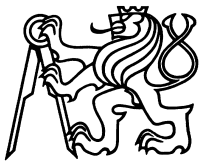
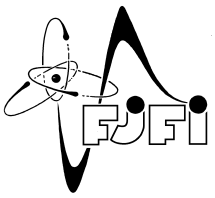


České vysoké učení technické v Praze
Fakulta jaderná a fyzikálně inženýrská



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

Teoretická část

1.1 HMM

1.2 K-means

1.3 Autoregresní model

1.4 Specifické 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	základoškolské
Miloš	51	muž	dům	29.000	1	středoškolské
			...			

Tabulka 1.1: Vzorová tabulka rysů k demonstračnímu příkaldu

1.4.1 První a druhá derivace

Jedním ze dvou rysů, jež jsem převzal z již existujícího kódu, je první derivace. Jelikož mám pouze jednotlivé body nemohu používat analytický vzorec na derivace, tedy konkrétně

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

Namísto toho ji 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).

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

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

Druhým převzatým rysem je druhá derivace, kterou získáme použitím vzorců pro první derivaci na již jednou zderivovaná data.

1.4.2 Úsekový aritmetický průměr

Prvním mnou vybraným rysem je úsekový aritmetický průměr naměřených dat, který budu nadále značit jako $\widetilde{X}_m$. Nechť máme vektor naměřených hodnot $\boldsymbol{X} = (x_1, x_2, ..., x_n)$, pak definuji úskový aritmetický průměr pro $m \in 1, 2, ..., n$ jako

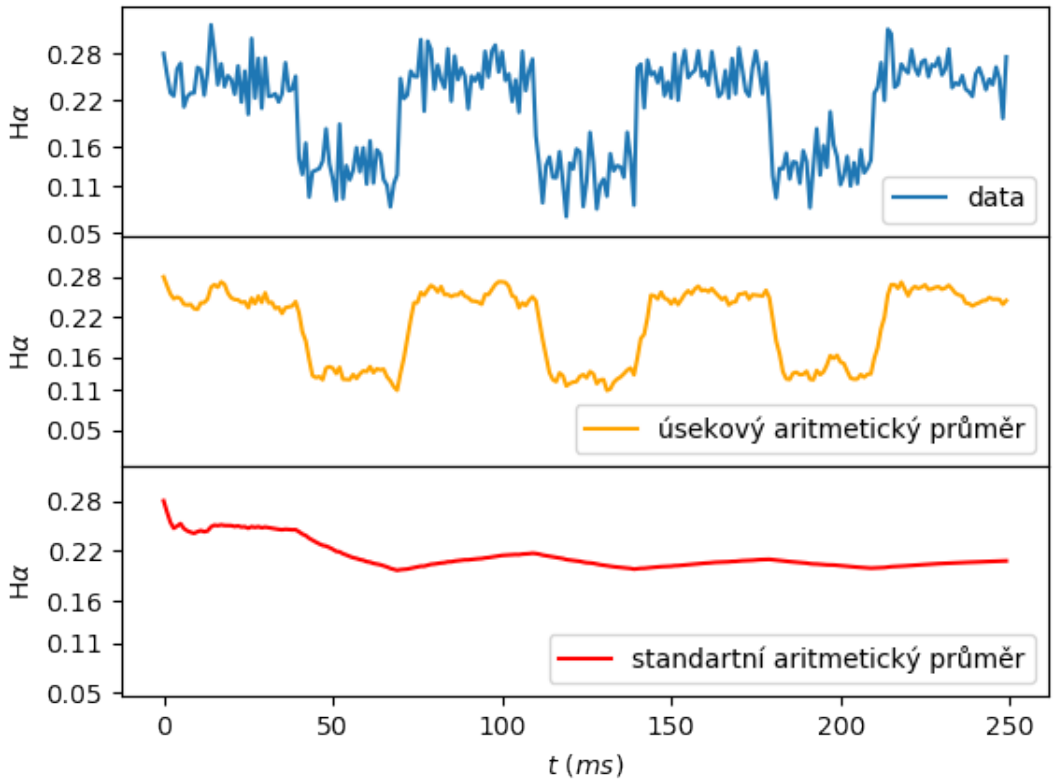
$$\widetilde{X}_m = \frac{1}{w} \sum_{k=m-w}^{m} x_k, \tag{1.4}$$

kde w je délka úseku. A 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 jsem ho vybral, protože předpokládám, že okamžitá hodnota je závislá na předchozích datech. Důvod proč využívám jen konečně dlouhý úsek předcházejících hodnot je ten, že ze zákona velkých čísel (standartní) 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 úsekovým a standartním aritmerickým průměrem aplikovaným na syntetická data (pro větší přehlednost je každá krivka zobrazena zvlášť)

1.4.3 Váhový součet zleva

Jak jsem se zmínil již dříve, předpoládám závislost na předchozích hodnotách. "Není však překvapením", že hodnoty naměřené s velkým časovým rozestupem na sebe mají mnohem menší vliv, než ty jež jsou naměřeny bezprostředně zasebou. Proto dalším mnou vybraným rysem je tedy váhový součet zleva S_m.

