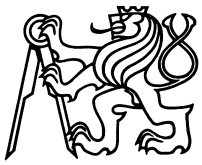
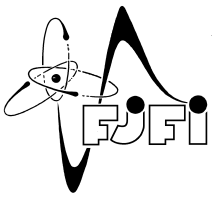


České vysoké učení technické v Praze
Fakulta jaderná a fyzikálně inženýrská



Těžba dat z experimentů na tokamaku COMPASS

Data mining on the COMPASS tokamak experiments

Bakalářská práce

Autor: **Matěj Zorek**

Vedoucí práce: **Ing. Vít Škvára**

Konzultant: **Ing. Jakub Urban, PhD**

Akademický rok: 2017/2018

- Zadání práce -

- Zadání práce (zadní strana) -

Poděkování:
Chtěl bych zde poděkovat především svému školiteli ..... za pečlivost, ochotu, vstřícnost a odborné i lidské zázemí při vedení mé diplomové práce. Dále děkuji svému konzultantovi ..... za .....

Čestné prohlášení:
Prohlašuji, že jsem tuto práci vypracoval samostatně a uvedl jsem všechnu použitou literaturu.

V Praze dne 2. července 2018 Matěj Zorek

Název práce:

Těžba dat z experimentů na tokamaku COMPASS

Autor: Matěj Zorek

Obor: Matematické inženýrství

Zaměření: Aplikované matematicko-stochastické metody

Druh práce: Bakalářská práce

Vedoucí práce: Ing. Vít Škvára, Ústav fyziky plazmatu, AV ČR Za Slovankou 1782/3 182 00 Praha 8

Konzultant: Ing. Jakub Urban, PhD., Ústav fyziky plazmatu, AV ČR, Za Slovankou 1782/3, 182 00 Praha 8

Abstrakt: Abstrakt max. na 10 řádků.

Klíčová slova: klíčová slova (nebo výrazy) seřazená podle abecedy a oddělená čárkou

Title:

Data mining on the COMPASS tokamak experiments

Author: Matěj Zorek

Abstract: Max. 10 lines of English abstract text.

Key words: keywords in alphabetical order separated by commas

Obsah

Úvod

Text úvodu....

Kapitola 1

Motivace

1.1 Strojové učení

1.2 Klasifikace

1.3 Rysy (Features)

Ve strojovém učení a rozpoznávání vzorů se pod pojmem "rys" (feature) rozumí individuální měřitelná vlastnost nebo charakteristika pozorovaného jevu. Výběr těchto rysů je naprosto zásadní pro efektivní rozpoznávací, regresní a klasifikační algoritmy. Čím relevantnější a charakterističtější rys, tím lépe jsme schopni docílít větší přesnosti modelu. Na druhou stranu vynechání zbytečných, případně méně důležitých rysů zase snižuje složitost modelu a urychluje jeho trénink. Nejčastější forma rysu je číselná hodnota, avšak při rozpoznávání syntetických vzorků se hojně používají i písmena, slova nebo grafy.

Selekci těchto rysů je možné demonstrovat na následujícím příkladu. Předpokládejme, že bychom chtěli předvídat typ domácího mazlíčka, jež si někdo koupí.

Do rysů můžeme zahrnout například věk osoby, pohlaví, jméno, bydlení (byt, dům, ...), rodinný příjem, vzdělání a počet dětí. Je zřejmě, že většina těchto rysů nám může při předvídání pomoci, ale některé jako třeba vzdělání nebo jméno jsou zjevně méně důležité.

jméno	věk	pohlaví	bydlení	příjem	počet dětí	vzdělání
Karel	25	muž	byt	30.000	0	středoškolské
Petr	30	muž	dům	45.000	2	vysokoškolské
Jana	42	žena	byt	23.000	1	základní
Miloš	51	muž	dům	29.000	1	středoškolské
			...			

