λ odhadnout co nejlépe. V případě, že víme úvodní hodnoty parametrů $\mathbf{P}, \mathbf{B}, \mathbf{p}^{(0)}$, chceme provést takovou změnu parametrů modelu, která povede k maximalizaci pravděpodobnosti generování posloupnosti. V druhém případě, kdy nemáme žádné informace o parametrech, pak se snažíme o co nejlepší odhad těchto parametrů.

I zde vysvětlení přiblížíme pomocí pokračování našeho příkladu se studentem.

Příklad 2.4.5. *Rodiče sice studenta znají a na základě jeho nálady odhadují výsledky zkoušky. Může se ale stát, že student přestane být pilný v takové míře jako doposud a výsledky zkoušek budou jiné než rodiče předpokládají. Učení by tedy v tomto příkladu znamenalo, že by rodiče museli přehodnotit názor na studenta. Řeknou si, že školu začal flákat, a tudíž není tolik úspěšný jako bývával.*

○

V úloze učení se využívá Baum-Welch algoritmus.

Baum-Welch algoritmus

Algoritmus se skládá z několik fází:

1. Inicializace

   V úvodní části se zvolí počáteční hodnoty pro parametry, které chceme odhadnout. Na tomto místě jsou tedy zadány matice $\mathbf{P}$, $\mathbf{B}$ a vektor $\mathbf{p}^{(0)}$ buďto na základě získaných informacích o těchto parametrech, nebo pokud nemáme žádné informace, snažíme se hodnoty $\mathbf{P}, \mathbf{B}, \mathbf{p}^{(0)}$ rovnoměrně rozložit. A to tak, aby jednotlivé prvky v maticích $\mathbf{P}$, $\mathbf{B}$ a vektoru $\mathbf{p}^{(0)}$ byly stejně pravděpodobné.

2. Dopředná fáze

   Tato fáze je také označována jako filtrování. V této části se využívá dopředný algoritmus, který jsme si uvedli dříve. Pomocí dopředného algoritmu se vypočítá $\alpha(\cdot)$ (2.7) až po čas k, $0 \leq k \leq T$.

3. Zpětná fáze

   Použijeme zpětný algoritmus, který jsme měli v úloze evaluace. Probíhá zde výpočet $\beta(\cdot)$ (2.11) od času $k+1$, $0 \leq k \leq T$.

4. Fáze aktualizace

   Využívá se zde vypočtené α, β, určené přechodové a emisní pravděpodobnosti.

   V této fázi se objevují veličiny ξ a γ, které si musíme zadefinovat.

$$\xi_t(i,j) = P(X_t = i, X_{t+1} = j|\mathbf{O}, \lambda). \tag{2.23}$$

   Tato pravděpodobnost odpovídá tomu, že řetězec bude ve skrytém stavu i v čase t a ve skrytém stavu j v čase $t+1$, vzhledem k λ a posloupnosti $\mathbf{O}$. Díky veličinám α a β z dopředného a zpětného algoritmu, můžeme pravděpodobnost přepsat do této podoby

$$\xi_t(i,j) = \frac{P(X_t = i, X_{t+1} = j, \mathbf{O}|\lambda)}{P(\mathbf{O}|\lambda)} = \frac{\alpha_t(i)p_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(\mathbf{O}|\lambda)}, \tag{2.24}$$

$$\xi_t(i,j) = \frac{\alpha_t(i)p_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)p_{ij}b_j(O_{t+1})\beta_{t+1}(j)}. \tag{2.25}$$

   Zadefinujeme si také druhou veličinu γ. Určuje pravděpodobnost, že bude řetězec ve skrytém stavu i v čase t při posloupnosti $\mathbf{O}$

$$\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j). \tag{2.26}$$

Nyní můžeme přikročit k samotnému odhadování parametrů modelu λ.

Nové počáteční pravděpodobnosti určíme položením této rovnosti

$$\hat{p}_i^{(0)} = \gamma_0(i). \tag{2.27}$$

Odhadované pravděpodobnosti přechodu získáme pomocí

$$\hat{p}_{ij} = \frac{\sum_{t=0}^{T-1} \xi_t(i,j)}{\sum_{t=0}^{T-1} \gamma_t(i)}, \tag{2.28}$$

kde čitatel určuje, kolik bylo přechodů ze skrytého stavu i do skrytého stavu j. Úplným očekávaným počtem výskytů ve skrytém stavu i je pak suma ve jmenovateli.

Jaké bude nové rozdělení pravděpodobnosti výstupu, určíme následovně

$$\hat{b}_j(k) = \frac{\sum_{t=0|O_t=v_k}^{T} \gamma_t(j)}{\sum_{t=0}^{T} \gamma_t(j)}, \tag{2.29}$$

v čitateli se objevuje očekávaný počet přítomnosti ve skrytém stavu j, kdy se sčítají pouze takové časy, kde pozorujeme v_k. Jmenovatel sčítá očekávaný výskyt ve skrytém stavu j.

Postup Baum-Welch metody tedy začíná vlastním nadefinováním modelu $\lambda(\mathbf{P}, \mathbf{B}, \mathbf{p}^{(0)})$. Parametry modelu spolu s hodnotami α, β, γ a ξ využijeme v rovnicích 2.27, 2.28, 2.29. Z těchto rovnic získáme parametry nového modelu $\lambda'(\hat{\mathbf{P}}, \hat{\mathbf{B}}, \hat{\mathbf{p}}^{(0)})$.

Nyní mohou nastat dvě situace:

1. Původní model λ definuje kritický bod funkce pravděpodobnosti a tedy $\lambda' = \lambda$.

2. Původní model λ má menší pravděpodobnost než λ'. A to znamená, že posloupnost generována původním modelem je ohodnocena menší pravděpodobností.

Uvedený postup můžeme opakovat dokud nedosáhneme požadované přesnosti. Jako kritérium si můžeme stanovit, že proces budeme dělat do doby, kdy pravděpodobnost pozorování **O** nedocílí stanoveného limitu nebo pokud pravděpodobnost nebude podobná dvěma po sobě následujících opakování.

Baum-Welch algoritmus je speciálním případem EM algoritmu (Expectation-Maximization).

EM metoda je iterativní algorimus, ve kterém se opakují kroky E (z anglického Expectation - očekávaná hodnota) a M (z anglického Maximization - maximalizace). EM algoritmus se používá, pokud chceme získat maximálně věrohodné odhady z neúplných nebo skrytých proměnných.

Výpočet α a β tvoří krok E, s ohledem na pozorovaná data a parametry $\mathbf{P}$, $\mathbf{B}$, $\mathbf{p}^{(0)}$. Krokem M je pak fáze aktualizace. Je tomu tak, protože vzorce 2.27, 2.28, 2.29 jsou odvozeny, aby nejvíce vyhovovaly očekávaným skrytým stavům.

Na tomto místě přepočítáme pomocí Baum-Welch algoritmu parametry našeho příkladu, které jsme si určili na začátku.

Doposud byly příklady programovány v softwaru R [11] pomocí vlastního zpracování. Také v následujícím příkladu bude vyžito softwaru R, avšak nyní využijeme již zabudovanou knihovnu s názvem `HMM` [2].

Příklad 2.4.6. *Příklad navazuje na minulé příklady, tudíž počáteční parametry zanecháme stejné.*

- *Počáteční rozdělení pravděpodobností* $\boldsymbol{p}^{(0)}$

$$\mathbf{p}^{(0)} = \begin{pmatrix} \textit{Úspěch} & \textit{Neúspěch} \\ 0.5 & 0.5 \end{pmatrix}.$$

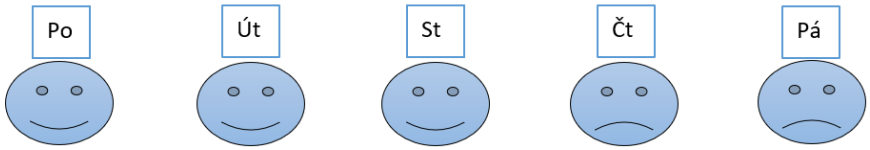
- *Matice pravděpodobností přechodu* $\boldsymbol{P}$

$$\mathbf{P} = \begin{array}{c} \\ \textit{Úspěch} \\ \textit{Neúspěch} \end{array} \begin{array}{c} \begin{array}{cc} \textit{Úspěch} & \textit{Neúspěch} \end{array} \\ \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{pmatrix} \end{array}.$$

- *Matice pravděpodobností výstupu* $\boldsymbol{B}$

$$\mathbf{B} = \begin{array}{c} \\ \textit{Úspěch} \\ \textit{Neúspěch} \end{array} \begin{array}{c} \begin{array}{cc} \textit{Veselý} & \textit{Smutný} \end{array} \\ \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix} \end{array}.$$

- *Posloupnost pozorování*



Obrázek 2.14: Studentova nálada v průběhu týdne.

Vysvětlení následujícího značení:

- Výstupy jsou označeny jako `$Symbols`.
- Počáteční pravděpodobnosti skrytých stavů jsou pojmenovány jako `$startprobs`.
- Pravděpodobnosti přechodu nesou označení `$transProbs`.
- Pravděpodobnosti výstupu se jmenují `$emissionProbs`.
- Situace, kdy je student veselý je označena jako `Úsměv`.
- Situaci, kdy je student smutný je pojmenována pomocí slova `Zamračení`.
- Pro úspěšnou zkoušku je použito slovo `Ano`.
- Pro neúspěšnou zkoušku je použito slovo `Ne`.