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

$$S_m = \frac{1}{w} \sum_{k=m-w}^{m} g_{m-k} \cdot x_k, \tag{1.6}$$

kde w je délka úseku a g_m je váhová funkce tvaru $g_m = \gamma^m$, přičemž $\gamma \in (0, 1)$. A vektor $\boldsymbol{S} = (S_1, S_2, .., S_n)$ je rys vektoru $\boldsymbol{X}$.

1.4.4 Rozptyl

Rozptyl asi vynechám protože je k ničemu

$$\sigma_n^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - E(x))^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \overline{X}_n)^2 \tag{1.7}$$

1.4.5 Úsekový rozptyl

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

$$Var(X) = E[(X - E[X])^2] \tag{1.8}$$

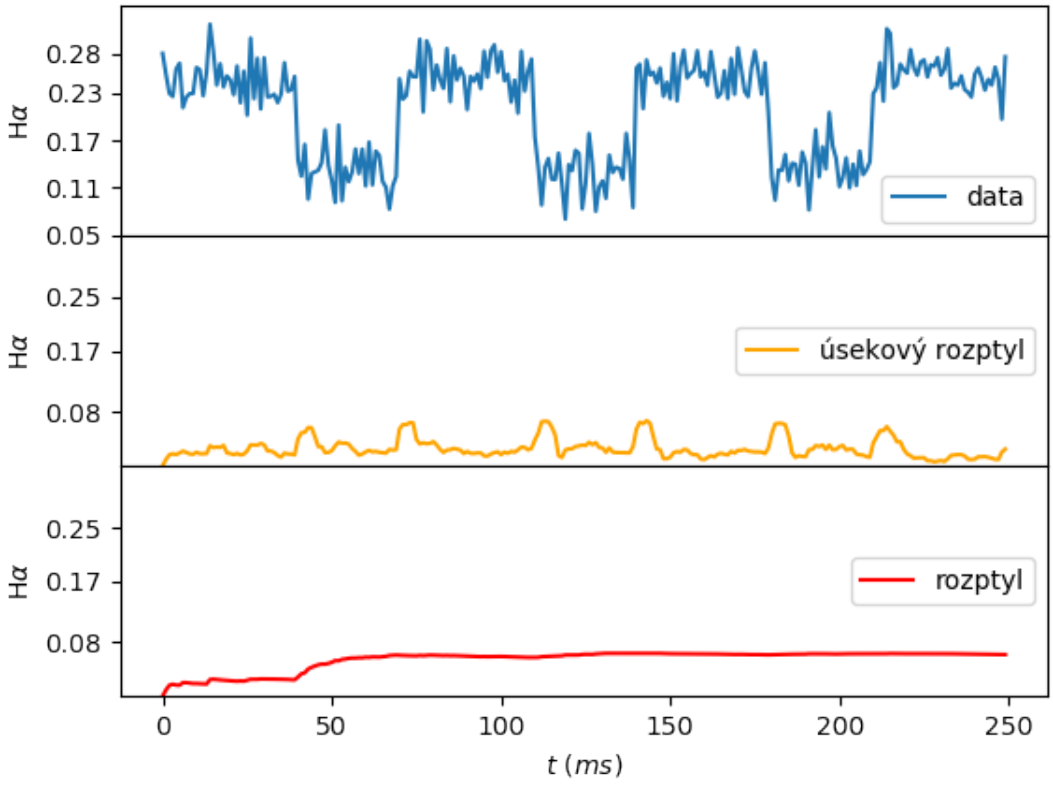
pro stejně rozdělené diskrétní náhodné veličiny $X = (x_1, x_2, ..., x_n)$ můžeme tento vzorec přepsat do tvaru

$$Var(X) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{X}_n)^2. \tag{1.9}$$

Když jsem se začínal vybírat rys, které budu potom používat při klasifikaci, byl rozptyl jedna z mých prvních voleb. Naneštěstí má stejný problém jako standartní skalární součin a to ten, že s postupem času bude různým stavům přiřazovat stejnou hodnotu. Proto jsem se rozhodl využít místo standartního skalárnho součinu již dříve definovaný úsekový skalární součin a výslednou veličinu jsem označil jako úsekový rozptyl. Úsekový rozptyl budu nadále značit jako D_m a pro náhodné veličiny $X = (x_1, x_2, ..., x_n)$ ho definuji jako

$$D_m = \frac{1}{w}\sum_{k=m-w}^{m}(x_k - \widetilde{X}_m)^2, \tag{1.10}$$

kde w je opět délka úseku a $\widetilde{X}_m$ je úsekový aritmetický průměr viz (1.4).



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

Kapitola 2

Způsoby vyhodnocení výsledků

V této kapitole se budu věnovat způsobům jak ohodnotit kvalitu (přsnost) mého algoritmu. "Vždy je potřeba nějak ohodnotit kvalitu našeho výsledku."

2.1 Přesnost

2.2 Správnost

2.3 Recall

2.4 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.