Tabulka 1.1: Vzorová tabulka rysů k demonstračnímu příkaldu

1.3.1 První a druhá derivace

Prvním rysem použitým při klasifikaci je první derivace. Jelikož jsou k dispozici pouze jednotlivé body, nelze použít analytický vzorec pro derivace funkce $f(x) : R \longrightarrow R$, tedy konkrétně

$$\frac{df(x)}{dx} = f'(x) = \lim_{h->0} \frac{f(x+h) - f(x)}{h}. \tag{1.1}$$

Namísto toho se musím počítat numericky, a to za použití centrální diference druhého řádu pomocí (1.2) a v krajních bodech pomocí jednostranných diferencí prvního nebo druhého řádu (1.3).

$$\hat{f}'_k = \frac{f(x_{k+1}) - f(x_{k-1})}{2h} \tag{1.2}$$

$$\hat{f}'_0 = \frac{f(x_1) - f(x_0)}{h} \text{ a } \hat{f}'_n = \frac{f(x_n) - f(x_{n-1})}{h} \tag{1.3}$$

Dalším použitým rysem je druhá derivace, kterou lze jesnáze získat použitím výše zmíněncým vzoruců (1.2) a (1.3) na již jednou zderivovaný signál.

1.3.2 Klouzavý průměr

Dalším vybraným rysem je klouzavý průměr. Klouzavý průměr je diskrétní lineární filtr s konečnou dobou odezvy, který slouží k vyhlazení signálu. Nadále ho budeme značit jako $\widetilde{X}_t$. Nechť máme vektor naměřených hodnot $\boldsymbol{X} = (x_1, x_2, ..., x_n)$, pak definuji klouzavý průměr pro $t \in 1, 2, ..., n$ jako

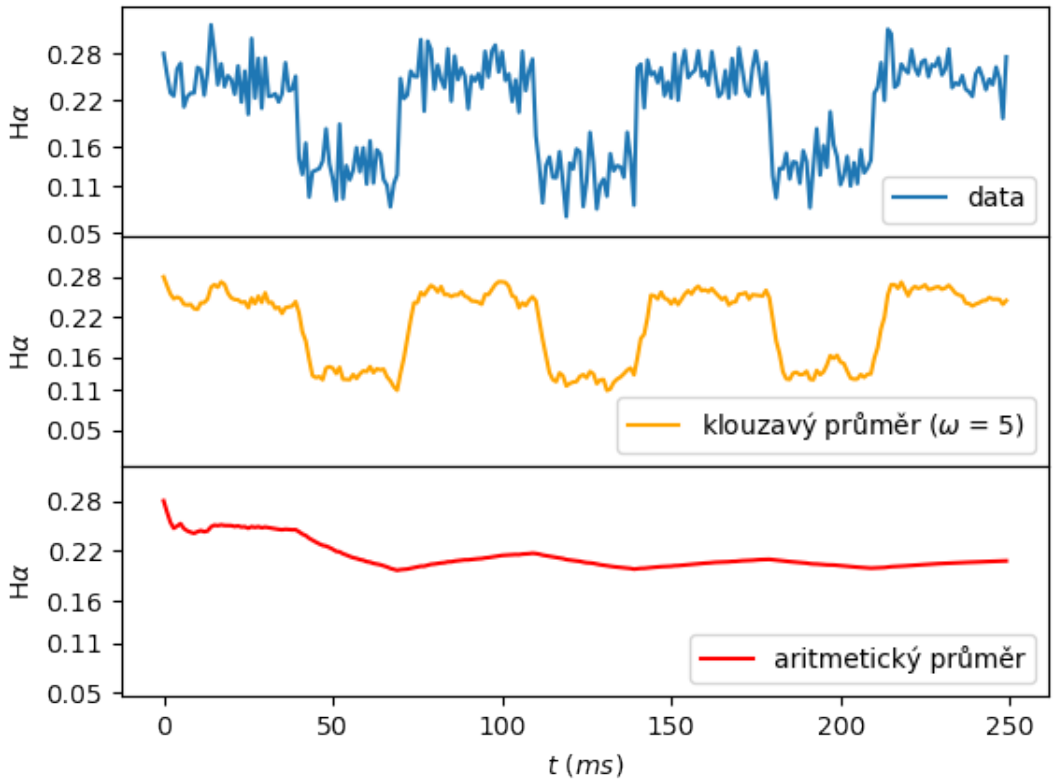
$$\widetilde{X}_t = \frac{1}{\tilde{\omega}} \sum_{k=t-\tilde{\omega}}^{t} x_k, \tag{1.4}$$

kde $\tilde{\omega} = min\{\omega, t\}$, přičemž ω je délka úseku. Pak vektor $\widetilde{\boldsymbol{X}} = (\widetilde{X}_1, \widetilde{X}_2, .., \widetilde{X}_n)$ je rys vektoru $\boldsymbol{X}$.