Pro spočítání nových přechodových, výstupních a počátečních pravděpodobností jsme v softwaru R využili funkci `baumWelch` [2]. Níže můžeme vidět výsledek Baum-Welch algoritmu.

```
> BaumWelchAlg$hmm
$States
[1] "Ano" "Ne"

$Symbols
[1] "Úsměv"     "Zamračení"

$startProbs
Ano  Ne
0.5 0.5

$transProbs
     to
from            Ano        Ne
  Ano  6.666667e-01 0.3333333
  Ne  5.807393e-292 1.0000000

$emissionProbs
      symbols
states        Úsměv  Zamračení
   Ano 1.000000e+00 2.4069e-52
   Ne  1.503024e-09 1.0000e+00
```

Můžeme říci, že rodiče studenta přehodnotili. Jelikož pravděpodobnost přechodu, že po úspěšné zkoušce bude následovat zase úspěšná zkouška, vyšla na 0.67 a to je menší než pravděpodobnost, která byla určena původně (0.80). Tím pádem je i vyšší pravděpodobnost neúspěšné zkoušky po úspěšné, z hodnoty 0.20 se zvýšila na 0.33.

Co se týče situace, kdy jeden den zkoušku neudělal, je téměř jisté, že další den bude zkouška také neúspěšná. Z toho vyplývá, že je zcela nepravděpodobné, aby následovala úspěšná zkouška po neúspěšné.

Pokud se podíváme na výstupní pravděpodobnosti, vidíme, že studentova nálada přímo souvisí se zkouškami. Jelikož se dá říci prakticky jistotou, že pokud student zkoušku zvládl úspěšně, pak je veselý a pokud zkoušku neudělal, pak je smutný.

Počáteční rozdělení pravděpodobností zůstalo stejné – studentova první zkouška může s pravděpodobností 0.5 dopadnout dobře a se stejnou pravděpodobností také nemusí dopadnout dobře.

○

Příklad 2.4.7. *Následující příklad bude podobný předchozímu. Rozdíl bude spočívat v tom, že nyní vyzkoušíme variantu, kdy o parametrech nic nevíme, tudíž nemáme úvodní zadání* $\mathbf{P}, \mathbf{B}$ *ani* $\mathbf{p}^{(0)}$*. Budeme tedy předpokládat, že všechny možnosti jsou stejně pravděpodobné.*

- *Počáteční rozdělení pravděpodobností* $\boldsymbol{p}^{(0)}$

$$\mathbf{p}^{(0)} = \begin{array}{c} \\ \end{array} \begin{array}{cc} \textit{Úspěch} & \textit{Neúspěch} \\ \end{array} \begin{pmatrix} 0.5 & 0.5 \end{pmatrix}.$$

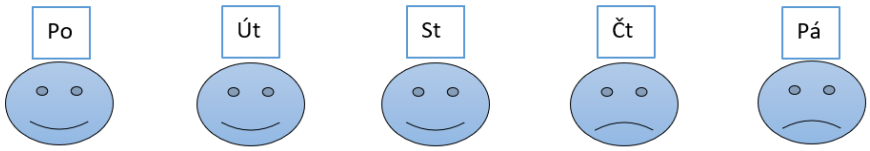
- *Matice pravděpodobností přechodu* $\boldsymbol{P}$

$$\mathbf{P} = \begin{array}{c} \\ \textit{Úspěch} \\ \textit{Neúspěch} \end{array} \begin{array}{c} \begin{array}{cc} \textit{Úspěch} & \textit{Neúspěch} \end{array} \\ \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} \end{array}.$$

- *Matice pravděpodobností výstupu* $\boldsymbol{B}$

$$\mathbf{B} = \begin{array}{c} \\ \textit{Úspěch} \\ \textit{Neúspěch} \end{array} \begin{array}{c} \begin{array}{cc} \textit{Veselý} & \textit{Smutný} \end{array} \\ \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} \end{array}.$$

- *Posloupnost pozorování*



Obrázek 2.15: Studentova nálada v průběhu týdne.

Výsledek Baum-Welch algoritmu při úvodním zadání stejně rozdělených přechodových, výstupních a počátečních pravděpodobností.

```
> BaumWelchAlg$hmm
$States
[1] "Ano" "Ne"

$Symbols
[1] "Úsměv"     "Zamračení"

$startProbs
Ano  Ne
0.5 0.5

$transProbs
     to
from  Ano  Ne
  Ano 0.5 0.5
  Ne  0.5 0.5

$emissionProbs
      symbols
states Úsměv Zamračení
   Ano   0.6       0.4
   Ne    0.6       0.4
```

Výstup ze softwaru R [11] je velmi podobný úvodnímu zadání. Počáteční pravděpodobnosti zůstaly neměnné – pravděpodobnost zůstala na úrovni 0.5 pro úvodní skrytý stav *Úspěch* a také *Neúspěch*.

Stejná situace nastala také u přechodových pravděpodobností, matice pravděpodobností přechodů zůstala stejná.

Nepatrná změna se objevila u matice pravděpodobností výstupu. Zvětšila se pravděpodobnost výstupu *Úsměvu* jak ze skrytého stavu *Úspěch*, tak ze skrytého stavu *Neúspěch* (původně 0.5, nyní 0.6 pro obě možnosti). Přímoúměrně se tedy zmenšila pravděpodobnost výstupu *Zamračení* z obou skrytých stavů *Úspěch* i *Neúspěch* (původně 0.5, nyní 0.4 pro obě možnosti).

Pokud bychom výsledky porovnali s výsledky z předchozího příkladu 2.4.6, kde jsme měli dány informace o hodnotách matic $\mathbf{P}, \mathbf{B}$ a vektoru $\mathbf{p}^{(0)}$, můžeme vidět, že výsledky se poměrně hodně liší. Hodnoty matic pravděpodobností přechodu a výstupu jsou v tomto příkladu 2.4.7 oproti předchozímu příkladu 2.4.6 více rovnoměrné. Například pravděpodobnost přechodu z *Neúspěšné zkoušky* na *Úspěšnou* je u příkladu 2.4.6 téměr nulová, kdežto u příkladu 2.4.7 je tato pravděpodobnost na hodnotě 0.5 (stejně jako všechny pravděpodobnosti v matici pravděpodobností přechodu). Podobně pravděpodobnost výstupu *Úsměvu* při *Úspěšné zkoušce* byla v příkladu 2.4.6 rovna 1, kdežto v tomto příkladu 2.4.7 je rovna hodnotě 0.6.

∘

Kapitola 3

Aplikace metod na reálná data

V následující praktické části práce ukážeme, jak se dají skryté Markovovy řetězce aplikovat v reálném použití. Uplatníme zde dříve představené metody na reálném datovém souboru. Jedná se o DNA sekvenci, kterou jsme získali pomocí nástroje BLAST [1]. Tato sekvence obsahuje nukleové báze A (Adenin), T (Thymin), G (Guanin) a C (Cytosin).

3.1. Úvod

Následující text byl vypracován podle literatury [4], [14] a [15].

DNA je všeobecné označení pro deoxyribonukleovou kyselinu. Již z názvu můžeme odvodit, že se jedná o nukleovou kyselinu. DNA obsahuje genetické informace téměř všech živých organismů. Výjimkou jsou nebuněčné organismy, kde tuto informaci nese RNA (ribonukleová kyselina). DNA je lineární řetězec nukleotidů. Každý nukleotid se skládá ze třech částí, kterými jsou deoxyribóza, fosfát a nukleová báze. Pro naše účely nás bude zajímat právě poslední zmíněná součást, tedy nukleová báze.

Nukleová báze je dusíkatá heterocyklická sloučenina. Nukleové báze jsou buď purinové, nebo pyrimidové. Mezi purinové se řadí adenin (A) a guanin (G). Pod pyrimidinové báze spadá thymin (T) a cytosin (C). Právě tyto čtyři nukleové báze jsou nejdůležitější součástí DNA pro přenos informací.

Nukleové báze tvoří komplementaritu bází. To znamená, že vytvářejí komple-

mentární páry, jež se skládají z jedné purinové báze (A, G) a jedné pyrimidové báze (T, C). Nukleová báze A se většinou páruje s T a tvoří pár AT. Naopak nukleová báze G spolu s C tvoří pár GC. Tyto vazby pak tvoří kód k zápisu genetické informace.

Výše jsme uvedli pouze stručný přehled, zájemce o tuto problematiku odkážeme na [4], [14] a [15].

Jak jsme zmínili na začátku kapitoly, praktická část bude předvedena na DNA sekvenci. Jedná se o DNA sekvenci nesoucí genetickou informaci mikroorganismu Aspergillus flavus. Použitá sekvence obsahuje 273 nukleových bází.

Následuje realizace praktické části práce, jež rozdělíme na dvě podkapitoly. V první pasáži budeme pracovat s celou sekvencí nukleových bází (Realizace první části aplikace metod na reálná data). V druhé praktické části rozdělíme sekvenci na dvě poloviny (Realizace druhé části aplikace metod na reálná data).