Ve skutečnosti se jedná o standartní aritmetický průměr (1.5), jež je aplikovaný pouze na úsek dat konečné délky.

$$\overline{X}_n = \frac{1}{n} \sum_{k=1}^{n} x_k, \tag{1.5}$$

Mezi rysy byl vybrán, protože předpokládáme, že okamžitá hodnota je závislá na předchozích datech. Důvod, proč využíváme jen konečně dlouhý úsek předcházejících hodnot je ten, že ze zákona velkých čísel aritmetický průměr konverguje ke střední hodnotě, a tedy ke konstantě. To znamená, že postupem času budou mít rozdílná data v odlišných stavech stejnou hodnotu rysu. Z čehož plyne, že takovýto rys by jen zkresloval a znepřesňoval výsledek, viz Obr. 1.1.



Obr. 1.1: Rozdíl mezi klouzavým a aritmerickým průměrem aplikovaným na syntetická data

1.3.3 Exponenciální klouzavý průměr

Jíž dříve jsme se zmínili, že předpokládáme závislost na předchozích hodnotách. Nicméně je zřejmé, že hodnoty naměřené s velkým časovým rozestupem na sebe mají mnohem menší vliv, než ty jež jsou naměřeny bezprostředně zasebou. Proto dalším vybraným rysem je tedy exponenciální klouzavý průměr S_t.

Nechť $\boldsymbol{X} = (x_1, x_2, ..., x_n)$ je vektor naměřených hodnot, pak definuji váhový součet zleva pro $t \in 1, 2, ..., n$ jako

$$S_t = \frac{1}{\tilde{\omega}} \sum_{k=t-\tilde{\omega}}^{t} g_{t-k} \cdot x_k, \tag{1.6}$$

kde $\tilde{\omega} = min\{\omega, t\}$, ω je délka úseku a g_m je váhová funkce tvaru $g_m = \gamma^m$, přičemž $\gamma \in (0, 1)$. Pak vektor $\boldsymbol{S} = (S_1, S_2, .., S_n)$ je rys vektoru $\boldsymbol{X}$. Díky exponenciální váhové funkci g_m, jsme tedy schopni snížit důležitost více vzdálených dat, což byl náš záměr.

1.3.4 Klouzavý rozptyl

Ve statistice a teorii pravděpodobnosti se pod pojmem rozptyl rozumí střední hodnota kvadrátu odchylky od střední hodnoty náhodné veličiny. Bývá reprezentován symbolem $Var(X)$ nebo σ^2 a definován

vzorcem

$$Var(X) = E[(X - E[X])^2] \tag{1.7}$$

pro stejně rozdělené diskrétní náhodné veličiny $\boldsymbol{X} = (x_1, x_2, ..., x_n)$ můžeme tento vzorec přepsat do tvaru

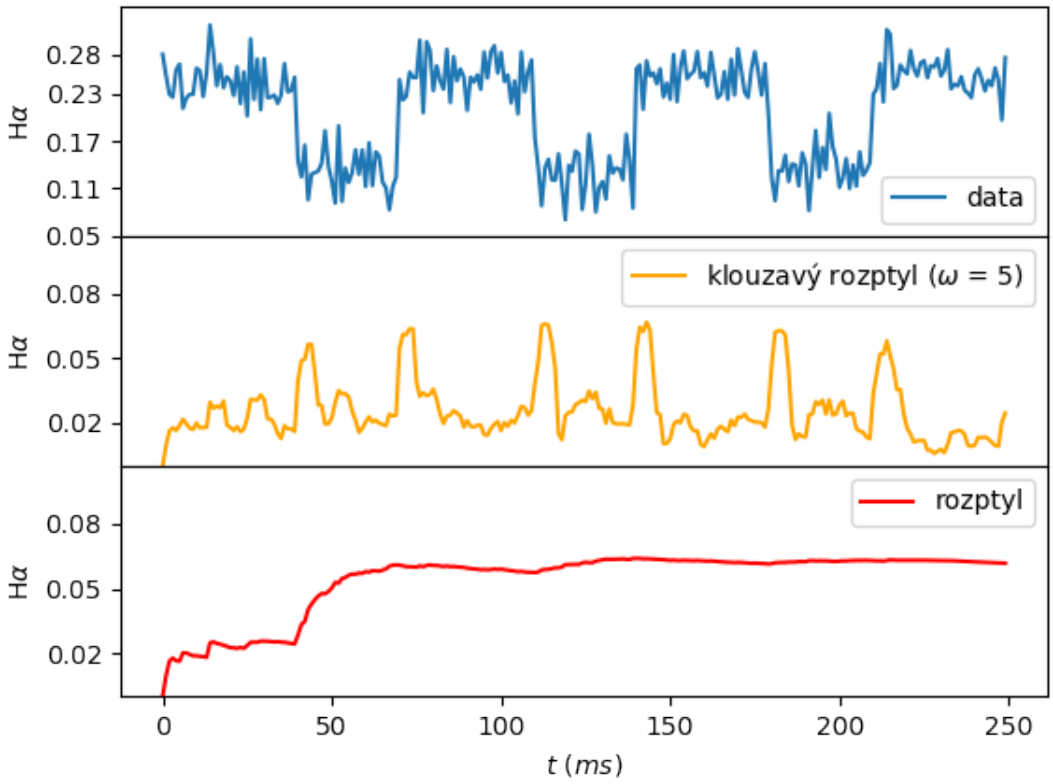
$$Var(X) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{X}_n)^2. \tag{1.8}$$

Bohužel, rozptyl nemůžeme jako rys použít ze stejného důvodu jako aritmetický průměr, protože s postupem času bude různým stavům přiřazovat stejnou hodnotu. Proto zed využijeme místo aritmetického průměru již dříve definovaný klouzavý průměr a výslednou veličinu budu dále nazývat klouzavým rozptylem a značit D_t.

Nechť $\boldsymbol{X} = (x_1, x_2, ..., x_n)$ je vektor naměřených hodnot, pak definuji úsekový rozptyl pro $t \in 1, 2, ..., n$ jako

$$D_m = \frac{1}{\tilde{\omega}}\sum_{k=t-\tilde{\omega}}^{t}(x_k - \widetilde{X}_t)^2, \tag{1.9}$$

kde $\tilde{\omega} = min\{\omega, t\}$, w je opět délka úseku a $\widetilde{X}_t$ je klouzavý průměr (1.4). Pak vektor $\boldsymbol{D} = (D_1, D_2, ..., D_n)$ je rys vektoru $\boldsymbol{X}$.



Obr. 1.2: Rozdíl mezi úsekovým a normálním rozptylem aplikovaným na syntetická data

Kapitola 2

Teoretická část

2.1 Skrytý Markovův model

Skrytý Markovův model známý spíše pod svým anglickým názvem "Hidden Markov model" je statistický model, který slouží k modelování Markovských procesů se skrytými stavy. Přesněji se jedná o dvojnásobně zapuštěný stochastický proces podložený dalším stochstickým procesem, který není pozorován, je tedy skrytý. Nicméně tento skrytý porces může být pozorován skrze jiné stochastické procesy, jež poskytují posloupnost pozorování.

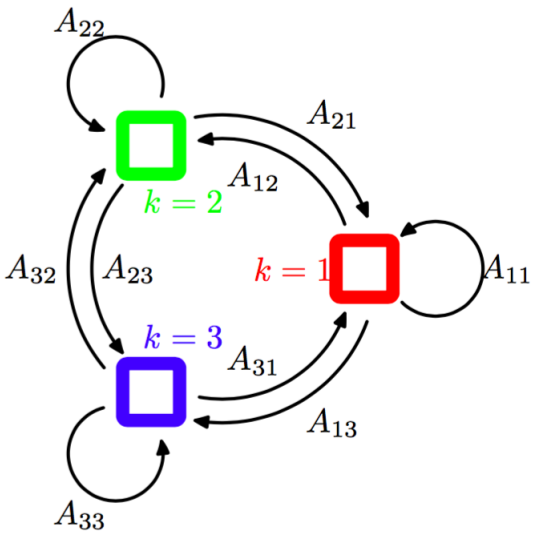
Skrytý Markovův model je široce používán v rozpoznávání řeči (speech recognition), modelování přirozeného jazyka, rozpoznávání ručněpsaného písma a analýza biologických sekvencí jako například DNA a protejnů.