3.2. Realizace první části aplikace metod na reálná data

V této praktické části budeme uvažovat celou sekvenci nukleových bází. Budeme zkoumat, jak se na základě sekvence tvoří páry nukleových bází. Prvně si ale shrneme informace, které máme zadané.

- Množina skrytých stavů $\mathbf{S}$:

  Množina skrytých stavů se skládá z párů bází AT a GC.

  $$\mathbf{S} = \{AT, GC\}$$

- Počáteční rozdělení pravděpodobností $\mathbf{p}^{(0)}$:

  $$\mathbf{p}^{(0)} = \begin{array}{c} \begin{array}{cc} \text{AT} & \text{GC} \end{array} \\ \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \end{array}.$$

- Matice pravděpodobností přechodu $\mathbf{P}$:

  K inspiraci rozložení pravděpodobností nám sloužila literatura [15].

$$\mathbf{P} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{c} \begin{array}{cc} \text{AT} & \text{GC} \end{array} \\ \begin{pmatrix} 0.7 & 0.3 \\ 0.1 & 0.9 \end{pmatrix} \end{array}$$

- Matice pravděpodobností výstupu **B**:

  Také zde jsme k rozložení pravděpodobností využili [15].

$$\mathbf{B} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} 0.39 & 0.10 & 0.10 & 0.41 \\ 0.10 & 0.41 & 0.39 & 0.10 \end{pmatrix} \end{array}$$

- Posloupnost pozorování **O**:

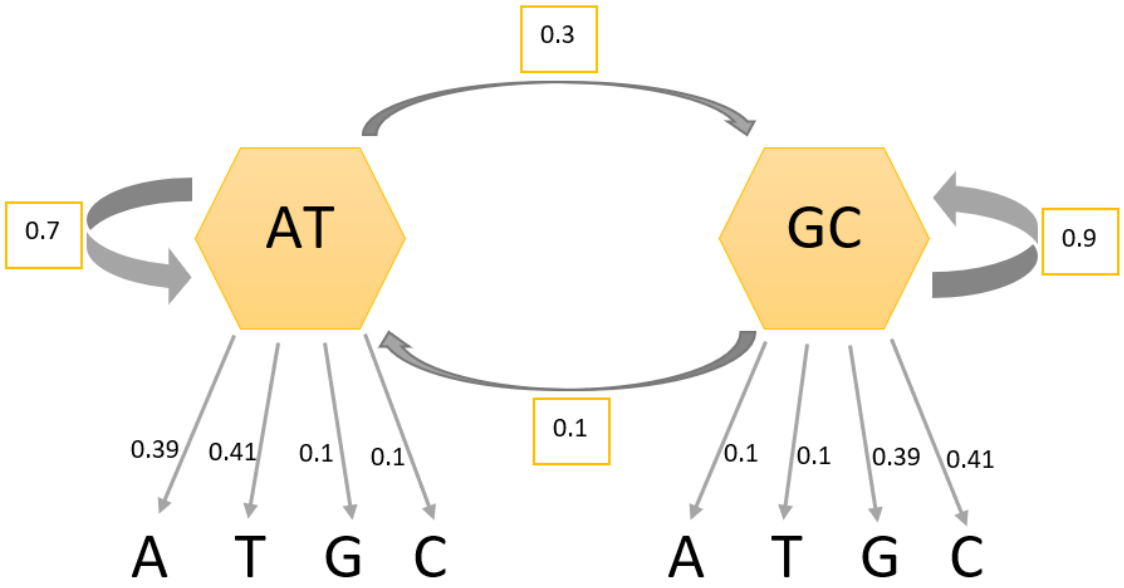
  Posloupnost pozorování **O** obsahuje celou sekvenci nukleových bází. Pro lepší přehlednost uvedeme posloupnost v tabulce (Obrázek 3.1), ve které jsme jednotlivé báze A, C, G a T barevně rozlišili.

G	A	T	T	T	G	C	G	T	T	C	G	G	C	A	A	G	C	G	C	C
G	G	C	C	G	G	G	C	C	T	A	C	A	G	A	G	C	G	G	G	T
G	A	C	A	A	A	G	C	C	C	C	A	T	A	C	G	C	T	C	G	A
G	G	A	T	C	G	G	A	C	G	C	G	G	T	G	C	C	G	C	C	G
C	T	G	C	C	T	T	T	G	G	G	G	C	C	C	G	T	C	C	C	C
C	C	C	G	G	A	G	A	G	G	G	G	A	C	G	A	C	G	A	C	C
C	A	A	C	A	C	A	C	A	A	G	C	C	G	T	G	C	T	T	G	A
T	G	G	G	C	A	G	C	A	A	T	G	A	C	G	C	T	C	G	G	A
C	A	G	G	C	A	T	G	C	C	C	C	C	C	G	G	A	A	T	A	C
C	A	G	G	G	G	G	C	G	C	A	A	T	G	T	G	C	G	T	T	C
A	A	A	G	A	C	T	C	G	A	T	G	A	T	T	C	A	C	G	G	A
A	T	T	C	T	G	C	A	A	T	T	C	A	C	A	C	T	A	G	T	T
A	T	C	G	C	A	T	T	T	C	G	C	T	G	C	G	T	T	C	T	T

Obrázek 3.1: Posloupnost pozorování **O**.

Nyní máme všechny potřebné informace a můžeme přejít k samotným úlohám, kterými jsou evaluace, dekódování a učení. Ke kalkulaci jednotlivých algoritmů využijeme statistický software R [11]. Před zahájením samotných výpočtů je potřeba si do softwaru R nainstalovat knihovnu `HMM` [2].

Na tomto místě si pro ilustraci uvedeme diagram skrytého Markovova řetězce (Obrázek 3.2), který obsahuje zadané parametry.



Obrázek 3.2: Skrytý Markovovův řetězec.

3.2.1. Evaluace

V této úloze chceme zjistit, jak je pravděpodobné vygenerování dané posloupnosti pozorování $\mathbf{O}$ daným modelem λ. K výpočtům evaluace využijeme algoritmy dopředný a zpětný, které jsme si zavedli dříve.

Dopředný algoritmus

K výpočtu dopředného algoritmu využijeme funkci `forward` v softwaru R [2]. Do funkce jsme zadali parametry modelu λ, tedy počáteční rozdělení pravděpodobnosti $\mathbf{p}^{(\mathbf{0})}$, matici pravděpodobností přechodu $\mathbf{P}$, matici pravděpodobností výstupu $\mathbf{B}$ a také posloupnost pozorování $\mathbf{O}$. Na základě těchto zadaných informací jsme získali výslednou pravděpodobnost přibližně $9.65e^{-163}$. Můžeme tedy

říci, že posloupnost **O** je generována pomocí takto nadefinovaného modelu λ s pravděpodobností $9.65e^{-163}$. Tato nízká pravděpodobnost je zapříčiněna velikostí sekvence neboli posloupnosti pozorování **O**.

Zpětný algoritmus

V softwaru R též spočítáme zpětný algoritmus, nyní prostřednictvím funkce `backward` [2]. Před spuštěním funkce jsme do softwaru R zadali všechny potřebné parametry, stejně jako u předchozího – dopředného algoritmu. Zadali jsme tedy počáteční rozdělení pravděpodobnosti $\mathbf{p^{(0)}}$, matici pravděpodobností přechodu **P**, matici pravděpodobností výstupu **B** a posloupnost pozorování **O**. Při využití zpětného algoritmu jsme získali stejnou pravděpodobnost generování naší posloupnosti **O** při námi zadanými parametry modelu. Pravděpodobnost je rovna přibližně hodnotě $9.65e^{-163}$.

3.2.2. Dekódování

Úloha dekódování odpovídá na otázku, jaký je nejvíce pravděpodobný průchod skrytými stavy. Tedy v jakém pořadí a počtu bude model generovat *AT*, respektive *GC*. S řešením této otázky nám pomůže Viterbiho metoda.

K výpočtům tohoto algoritmu jsme využili v softwaru R již nadefinovanou funkci `viterbi` [2]. Před zahájením výpočtu jsme do sotwaru R zadali potřebné informace. Těmito informacemi jsou myšleny hodnoty matice pravděpodobností přechodu **P**, matice pravděpodobností výstupu **B**, počáteční rozdělení pravděpodobnosti $\mathbf{p}^{(0)}$ a posloupnost pozorování **O**. Pomocí této funkce `viterbi` jsme získali následující posloupnost skrytých stavů **X**, viz Obrázek 3.3.

AT	AT	AT	AT	AT	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	AT	AT	AT	GC	GC	GC	GC	GC	AT	AT	AT	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	AT	AT	AT	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	AT	AT	AT	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	AT	AT	AT	AT	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	AT	AT	AT	GC	GC	GC	GC	GC	AT	AT	AT
AT	AT	AT	GC	GC	GC	GC	GC	GC	AT	AT	AT	AT	AT	AT	GC	GC	GC	GC	GC	AT
AT	AT	AT	GC	GC	GC	GC	AT	AT	AT	AT	AT	AT	AT	AT	AT	AT	AT	AT	AT	AT
AT	AT	GC	GC	GC	AT	AT	AT	AT	GC	GC	GC	GC	GC	GC	GC	AT	AT	AT	AT	AT

Obrázek 3.3: Posloupnost skrytých stavů $\mathbf{X}$.