Pro lepší představu si tento model ukážeme na příkladě uren a míčků. "Předpokládejme, že v místnosti je N velkých skleněných uren. V každé urně je velký počet barevných míčků. Předpokládejme, že máme M odlišných barev míčků. Fyzikální proces pro získání pozorování je následující. Džin je v místnosti a on (nebo ona) podle nějakého náhodného procesu vybírá počáteční urnu. Z této urny vybere náhodně míček a jeho barva je nahrána jako pozorování. Míček je pak nahrazen v urně, z níž byl vybrán. Další urna je vybrána porocesem náhodného výběru spojeného se současnou urnou a výběr míčku je opakován. Celý proces generuje konečný počet pozorování posloupnosti barev, který bychom rádi modelovaly jako pozorovaný výstup skrytého markovského modelu."[3]

Dále se na tento model můžeme dívat jako na specifický případ stavového prostorového modelu, ve kterém jsou skryté proměnné diskrétní. Nicméně když prozkoumáme "jednorázový"řez modelu, vidíme že odpovídá Mixture distribution [1] s hustotou pravděpodobnosti danou $\mathbb{P}(\boldsymbol{x}|\boldsymbol{z})$. Proto ho můžeme interpretovat jako rozšíření Mixture modelu, kde výběr složky směsi, pro každé pozorování, není nezávislý, ale závisí na volbě složek z předchozího pozorování.

Jako v případě standartního Mixture modelu, skryté proměnné jsou diskrétní multinomické proměnné $\boldsymbol{z}_n$ popisující složku směsi jež je zodpovědná za generování příslušného pozorování $\boldsymbol{x}_n$. Od této chvíle budeme předpokládat, že skrytý proces je Markovův, tzn. splňuje Markovu vlastnost (Markov property)[4]. Z tohoto předpokladu nyní plyne, že budoucí stav skryté proměnné $\boldsymbol{z}_n$ závisí pouze na předchozím stavu $\boldsymbol{z}_{n-1}$ skrze podmíněnou pravděpodobnost $\mathbb{P}(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})$. Pak zavedeme matici přechodů $\mathbb{T}$ danou předpisem $T_{i,j} \equiv \mathbb{P}(\boldsymbol{z}_{n,j}|\boldsymbol{z}_{n-1,i} = 1)$, navíc matice splňuje, že $T_{i,j} \in (0, 1)$ a skoupce jsou normovány na 1, tzn $\sum_j T_{i,j} = 1$. Dále můžeme tedy explicitně napsat podmíněné rozdělení pro K skrytých stavů ve tvaru

$$\mathbb{P}(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}, \mathbb{T}) = \prod_{i=1}^{K} \prod_{j=1}^{K} T_{i,j}^{z_{n-1,i}z_{n,j}}. \tag{2.1}$$



Obr. 2.1: Diagram přechodů, kde skryté proměnné mají tři možné stavy odpovídající třem boxům. Šipky označují prvky matice přechodů $T_{i,j}$. [1]

Tímto vzorcem můžeme vyjádřit všechny skryté stavy až na počáteční z_1. Tento stav má pouze marginální rozdělení $\mathbb{P}(z_1)$ reprezentované vektorem pravděpodobností π, kterýž má tvar $\pi \equiv \mathbb{P}(z_{1,j} = 1)$ a tedy

$$\mathbb{P}(z_1|\pi) = \prod_{j=1}^{K} \pi_j^{z_{1,j}}, \tag{2.2}$$

kde $\sum_j \pi_j = 1$. Kdybychom chtěli matici $\mathbb{T}$ vysvětlit i jinak, mohli bychom toho docílit graficky pomocí diagramu na Obr. 2.1. Když tento diagram dále rozvyneme s průběhem času, získáme takzvaný mřížový diagram, který nám poskytuje alternativní reprezentaci přechodů mezi jednotlivými skrytými stavy. Pro případ Skritého Markovského modelu pak tento diagram nabývá tvaru Obr. 2.2.

Abychom měli pravděpodobnostní model kompletní, je třeba ještě zavést emisní pravděpodobnosti (emission probabilities) $\mathbb{P}(\boldsymbol{x}_n|\boldsymbol{z}_n, \Theta)$, kde Θ představuje soubor parametrů řídícího rozdělení. Pro pozorované proměnné $\boldsymbol{x}_n$ se rozdělení $\mathbb{P}(\boldsymbol{x}_n|\boldsymbol{z}_n, \Theta)$ skládá z K-dimenzionálního vekoru odpovídajícího K potenciálním stavům $\boldsymbol{z}_n$. Emisní pravděpodobnosti můžeme pak zapsat jako