Z posloupnosti skrytých stavů $\mathbf{X}$ (Obrázek 3.3) je již na první pohled viditelná převahaha skrytých stavů GC. Skryté stavy GC se v posloupnosti vyskytují hned 208 krát. Naopak skryté stavy AT jsou v posloupnosti zastoupeny v počtu 65.

3.2.3. Učení

Tato úloha napomáhá k lepšímu nadefinování modelu λ. Jak jsme si uvedli již dříve, v úloze učení se využívá Baum-Welchův algoritmus. Stejně jako k řešení předchozích úloh, i v tomto případě jsme použili již zadefinovanou funkci v softwaru R a sice `baumWelch` [2]. Také tato funkce vyžadovala zadání úvodních hodnot počátečního rozdělení pravděpodobnosti $\mathbf{p}^{(0)}$, matice pravděpodobností přechodu $\mathbf{P}$ a matice pravděpodobností výstupu $\mathbf{B}$. Dále jsme do funkce zadali posloupnost pozorování $\mathbf{O}$ a posloupnost skrytých stavů $\mathbf{X}$, kterou jsme získali v úloze dekódování prostřednictvím Viterbiho metody. Navíc jsme ponechali výchozí nastavení funkce počtu iterací na hodnotě 100. Výsledkem je nové počáteční rozdělení pravděpodobnosti $\hat{\mathbf{p}}^{(0)}$, nová matice pravděpodobností přechodu $\hat{\mathbf{P}}$ a nová matice pravděpodobností výstupu $\hat{\mathbf{B}}$.

Prostřednictvím funkce `baumWelch` v softwaru R jsme obdrželi nové počáteční rozdělení pravděpodobnosti $\hat{\mathbf{p}}^{(0)}$ mající tvar:

$$\hat{\mathbf{p}}^{(0)} = \begin{array}{c} \\ \end{array} \begin{array}{c} \begin{array}{cc} \text{AT} & \text{GC} \end{array} \\ \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \end{array}.$$

Abychom mohli provést srovnání, uvedeme si zde také původní počáteční rozdělení pravděpodobnosti $\mathbf{p}^{(0)}$:

$$\mathbf{p}^{(0)} = \begin{array}{c} \\ \end{array} \begin{array}{c} \begin{array}{cc} \text{AT} & \text{GC} \end{array} \\ \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \end{array}.$$

Získali jsme také novou matici pravděpodobností přechodu $\hat{\mathbf{P}}$, jež má podobu:

$$\hat{\mathbf{P}} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{c} \begin{array}{cc} \text{AT} & \text{GC} \end{array} \\ \begin{pmatrix} 0.682 & 0.318 \\ 0.124 & 0.876 \end{pmatrix} \end{array}.$$

Původní matice pravděpodobností přechodu $\mathbf{P}$ měla tvar:

$$\mathbf{P} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{c} \begin{array}{cc} \text{AT} & \text{GC} \end{array} \\ \begin{pmatrix} 0.700 & 0.300 \\ 0.100 & 0.900 \end{pmatrix} \end{array}.$$

Nová matice pravděpodobností výstupu $\hat{\mathbf{B}}$ získala následující podobu:

$$\hat{\mathbf{B}} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} 0.384 & 0.124 & 0.113 & 0.379 \\ 0.136 & 0.383 & 0.378 & 0.103 \end{pmatrix} \end{array}.$$

Původní matice pravděpodobností výstupu $\mathbf{B}$, ke srovnání s novou maticí, měla podobu:

$$\mathbf{B} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} 0.390 & 0.100 & 0.100 & 0.410 \\ 0.100 & 0.410 & 0.390 & 0.100 \end{pmatrix} \end{array}.$$

Jak můžeme vidět v porovnání výše, počáteční rozdělení pravděpodobnosti zůstalo neměnné. Co se však změnilo, jsou matice pravděpodobností přechodu a výstupu. U první zmíněné matice (matice pravděpodobností přechodu) se hodnoty pravděpodobností změnily pouze nepatrně. Pravděpodobnost přechodu ze skrytého stavu AT do skrytého stavu AT se snížila přibližně na 0.682 (původní hodnota byla 0.7). Z toho vyplývá, že se přímoúměrně zvýšila druhá možnost přechodu ze skrytého stavu AT, tedy do skrytého stavu GC, přibližně na 0.318 (z původní hodnoty 0.3). Možnost přechodu ze skrytého stavu GC se také změnila, při přechodu z GC do AT se pravděpodobnost zvýšila přibližně na hodnotu 0.124 z původní hodnoty 0.1. Tím pádem pravděpodobnost přechodu ze stejného stavu GC do GC se snížila na hodnotu 0.876 z počáteční hodnoty 0.9. Také u matice pravděpodobností výstupu se změnily hodnoty pravděpodobností. Pravděpodobnosti výstupů A, C, G, resp. T ze skrytého stavu AT jsou nově přibližně 0.384, 0.124, 0.113, resp. 0.379 (původní hodnoty byly 0.390, 0.100, 0.100, resp. 0.410). Z druhého skrytého stavu GC jsou nově výstupy A, C, G, resp. T ohodnoceny pravděpodobnostmi přibližně 0.136, 0.383, 0.378, resp. 0.103 (původní pravděpodobnosti měli hodnoty 0.100, 0.410, 0.390, resp. 0.100).

3.2.4. Evaluace s novými parametry

V úloze učení jsme prostřednictvím Baum-Welchova algoritmu získali novou matici pravděpodobností přechodu $\hat{\mathbf{P}}$, novou matici pravděpodobností výstupu $\hat{\mathbf{B}}$ a nové počáteční rozdělení pravděpodobnosti $\hat{\mathbf{p}}^{(0)}$. Provedeme tedy přepočet dopředného a zpětného algoritmu na základě nově získaných parametrů.

Potřebné parametry vypadají následovně:

- Množina skrytých stavů $\mathbf{S}$:

$$\mathbf{S} = \{AT, GC\}.$$

- Nové počáteční pravděpodobnosti $\hat{\mathbf{p}}^{(0)}$:

$$\hat{\mathbf{p}}^{(0)} = \begin{array}{c} \\ \end{array} \begin{array}{cc} \text{AT} & \text{GC} \\ \end{array} \begin{pmatrix} 0.5 & 0.5 \end{pmatrix}.$$

- Nová matice pravděpodobností přechodu $\hat{\mathbf{P}}$:

$$\hat{\mathbf{P}} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{cc} \text{AT} & \text{GC} \\ \end{array} \begin{pmatrix} 0.682 & 0.318 \\ 0.124 & 0.876 \end{pmatrix}.$$

- Nová matice pravděpodobností výstupu $\hat{\mathbf{B}}$:

$$\hat{\mathbf{B}} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \\ \end{array} \begin{pmatrix} 0.384 & 0.124 & 0.113 & 0.379 \\ 0.136 & 0.383 & 0.378 & 0.103 \end{pmatrix}.$$

- Posloupnost pozorování $\mathbf{O}$:

G	A	T	T	T	G	C	G	T	T	C	G	G	C	A	A	G	C	G	C	C
G	G	C	C	G	G	G	C	C	T	A	C	A	G	A	G	C	G	G	G	T
G	A	C	A	A	A	G	C	C	C	C	A	T	A	C	G	C	T	C	G	A
G	G	A	T	C	G	G	A	C	G	C	G	G	T	G	C	C	G	C	C	G
C	T	G	C	C	T	T	T	G	G	G	G	C	C	C	G	T	C	C	C	C
C	C	C	G	G	A	G	A	G	G	G	G	A	C	G	A	C	G	A	C	C
C	A	A	C	A	C	A	C	A	A	G	C	C	G	T	G	C	T	T	G	A
T	G	G	G	C	A	G	C	A	A	T	G	A	C	G	C	T	C	G	G	A
C	A	G	G	C	A	T	G	C	C	C	C	C	C	G	G	A	A	T	A	C
C	A	G	G	G	G	G	C	G	C	A	A	T	G	T	G	C	G	T	T	C
A	A	A	G	A	C	T	C	G	A	T	G	A	T	T	C	A	C	G	G	A
A	T	T	C	T	G	C	A	A	T	T	C	A	C	A	C	T	A	G	T	T
A	T	C	G	C	A	T	T	T	C	G	C	T	G	C	G	T	T	C	T	T

Obrázek 3.4: Posloupnost pozorování $\mathbf{O}$.

Dopředný algoritmus

Stejně jako v předchozím výpočtu dopředného algoritmu i zde použijeme v softwaru R funkci `forward` [2]. Do funkce nyní zadáme nové matice $\hat{\mathbf{P}}$, $\hat{\mathbf{B}}$, nový vektor $\hat{\mathbf{p}}^{(0)}$ a posloupnost pozorování $\mathbf{O}$. Po spuštění funkce `forward` jsme obdrželi výslednou pravděpodobnost, která činí přibližně $1.37e^{-161}$. Hodnota tohoto výsledku je vyšší nežli hodnota výsledku při zadání původních maticí $\mathbf{P}$, $\mathbf{B}$ a vektoru $\mathbf{p^{(0)}}$, jež byla přibližně $9.65e^{-163}$.