$$\mathbb{P}(\boldsymbol{x}_n|\boldsymbol{z}_n, \Theta) = \prod_{j=1}^{K} \mathbb{P}(\boldsymbol{x}_n|\Theta_j)^{z_{n,j}}, \tag{2.3}$$

přičemž mohou mít například tvar Gaussova (R-rozměrného) rozdělení

$$\mathbb{P}(\boldsymbol{x}_n|\boldsymbol{z}_n) = \prod_{j=1}^{K} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \Sigma_j)^{z_{n,j}} = \prod_{j=1}^{K} \left(\frac{1}{(2\pi)^{R/2}} \frac{1}{|\Sigma_j|^{1/2}} exp\left\{ -\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_j)^T \Sigma_j^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_j) \right\} \right)^{z_{n,j}} \tag{2.4}$$

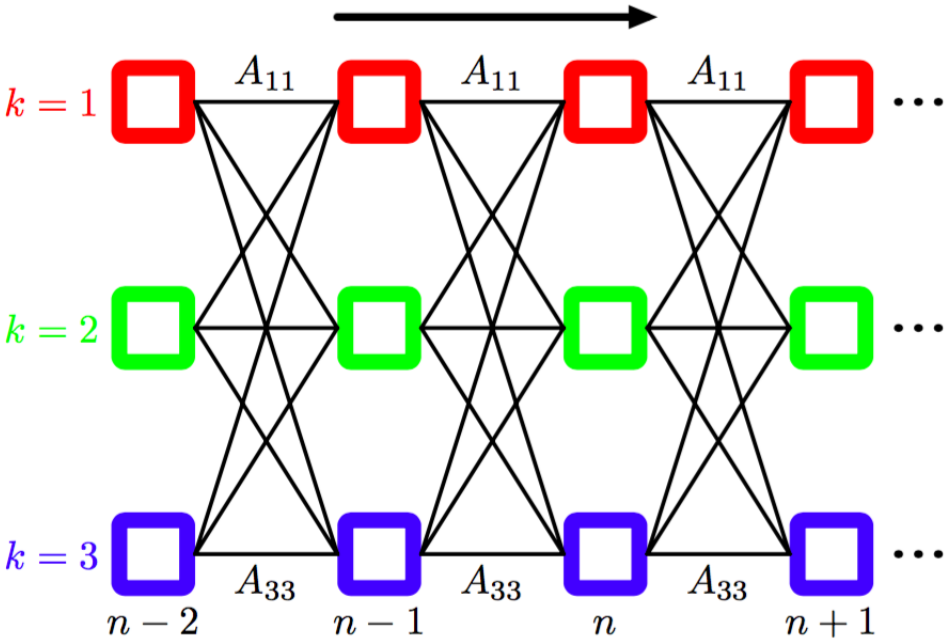
pakud prvky $\boldsymbol{x}_n$ jsou spojité.

Nyní se zaměříme na homogenní modely, pro které všechna podmíněná rozdělení řídící skryté proměnné sdílí stejné parametry $\mathbb{T}$ a podobně všechna emisní rozdělení sdílí stejné parametry Θ.

Sdružené pravděpodobnostní rozdělení přes skryté i pozorované proměnné jsou pak dány vzorcem

$$\mathbb{P}(\boldsymbol{X}, \boldsymbol{Z}|\widetilde{\boldsymbol{\Theta}}) = \mathbb{P}(z_1|\pi) \prod_{n=2}^{N} \mathbb{P}(z_n|z_{n-1}, \mathbb{T}) \prod_{m=1}^{N} \mathbb{P}(\boldsymbol{x}_m|z_m, \Theta), \tag{2.5}$$

kde $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_N)$, $\boldsymbol{Z} = (z_1, ..., z_N)$ a $\widetilde{\boldsymbol{\Theta}} = (\pi, \mathbb{T}, \Theta)$ jsou parametry řídícího modelu.



Obr. 2.2: Stavový diagram Obr. 2.1 rozvynutý s časem. Každý sloupec diagramu odpovídá jedné skryté proměnné z_n. [1]

2.1.1 Viterbiho algoritmus

2.2 K-means Clustering

K-means clustering je typ strojového učení bez učitele, který se používá v případě, že chceme zpracovat neoznačená data, tzn. nemáme předem definované skupiny či kategorie. Právě hlavním cílem algoritmu je najít tyto skupiny.