Zpětný algoritmus

Podobně budeme postupovat také u zpětného algoritmu. Ve funkci `backward` v softwaru R [2] použijeme hodnoty nových matic $\hat{\mathbf{P}}$, $\hat{\mathbf{B}}$, nového vektoru $\hat{\mathbf{p}}^{(0)}$ a posloupnost pozorování $\mathbf{O}$. Výsledek, který jsme získali prostřednictvím funkce `backward` je pravděpodobnost s hodnotou $1.37e^{-161}$. Pravděpodobnost je totožná s výsledkem získaným pomocí dopředného algoritmu.

V důsledku můžeme konstatovat, že úloha učení splnila svůj účel. Pomocí Baum-Welch algoritmu jsme předefinovali původní model $\lambda(\mathbf{P}, \mathbf{B}, \mathbf{p^{(0)}})$ na nový

model $\lambda'(\hat{\mathbf{P}}, \hat{\mathbf{B}}, \hat{\mathbf{p}}^{(0)})$. Díky čemuž jsme následně obdrželi lepší výsledek v úloze evaluace. To znamená, že nový model je nadefinovaný lépe nežli ten původní.

3.2.5. Dekódování s novými parametry

V této podkapitole provedeme výpočet pomocí stejné funkce `viterbi` jako tomu bylo v podkapitole *Dekódování*. Rozdíl bude v parametrech, které do funkce zadáme. Nyní využijeme získanou novou matici pravděpodobností přechodu $\hat{\mathbf{P}}$, novou matici pravděpodobností výstupu $\hat{\mathbf{B}}$ a nový vektor počátečních pravděpodobností $\hat{\mathbf{p}}^{(0)}$. Tyto parametry jsou nám známy z úlohy učení díky Baum-Welchovému algoritmu.

Pro výpočet Viterbiho metody tedy máme tyto informace:

- Množina skrytých stavů $\mathbf{S}$:

$$\mathbf{S} = \{AT, GC\}.$$

- Nové počáteční pravděpodobnosti $\hat{\mathbf{p}}^{(0)}$:

$$\hat{\mathbf{p}}^{(0)} = \begin{array}{c} \begin{array}{cc} \text{AT} & \text{GC} \end{array} \\ \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \end{array}.$$

- Nová matice pravděpodobností přechodu $\hat{\mathbf{P}}$:

$$\hat{\mathbf{P}} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{c} \begin{array}{cc} \text{AT} & \text{GC} \end{array} \\ \begin{pmatrix} 0.682 & 0.318 \\ 0.124 & 0.876 \end{pmatrix} \end{array}.$$

- Nová matice pravděpodobností výstupu $\hat{\mathbf{B}}$:

$$\hat{\mathbf{B}} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} 0.384 & 0.124 & 0.113 & 0.379 \\ 0.136 & 0.383 & 0.378 & 0.103 \end{pmatrix} \end{array}.$$

- Posloupnost pozorování $\mathbf{O}$:

G	A	T	T	T	G	C	G	T	T	C	G	G	C	A	A	G	C	G	C	C
G	G	C	C	G	G	G	C	C	T	A	C	A	G	A	G	C	G	G	G	T
G	A	C	A	A	A	G	C	C	C	C	A	T	A	C	G	C	T	C	G	A
G	G	A	T	C	G	G	A	C	G	C	G	G	T	G	C	C	G	C	C	G
C	T	G	C	C	T	T	T	G	G	G	G	C	C	C	G	T	C	C	C	C
C	C	C	G	G	A	G	A	G	G	G	G	A	C	G	A	C	G	A	C	C
C	A	A	C	A	C	A	C	A	A	G	C	C	G	T	G	C	T	T	G	A
T	G	G	G	C	A	G	C	A	A	T	G	A	C	G	C	T	C	G	G	A
C	A	G	G	C	A	T	G	C	C	C	C	C	C	G	G	A	A	T	A	C
C	A	G	G	G	G	G	C	G	C	A	A	T	G	T	G	C	G	T	T	C
A	A	A	G	A	C	T	C	G	A	T	G	A	T	T	C	A	C	G	G	A
A	T	T	C	T	G	C	A	A	T	T	C	A	C	A	C	T	A	G	T	T
A	T	C	G	C	A	T	T	T	C	G	C	T	G	C	G	T	T	C	T	T

Obrázek 3.5: Posloupnost pozorování $\mathbf{O}$.

Prostřednictvím funkce `viterbi` v softwaru R jsme obdrželi následující posloupnost skrytých stavů $\hat{\mathbf{X}}$ (Obrázek 3.6).

AT	AT	AT	AT	AT	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	AT	AT	AT	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	AT	AT	AT	AT	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	AT	AT	AT	AT
AT	AT	AT	GC	GC	GC	GC	GC	GC	AT	AT	AT	AT	AT	AT	GC	GC	GC	GC	GC	AT
AT	AT	AT	GC	GC	GC	GC	AT	AT	AT	AT	AT	AT	AT	AT	AT	AT	AT	AT	AT	AT
AT	AT	GC	GC	GC	AT	AT	AT	AT	GC	GC	GC	GC	GC	GC	GC	AT	AT	AT	AT	AT

Obrázek 3.6: Posloupnost skrytých stavů $\hat{\mathbf{X}}$.

Získaná posloupnost skrytých stavů (Obrázek 3.6) je ve své podstatě velice podobná posloupnosti (Obrázek 3.3), kterou jsme obdrželi prostřednictvím původních hodnot matic $\mathbf{P}$, $\mathbf{B}$ a vektoru $\mathbf{p}^{(\mathbf{0})}$. Rozdíl, který se zde objevuje, je v

nahrazení několika skrytých stavů AT skrytými stavy GC, tyto skryté stavy jsou v posloupnosti označeny červenou barvou. V této podobě posloupnosti skrytých stavů $\mathbf{X}$ se vyskytuje 53 skrytých stavů AT a 220 skrytých stavů GC.

3.2.6. Učení s rovnoměrně rozloženými parametry

Nyní předvedeme variantu výpočtu Baum-Welchova algoritmu, kdy jsou všechny pravděpodobnosti rozloženy rovnoměrně. Předpokládejme tedy následující podobu vektoru $\mathbf{p^{(0)}}$, matice $\mathbf{P}$ a matice $\mathbf{B}$.

- Počáteční rozdělení pravděpodobností $\mathbf{p}^{(0)}$:

$$\mathbf{p}^{(0)} = \begin{array}{c} \\ \end{array} \begin{array}{cc} \text{AT} & \text{GC} \\ \end{array} \begin{pmatrix} 0.5 & 0.5 \end{pmatrix}.$$

- Matice pravděpodobností přechodu $\mathbf{P}$:

$$\mathbf{P} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{cc} \text{AT} & \text{GC} \\ \end{array} \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}.$$

- Matice pravděpodobností výstupu $\mathbf{B}$:

$$\mathbf{B} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \\ \end{array} \begin{pmatrix} 0.250 & 0.250 & 0.250 & 0.250 \\ 0.250 & 0.250 & 0.250 & 0.250 \end{pmatrix}.$$

- Množina skrytých stavů $\mathbf{S}$:

$$\mathbf{S} = \{AT, GC\}.$$

- Posloupnost pozorování $\mathbf{O}$:

G	A	T	T	T	G	C	G	T	T	C	G	G	C	A	A	G	C	G	C	C
G	G	C	C	G	G	G	C	C	T	A	C	A	G	A	G	C	G	G	G	T
G	A	C	A	A	A	G	C	C	C	C	A	T	A	C	G	C	T	C	G	A
G	G	A	T	C	G	G	A	C	G	C	G	G	T	G	C	C	G	C	C	G
C	T	G	C	C	T	T	T	G	G	G	G	C	C	C	G	T	C	C	C	C
C	C	C	G	G	A	G	A	G	G	G	G	A	C	G	A	C	G	A	C	C
C	A	A	C	A	C	A	C	A	A	G	C	C	G	T	G	C	T	T	G	A
T	G	G	G	C	A	G	C	A	A	T	G	A	C	G	C	T	C	G	G	A
C	A	G	G	C	A	T	G	C	C	C	C	C	C	G	G	A	A	T	A	C
C	A	G	G	G	G	G	C	G	C	A	A	T	G	T	G	C	G	T	T	C
A	A	A	G	A	C	T	C	G	A	T	G	A	T	T	C	A	C	G	G	A
A	T	T	C	T	G	C	A	A	T	T	C	A	C	A	C	T	A	G	T	T
A	T	C	G	C	A	T	T	T	C	G	C	T	G	C	G	T	T	C	T	T

Obrázek 3.7: Posloupnost pozorování **O**.

Po dosazení výše zmíněných parametrů do funkce `baumWelch` v softwaru R jsme obdrželi přepočítané pravděpodobnosti.

Nový tvar počátečního rozdělení pravděpodobností $\hat{\mathbf{p}}^{(0)}$ vypadá následovně:

$$\hat{\mathbf{p}}^{(0)} = \begin{array}{c} \\ \end{array} \begin{array}{c} \begin{array}{cc} \text{AT} & \text{GC} \end{array} \\ \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \end{array}.$$

Přepočtená matice pravděpodobností přechodu $\hat{\mathbf{P}}$ má následující podobu:

$$\hat{\mathbf{P}} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{c} \begin{array}{cc} \text{AT} & \text{GC} \end{array} \\ \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} \end{array}.$$

Nově získaná matice pravděpodobností výstupu $\hat{\mathbf{B}}$ má tvar:

$$\hat{\mathbf{B}} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} 0.206 & 0.310 & 0.303 & 0.181 \\ 0.206 & 0.310 & 0.303 & 0.181 \end{pmatrix} \end{array}.$$

Nyní provedeme porovnání výsledků s původními hodnotami. Dojdeme k závěru, že původní vektor počátečních pravděpodobností $\mathbf{p}^{(\mathbf{0})}$ je totožný s novým

vektorem $\hat{\mathbf{p}}^{(0)}$, jehož pravděpodobnosti zůstaly na hodnotě 0.5. To samé můžeme říci o původní matici pravděpodobností přechodu $\mathbf{P}$ a nově získané matici $\hat{\mathbf{P}}$, obě matice nesou stejnou podobu s pravděpodobnostmi 0.5 pro všechny možnosti. Naopak odlišné hodnoty mají původní matice pravděpodobností výstupu $\mathbf{B}$ spolu s novou maticí pravděpodobností výstupu $\hat{\mathbf{B}}$. Pravděpodobnosti výstupů A, C, G, resp. T ze skrytého stavu AT jsou nově ohodnoceny přibližně 0.206, 0.310, 0.303, resp. 0.181. Pravděpodobnosti výstupů A, C, G, resp. T z druhého skrytého stavu GC mají nově hodnoty 0.206, 0.310, 0.303, resp. 0.181. Původní pravděpodobnosti byly ohodnoceny rovnoměrně, tedy hodnotou 0.250 pro výstupy A, C, G, T z obou skrytých stavů AT i GC.

Na základě získaných výsledků lze usoudit, že rovnoměrné rozdělení vstupních parametrů není ideální volbou. Algoritmus pak není schopen odhalit rozdíly mezi jednotlivými skrytými stavy AT a GC.

3.3. Realizace druhé části aplikace metod na reálná data

V této praktické části práce rozdělíme sekvenci nukleových bází na dvě poloviny, získáme tedy dvě posloupnosti pozorování $\mathbf{O}_1$, $\mathbf{O}_2$. První část posloupnosti pozorování $\mathbf{O}_1$ použijeme k provedení výpočtů jednotlivých algoritmů. Na základě odhadnutých parametrů z první poloviny posloupnosti pozorování $\mathbf{O}_1$ najdeme také optimální posloupnost skrytých stavů pro druhou polovinu posloupnosti pozorování $\mathbf{O}_2$.

Nyní si ještě uvedeme vstupní parametry, jež pro tuto chvíli máme k dispozici.

- Množina skrytých stavů $\mathbf{S}$:

$$\mathbf{S} = \{AT, GC\}.$$

- Počáteční rozdělení pravděpodobnosti $\mathbf{p}^{(0)}$:

$$\mathbf{p}^{(0)} = \begin{array}{c} \begin{array}{cc} \text{AT} & \text{GC} \end{array} \\ \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \end{array}.$$

- Matice pravděpodobností přechodu $\mathbf{P}$:

$$\mathbf{P} = \begin{array}{c} \begin{array}{cc} & \begin{array}{cc} \text{AT} & \text{GC} \end{array} \\ \begin{array}{c} \text{AT} \\ \text{GC} \end{array} & \begin{pmatrix} 0.7 & 0.3 \\ 0.1 & 0.9 \end{pmatrix} \end{array} \end{array}.$$

- Matice pravděpodobností výstupu $\mathbf{B}$:

$$\mathbf{B} = \begin{array}{c} \begin{array}{cccc} & \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{array}{c} \text{AT} \\ \text{GC} \end{array} \begin{pmatrix} 0.39 & 0.10 & 0.10 & 0.41 \\ 0.10 & 0.41 & 0.39 & 0.10 \end{pmatrix} \end{array}.$$

- První část posloupnosti pozorování $\mathbf{O}_1$:

G	A	T	T	T	G	C	G	T	T	C	G	G	C	A	A	G	C	G	C	C
G	G	C	C	G	G	G	C	C	T	A	C	A	G	A	G	C	G	G	G	T
G	A	C	A	A	A	G	C	C	C	C	A	T	A	C	G	C	T	C	G	A
G	G	A	T	C	G	G	A	C	G	C	G	G	T	G	C	C	G	C	C	G
C	T	G	C	C	T	T	T	G	G	G	G	C	C	C	G	T	C	C	C	C
C	C	C	G	G	A	G	A	G	G	G	G	A	C	G	A	C	G	A	C	C
C	A	A	C	A	C	A	C	A	A	G										

Obrázek 3.8: První část posloupnosti pozorování.

- Druhá část posloupnosti pozorování $\mathbf{O}_2$:

C	C	G	T	G	C	T	T	G	A	T	G	G	G	C	A	G	C	A	A	T
G	A	C	G	C	T	C	G	G	A	C	A	G	G	C	A	T	G	C	C	C
C	C	C	G	G	A	A	T	A	C	C	A	G	G	G	G	G	C	G	C	A
A	T	G	T	G	C	G	T	T	C	A	A	A	G	A	C	T	C	G	A	T
G	A	T	T	C	A	C	G	G	A	A	T	T	C	T	G	C	A	A	T	T
C	A	C	A	C	T	A	G	T	T	A	T	C	G	C	A	T	T	T	C	G
C	T	G	C	G	T	T	C	T	T											

Obrázek 3.9: Druhá část posloupnosti pozorování.

3.3.1. Evaluace

Dopředný algoritmus

Výpočet dopředného algoritmu jsme provedli, stejně jako v předchozí praktické části práce, pomocí funkce `forward` v softwaru R. Do funkce jsme zadali stejné parametry jako dříve, jediná změna nastala u posloupnosti pozorování. Namísto celé posloupnosti pozorování $\mathbf{O}$ jsme do funkce zadali její první polovinu $\mathbf{O}_1$. Výsledná pravděpodobnost, kterou jsme obdrželi nyní činí přibližně $3.21e^{-79}$.

Zpětný algoritmus

Pro spočtení zpětného algoritmu jsme použili funkci `backward` v softwaru R. Funkce je stejná s funkcí, kterou jsme použili pro zpětný algoritmus v první praktické části práce. Totožné jsou také parametry, jež jsme do funkce zadali, rozdíl je pouze v užité posloupnosti pozorování. Stejně jako výše v dopředném algoritmu, využijeme první polovinu posloupnosti pozorování $\mathbf{O}_1$. Za pomoci funkce `backward` jsme pro posloupnost pozorování $\mathbf{O}_1$ získali stejnou výslednou pravděpodobnost jako s využitím funkce `forward`, tedy pravděpodobnost je rovna přibližně hodnotě $3.21e^{-79}$.

Výsledná pravděpodobnost $3.21e^{-79}$ generování posloupnosti $\mathbf{O}_1$ je tedy vyšší

nežli výsledná pravděpodobnost vygenerování posloupnosti **O** v první praktické části, která činila 9.65^{-163}. Pravděpodobnost je tedy tím větší, čím menší je posloupnost pozorování.

3.3.2. Dekódování

Také v úloze dekódování jsme využili rozdělené posloupnosti pozorování **O**. Funkci `viterbi` jsme spustili pro první část posloupnosti pozorování $\mathbf{O}_1$. Obdrželi jsme výsledek v podobě posloupnosti skrytých stavů $\mathbf{X}_1$ (Obrázek 3.10), který si zde znázorníme.

AT	AT	AT	AT	AT	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	AT	AT	AT	GC	GC	GC	GC	GC	AT	AT	AT	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	AT	AT	AT	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC										

Obrázek 3.10: Posloupnost skrytých stavů.

Provedeme porovnání této posloupnosti skrytých stavů $\mathbf{X}_1$ (Obrázek 3.10) s posloupností skrytých stavů **X** (Obrázek 3.3), již jsme získali v první praktické části práce. Zjistíme, že posloupnost skrytých stavů $\mathbf{X}_1$ odpovídá první polovině posloupnosti skrytých stavů **X**, viz Obrázek 3.3.

3.3.3. Učení

V této sekci učení spočítáme pomocí funkce `baumWelch` v softwaru R Baum-Welchův algoritmus. Do této funkce zadáme stejné parametry jako v předchozím výpočtu Baum-Welchova algoritmu, změna bude pouze v použité posloupnosti pozorování. Nyní namísto posloupnosti pozorování **O**, použijeme pouze její první polovinu $\mathbf{O}_1$.

Výsledek, který jsme obdrželi se skládá z nového vektoru počátečních pravděpodobností $\hat{\mathbf{p}}^{(0)}$, nové matice pravděpodobností přechodu $\hat{\mathbf{P}}$ a nové matice pravděpodobností výstupu $\hat{\mathbf{B}}$.

Tyto matice $\hat{\mathbf{P}}$ a $\hat{\mathbf{B}}$ a vektor $\hat{\mathbf{p}}^{(0)}$ si zde uvedeme spolu s původními maticemi $\mathbf{P}$ a $\mathbf{B}$ a vektorem $\mathbf{p^{(0)}}$ pro snažší porovnání výsledků.