Předpokládejme, že máme N pozorování $(\boldsymbol{x}_1, ..., \boldsymbol{x}_N)$ náhodné R rozměrné veličiny $\boldsymbol{X}$ a chceme je rozdělit do nějakých K shluků. Pod pojmem shluk budeme rozumět skupinu bodů, jejichž vzájemné vzdálenosti jsou mnohem menší v porovnání se vzdálenostmi k bodům vně skupiny. Dále je pak potřeba zavést si vektory $\boldsymbol{\phi}_j$, kde $j \in 1, .., K$. Tyto vektory představují středy našich shluků a naším cílem se nyní stává nalezení množiny $\{\boldsymbol{\phi}_j\}$, tak aby bylo splněno, že součet čtverců vzdáleností každého bodu shluku k nejbližšímu vektoru $\boldsymbol{\phi}_j$ je minimální. Jinými slovy, potřebujeme minimalizovat objektovou funkci ρ definovanou jako

$$\rho = \sum_{n=1}^{N} \sum_{j=1}^{K} b_{n,j} \, \|\boldsymbol{x}_n - \boldsymbol{\phi}_j\|^2, \tag{2.6}$$

kde proměnné $b_{n,j} \in \{0, 1\}$ příslušící každému bodu $\boldsymbol{x}_n$ indikují, zda tento bod patří j-tého shluku nebo nikoli. Pokud ano, $b_{n,j}$ je rovno 1, v opačném případě nabývá $b_{n,j}$ hodnoty 0. Minimálního ρ lze dosáh-

nout pomocí dvoufázového iteračního procesu. Nejprve vybereme, nejlépe náhodně, počáteční vektory $\boldsymbol{\phi}_j$. V první fázy bereme $\boldsymbol{\phi}_j$ jako fixní a minimalizovat budeme s ohledem na $b_{n,j}$. Jelikož ρ je vůči $b_{n,j}$ lineární a pro rozdílná n jsou $b_{n,j}$ nezávislé, můžeme je minimalizovat pro každé n zvlášť. Jednoduše bereme $b_{n,j}$ následně

$$b_{n,j} = \begin{cases} 1 & \text{pokud } argmin_m \|\boldsymbol{x}_n - \boldsymbol{\phi}_m\|^2 = j \\ 0 & \text{jinde.} \end{cases} \tag{2.7}$$

V druhé fázi je naopak $b_{n,j}$ pevné a minimalizujeme s ohledem na $\boldsymbol{\phi}_j$. Proto zderivujeme funkci ρ podle $\boldsymbol{\phi}_j$ a tuto derivací položíme rovnu 0, tzn.

$$2\sum_{n=1}^{N} b_{n,j}(\boldsymbol{x}_n - \boldsymbol{\phi}_j) = 0. \tag{2.8}$$

Načež po osamostatnění $\boldsymbol{\phi}_j$. získáváme konečný tvar

$$\boldsymbol{\phi}_j = \frac{\sum_{n=1}^{N} b_{n,j}\boldsymbol{x}_n}{\sum_{n=1}^{N} b_{n,j}}, \tag{2.9}$$

kde jmenovatel představuje celkový počet bodů patřících do j-tého shluku. Vzorec (2.9) lze ovšem interpretovat také jako střední hodnotu (anglicky: mean) všech bodů příslušících do shluku j, odtud plyne i název "K-means".

přidat demonstrační obrázek

2.3 kNN

2.4 Autoregresní model

Kapitola 3

Postup práce a výsledky

3.1 Způsoby vyhodnocení výsledků

V této kapitole se budu věnovat způsobům jak ohodnotit kvalitu (přsnost) mého algoritmu.

3.1.1 Přesnost

3.1.2 Správnost

3.1.3 Recall

3.1.4 F míra

Závěr

Text závěru....

Literatura

[1] C. M. Bishop, Pattern recognition and machine learning. Springer, New York, 2013.

[2] M. Kikuchi, K. Lackner, M. Q. Tran, Fusion physics. International Atomic Energy Agency, Vienna, 2012.

[3] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 1989, 257-286.

[4] N. Privault, Understanding Markov Chains: Examples and Applications. Springer, 2013.

[5] G. D. Forney, The Viterbi Algorithm. Proceedings of the IEEE 61(3) 1973, 268-278.

[6] P. Harrington, Machine Learning in Action. Minning Publications Co. Greenwich, 2012.