Nové a původní počáteční rozdělení pravděpodobnosti $\mathbf{p^{(0)}}$ a $\hat{\mathbf{p}}^{(0)}$ mají tvar:

$$\hat{\mathbf{p}}^{(0)} = \begin{array}{c} \\ \end{array} \begin{array}{cc} \text{AT} & \text{GC} \\ \end{array} \begin{pmatrix} 0.5 & 0.5 \end{pmatrix},$$

$$\mathbf{p}^{(0)} = \begin{array}{c} \\ \end{array} \begin{array}{cc} \text{AT} & \text{GC} \\ \end{array} \begin{pmatrix} 0.5 & 0.5 \end{pmatrix}.$$

Matice pravděpodobností přechodu $\hat{\mathbf{P}}$ a $\mathbf{P}$ jsou v následující podobě:

$$\hat{\mathbf{P}} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{cc} \text{AT} & \text{GC} \\ \end{array} \begin{pmatrix} 0.602 & 0.398 \\ 0.097 & 0.903 \end{pmatrix},$$

$$\mathbf{P} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{cc} \text{AT} & \text{GC} \\ \end{array} \begin{pmatrix} 0.700 & 0.300 \\ 0.100 & 0.900 \end{pmatrix}.$$

Přepočtená a původní matice pravděpodobností výstupu $\hat{\mathbf{B}}$ a $\mathbf{B}$ vypadají následovně:

$$\hat{\mathbf{B}} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \\ \end{array} \begin{pmatrix} 0.411 & 0.145 & 0.144 & 0.300 \\ 0.136 & 0.397 & 0.389 & 0.078 \end{pmatrix},$$

$$\mathbf{B} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \\ \end{array} \begin{pmatrix} 0.390 & 0.100 & 0.100 & 0.410 \\ 0.100 & 0.410 & 0.390 & 0.100 \end{pmatrix}.$$

Již na první pohled je viditelná rovnost nového a původního počáteční rozdělení pravděpodobnosti ($\hat{\mathbf{p}}^{(0)} = \mathbf{p}^{(0)}$). Naopak při srovnání nové matice pravděpodobností přechodu $\hat{\mathbf{P}}$ a původní matice pravděpodobností přechodu $\mathbf{P}$ je vidět změna v hodnotách pravděpodobností. Přechod ze skrytého stavu AT do stejného skrytého stavu AT je nově ohodnocen pravděpodobností cca 0.602. Tato pravděpodobnost je nižší než původní pravděpodobnost, která byla na hodnotě 0.700. Druhou možností přechodu ze skrytého stavu AT je přechod do skrytého stavu GC, jehož pravděpodobnost se zvýšila na hodnotu 0.398 (z původní hodnoty 0.300). Pravděpodobnosti přechodu ze skrytého stavu GC zůstaly téměř neměnné. Pravděpodobnost přechodu ze skrytého stavu GC do AT se nepatrně snížila na hodnotu 0.097 (původní pravděpodobnost byla 0.100). Pravděpodobnost přechodu ze skrytého stavu GC do AT se zvýšila na 0.903 z původní hodnoty 0.900.

Rozdíl pozorujeme také u matic pravděpodobností výstupu ($\hat{\mathbf{B}}$ a $\mathbf{B}$). Pravděpodobnosti výstupů A, C, G, resp. T ze skrytého stavu AT jsou nyní ohodnoceny pravděpodobnostmi cca 0.411, 0.145, 0.144, resp. 0.300 (původní hodnoty byly 0.390, 0.100, 0.100, resp. 0.410). Pravděpodobnosti výstupů A, C, G, resp. T z druhého skrytého stavu GC jsou nově přibližně 0.136, 0.397, 0.389, resp. 0.078 (původní hodnoty pravděpodobností byly 0.100, 0.410, 0.390, resp. 0.100).

Na tomto místě uvedeme shrnující tabulku (Obrázek 3.11). V tabulce je znázorněno, jak se liší původní hodnoty od hodnot odhadnutých z celé sekvence a z její první poloviny.

Parametry			Původní hodnoty	Odhadnuté hodnoty z celé sekvence	Odhadnuté hodnoty z první poloviny sekvence
Počáteční rozdělení pravděpodobností	$\mathbf{p}^{(0)}$	$p_1^{(0)}$	0.500	0.500	0.500
		$p_2^{(0)}$	0.500	0.500	0.500
Matice pravděpodobností přechodu	**P**	p_{11}	0.700	0.682	0.602
		p_{12}	0.300	0.318	0.398
		p_{21}	0.100	0.124	0.097
		p_{22}	0.900	0.876	0.903
Matice pravděpodobností výstupu	**B**	$b_1(1)$	0.390	0.384	0.411
		$b_1(2)$	0.100	0.124	0.145
		$b_1(3)$	0.100	0.113	0.144
		$b_1(4)$	0.410	0.379	0.300
		$b_2(1)$	0.100	0.136	0.136
		$b_2(2)$	0.410	0.383	0.397
		$b_2(3)$	0.390	0.378	0.389
		$b_2(4)$	0.100	0.103	0.078

Obrázek 3.11: Srovnání parametrů.

3.3.4. Evaluace s novými parametry

Prostřednictvím Baum-Welchova algoritmu jsme obdrželi nové parametry modelu λ', kterými jsou počáteční rozdělení pravděpodobnosti $\hat{\mathbf{p}}^{(0)}$, matice pravděpodobností přechodu $\hat{\mathbf{P}}$ a matice pravděpodobností výstupu $\hat{\mathbf{B}}$. Provedeme znovu výpočet dopředného a zpětného algoritmu za pomoci následujících parametrů.

- Množina skrytých stavů $\mathbf{S}$:

$$\mathbf{S} = \{AT, GC\}.$$

- Nové počáteční pravděpodobnosti $\hat{\mathbf{p}}^{(0)}$:

$$\hat{\mathbf{p}}^{(0)} = \begin{array}{c} \begin{array}{cc} \text{AT} & \text{GC} \end{array} \\ \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \end{array}.$$

- Nová matice pravděpodobností přechodu $\hat{\mathbf{P}}$:

$$\hat{\mathbf{P}} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{c} \begin{array}{cc} \text{AT} & \text{GC} \end{array} \\ \begin{pmatrix} 0.602 & 0.398 \\ 0.097 & 0.903 \end{pmatrix} \end{array}.$$

- Nová matice pravděpodobností výstupu $\hat{\mathbf{B}}$:

$$\hat{\mathbf{B}} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} 0.411 & 0.145 & 0.144 & 0.300 \\ 0.136 & 0.397 & 0.389 & 0.078 \end{pmatrix} \end{array}.$$

- Posloupnost pozorování $\mathbf{O}_1$:

G	A	T	T	T	G	C	G	T	T	C	G	G	C	A	A	G	C	G	C	C
G	G	C	C	G	G	G	C	C	T	A	C	A	G	A	G	C	G	G	G	T
G	A	C	A	A	A	G	C	C	C	C	A	T	A	C	G	C	T	C	G	A
G	G	A	T	C	G	G	A	C	G	C	G	G	T	G	C	C	G	C	C	G
C	T	G	C	C	T	T	T	G	G	G	G	C	C	C	G	T	C	C	C	C
C	C	C	G	G	A	G	A	G	G	G	G	A	C	G	A	C	G	A	C	C
C	A	A	C	A	C	A	C	A	A	G										

Obrázek 3.12: Posloupnost pozorování $\mathbf{O}1$.

Dopředný algoritmus

Výše uvedené parametry nám slouží jako vstupní informace pro funkci `forward` v softwaru R, pomocí které spočítáme dopředný algoritmus. Obdržíme výslednou pravděpodobnost, jež činí cca $4.40e^{-78}$. Nyní ještě provedeme porovnání s výsledkem z přechozího výpočtu s původním modelem $\lambda(\mathbf{p}^{(\mathbf{0})}, \mathbf{P}, \mathbf{B})$, kde výsledná pravděpodobnost byla přibližně $3.21e^{-79}$. Hodnota výsledné pravděpodobnosti získané za pomocí nového modelu λ' je vyšší nežli hodnota výsledné pravděpodobnosti

získané prostřednictvím původního modelu λ.

Zpětný algoritmus

Zpětný algoritmus provedeme se stejnými parametry, avšak za pomocí funkce `backward` v softwaru R. Výsledná pravděpodobnost $4.40e^{-78}$ je totožná s pravděpodobností, kterou jsme získali prostřednictvím funkce `forward` u dopředného algoritmu.

Na základě obdržených výsledků lze konstatovat, že nový model $\lambda'(\hat{\mathbf{p}}^{(0)}, \hat{\mathbf{P}}, \hat{\mathbf{B}})$ má parametry nadefinované lépe oproti původnímu modelu $\lambda(\mathbf{p^{(0)}}, \mathbf{P}, \mathbf{B})$.

3.3.5. Dekódování s novými parametry

V této části práce najdeme za pomoci Viterbiho algoritmu optimální posloupnost skrytých stavů $\hat{\mathbf{X}}_1$ odpovídající první polovině posloupnosti pozorování $\mathbf{O}_1$. Následně také najdeme optimální posloupnost skrytých stavů $\hat{\mathbf{X}}_2$, která bude odpovídat druhé polovině posloupnosti pozorování $\mathbf{O}_2$. Stejně jako v předchozí podkapitole také zde využijeme nově získaných parametrů modelu $\lambda'(\hat{\mathbf{p}}^{(0)}, \hat{\mathbf{P}}, \hat{\mathbf{B}})$. Tyto parametry zadáme do funkce `viterbi` v softwaru R.

Pro spuštění funkce `viterbi` využijeme následující informace.

- Množina skrytých stavů $\mathbf{S}$:

$$\mathbf{S} = \{AT, GC\}.$$

- Nové počáteční pravděpodobnosti $\hat{\mathbf{p}}^{(0)}$:

$$\hat{\mathbf{p}}^{(0)} = \begin{array}{c} \begin{array}{cc} \text{AT} & \text{GC} \end{array} \\ \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \end{array}.$$

- Nová matice pravděpodobností přechodu $\hat{\mathbf{P}}$:

$$\hat{\mathbf{P}} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{c} \begin{array}{cc} \text{AT} & \text{GC} \end{array} \\ \begin{pmatrix} 0.602 & 0.398 \\ 0.097 & 0.903 \end{pmatrix} \end{array}.$$

- Nová matice pravděpodobností výstupu $\hat{\mathbf{B}}$:

$$\hat{\mathbf{B}} = \begin{array}{c} \\ \text{AT} \\ \text{GC} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} 0.411 & 0.145 & 0.144 & 0.300 \\ 0.136 & 0.397 & 0.389 & 0.078 \end{pmatrix} \end{array}.$$

- Posloupnost pozorování $\mathbf{O}_1$:

G	A	T	T	T	G	C	G	T	T	C	G	G	C	A	A	G	C	G	C	C
G	G	C	C	G	G	G	C	C	T	A	C	A	G	A	G	C	G	G	G	T
G	A	C	A	A	A	G	C	C	C	C	A	T	A	C	G	C	T	C	G	A
G	G	A	T	C	G	G	A	C	G	C	G	G	T	G	C	C	G	C	C	G
C	T	G	C	C	T	T	T	G	G	G	G	C	C	C	G	T	C	C	C	C
C	C	C	G	G	A	G	A	G	G	G	G	A	C	G	A	C	G	A	C	C
C	A	A	C	A	C	A	C	A	A	G										

Obrázek 3.13: Posloupnost pozorování **O**1.

- Posloupnost pozorování $\mathbf{O}_2$:

C	C	G	T	G	C	T	T	G	A	T	G	G	G	C	A	G	C	A	A	T
G	A	C	G	C	T	C	G	G	A	C	A	G	G	C	A	T	G	C	C	C
C	C	C	G	G	A	A	T	A	C	C	A	G	G	G	G	G	C	G	C	A
A	T	G	T	G	C	G	T	T	C	A	A	A	G	A	C	T	C	G	A	T
G	A	T	T	C	A	C	G	G	A	A	T	T	C	T	G	C	A	A	T	T
C	A	C	A	C	T	A	G	T	T	A	T	C	G	C	A	T	T	T	C	G
C	T	G	C	G	T	T	C	T	T											

Obrázek 3.14: Posloupnost pozorování **O**2.

Prostřednictvím Viterbiho algoritmu jsme získali posloupnost skrytých stavů $\hat{\mathbf{X}}_1$ (Obrázek 3.15) mající tvar:

AT	AT	AT	AT	AT	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	**GC**	**GC**	**GC**	GC	GC	GC	GC	GC	**GC**	**GC**	**GC**	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	AT	AT	AT	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC										

Obrázek 3.15: Posloupnost skrytých stavů $\hat{\mathbf{X}}1$.

Uvedená posloupnost skrytých stavů $\hat{\mathbf{X}}_1$ (Obrázek 3.15) se podobá posloupnosti skrytých stavů $\mathbf{X}_1$ (viz Obrázek 3.10), jež jsme obdrželi za pomocí původního modelu $\lambda(\mathbf{p}^{(\mathbf{0})}, \mathbf{P}, \mathbf{B})$. Odlišnost tvoří pouze šest skrytých stavů, které jsou v posloupnosti vyznačeny červenou barvou. Jedná se o nahrazení skrytých stavů *AT* skrytými stavy *GC*.

Na základě parametrů odhadnutých z první poloviny posloupnosti pozorování $\mathbf{O}_1$ jsme našli také posloupnost skrytých stavů $\hat{\mathbf{X}}_2$ (Obrázek 3.16), která odpovídá druhé polovině posloupnosti pozorování $\mathbf{O}_2$.

GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	AT	AT	AT	AT	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC	GC
GC	GC	GC	GC	GC	GC	GC	**GC**	**GC**	**GC**	**GC**	**GC**	**GC**	GC	GC	GC	GC	GC	GC	AT	AT
AT	AT	AT	AT	GC	GC	GC	GC	GC	AT	AT	AT	AT	GC	GC	GC	GC	AT	AT	AT	AT
GC	**GC**	**GC**	**GC**	**GC**	AT	AT	AT	AT	AT	AT	AT	GC	GC	GC	AT	AT	AT	AT	GC	GC
GC	GC	GC	GC	GC	AT	AT	AT	AT	AT											

Obrázek 3.16: Posloupnost skrytých stavů $\hat{\mathbf{X}}2$.

Posloupnost skrytých stavů $\hat{\mathbf{X}}_2$ (Obrázek 3.16) je podobná druhé polovině posloupnosti skrytých stavů $\mathbf{X}$, již jsme našli odhadem parametrů z celé sekvence (viz Obrázek 3.5). Neshoduje se pouze pár skrytých stavů. Skryté stavy *GC* v posloupnosti $\hat{\mathbf{X}}_2$, na jejichž místě byly v posloupnosti $\mathbf{X}$ skryté stavy *AT*, jsou opět vyznačeny červenou barvou.

Závěr

Diplomová práce si kladla za cíl seznámit čtenáře se skrytými Markovovými řetězci. Práce je zkompletována takovým způsobem, aby téma bylo co nejvíce pochopitelné i pro čtenáře, který se s touto tématikou dříve nesetkal.

V první části práce se nachází seznámení s pojmy, které jsou se skrytými Markovovými řetězci úzce spjaty. Mluvíme zde o náhodném procesu a Markovových řetězcích. V další části diplomové práce je popsán samotný princip skrytých Markovových řetězců. Následuje rozebrání úloh, které jsou řešeny právě s pomocí skrytých Markovových řetězců. Úlohy, jež jsou také někdy označovány jako problémy, se nazývají evaluace, dekódování a učení. Každý problém je řešitelný pomocí určitého algoritmu. Tyto algoritmy jsou v práci také popsány. Veškerá teorie je proložena příklady, které mají napomoci k lepšímu pochopení dané problematiky. Ve zbylé části práce jsou jednotlivé úlohy spolu s algoritmy aplikovány na reálná data z biologické sféry.

Velkou výzvou pro mne byla práce se softwary LaTeX a R, se kterými jsem do této doby pracovala pouze okrajově. Psaní práce bylo pro mě velice prospěšné a přínosné, jelikož jsem se s danou problematikou doposud nesetkala. Z poznatků, které jsem při psaní práce získala, si troufám říci, že dané téma je opravdu zajímavé a užitečné. Byla bych velice potěšena, pokud by tato diplomová práce posloužila zájemcům k pochopení probrané tématiky.

Literatura

[1] *BLAST – Basic Local Alignment Search Tool [online].* NCBI. [cit. 2020-06-09]. dostupné z: https://blast.ncbi.nlm.nih.gov/Blast.cgi.

[2] Himmelmann, L.: *Package 'HMM' [online].* 2010, [cit. 2020-03-17]. dostupné z: https://cran.r-project.org/web/packages/HMM/HMM.pdf.

[3] Hron, K., Kunderová, P.: *Základy počtu pravděpodobnosti a metod matematické statistity.* 2. dopl.vyd. Olomouc, Univerzita Palackého v Olomouci, 2015.

[4] Jelínek, J., Zicháček, V. *Biologie pro gymnázia: (teoretická a praktická část).* 9. vyd. Olomouc: Nakladatelství Olomouc, 2007.

[5] Kopka, H., Daly, P.: *LATEX: podrobný průvodce.* Computer Press, Brno, 2004.

[6] Lachout, P., Prášková, Z.: *Základy náhodných procesů.* Praha, Karolinum, 2001.

[7] *Latent Gold [online].* Statistical Innovations. [cit. 2020-06-07]. dostupné z: https://www.statisticalinnovations.com/latent-gold-5-1/.

[8] Leeflang, P.: *Advanced Methods for Modeling Markets.* New York, Springer, 2017.

[9] *Letian: Hidden Markov Chain.* [online]. 2019, [cit. 2020-01-21]. dostupné z: https://letianquant.com/hidden-markov-chain.html.

[10] Rabiner, L., Juang, B. H. *Fundamentals of speech recognition.* Englewood Cliffs, N.J.: PTR Prentice Hall, 1993.

[11] R Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017. dostupné z: https://www.R-project.org/.

[12] Rogalewicz, V.: *Stochastické procesy (Analýza časových řad). 1. vydání.* Praha, Vydavatelství ČVUT, 1993.

[13] Rybička, J.: *LATEX pro začátečníky (3. vydání).* Konvoj, Brno, 2003.

[14] Štípek, S.:*Stručná biochemie : uchování a exprese genetické informace : učební texty.* Praha: Medprint, 1997.

[15] Voet, D., Voet, J.G. *Biochemistry.* Victoria Publishing, 1995.

[16] Zucchini, W., MacDonald, I. L., Langrock, R.: *Hidden Markov Models for Time Series, An Introduction Using R (Second Edition).* Canada, CRC Press, 2016.