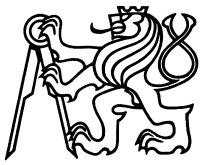
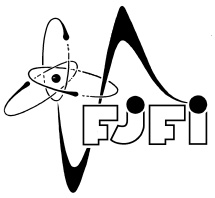


České vysoké učení technické v Praze
Fakulta jaderná a fyzikálně inženýrská



Generativní modely dat popsaných stromovou strukturou

Generative models of tree structured data

Bakalářská práce

Autor: **Jakub Bureš**

Vedoucí práce: **Doc. Ing. Václav Šmídl, Ph.D.**

Konzultant: **Doc. Ing. Tomáš Pevný, Ph.D.**

Akademický rok: 2019/2020

1. Seznamte se s popisem dat pomocí stromové struktury. Zvláštní pozornost věnujte metodám více instančního učení (multiple instance learning). Seznamte se s konceptem vnořeného prostoru (embedded space) a jeho reprezentace pomocí neuronových sítí.

2. Seznamte se se základními generativními modely dat popsaných vektorem příznaků. Zvláštní pozornost věnujte metodám typu autoencoder a jejich variační formě. Demonstrujte vlastnosti modelů na jednoduchých příkladech. V maximální míře využijte dostupné knihovny pro generativní modely.

3. Navrhněte několik příkladů typů dat se stromovou strukturou a pro každý z nich navrhněte generativní model. Navrhněte algoritmus pro určení jeho parametrů z dat a diskutujte vhodnost jednotlivých architektur neuronových sítí.

4. Seznamte se s různými druhy apriorních rozložení používaných na latentní proměnné autoencoderu. Odvoďte algoritmy odhadu jejich parametrů a srovnejte jejich výsledky se základním modelem. Diskutujte výsledné odhady.

5. Vyvinutou metodu aplikujte na vhodně zvolená reálná data a diskutujte vliv zvoleného apriorního rozložení na výsledky.

- Zadání práce (zadní strana) -

Poděkování:
Chtěl bych zde poděkovat především svému školiteli panu Doc. Ing. Václavu Šmídlovi, Ph.D. za pečlivost, ochotu, vstřícnost a odborné i lidské zázemí při vedení mé bakalářské práce. Dále děkuji svému konzultantovi panu Doc. Ing. Tomáši Pevnému, Ph.D.

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

V Praze dne 7. července 2020 Jakub Bureš

Název práce:

Generativní modely dat popsaných stromovou strukturou

Autor: Jakub Bureš

Obor: Matematické inženýrství

Zaměření: Aplikované matematicko-stochastické metody

Druh práce: Bakalářská práce

Vedoucí práce: Doc. Ing. Václav Šmídl, Ph.D.
ÚTIA AV ČR Pod vodárenskou věží 4 182 00 Praha 8

Konzultant: Doc. Ing. Tomáš Pevný, Ph.D.
Katedra počítačů FEL ČVUT Praha Technická 1902/2 166 27 Praha 6 - Dejvice

Abstrakt: Abstrakt max. na 10 řádků.

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

*Title:*Generative models of tree structured data

Generative models of tree structured data

Author: Jakub Bureš

Abstract: Max. 10 lines of English abstract text.

Key words: keywords in alphabetical order separated by commas

Obsah

Kapitola 1

Teorie

1.1 Optimalizace

Optimalizace je matematická úloha, jejíž snahou je nalezení takových hodnot proměnných, pro které daná funkce nabývá minima či maxima. My se budeme snažit najít minimální hodnoty parametrů θ tzv. ztrátové funkce, kterou budeme značit $L(\theta)$ z anglického výrazu loss function. Minimalizací ztrátové funkce získáme

$$\hat{\theta} = \arg\min_{\theta} L(\theta) \tag{1.1}$$

Existuje nespočet způsobů jak danou funkci minimalizovat. My budeme výhradně používat metodu zvanou Gradient Descent.

1.1.1 Gradient Descent

Jedná se iterativní optimalizační metodu. Minimalizujeme $L(\theta)$, tedy derivujeme dle vektoru parametrů θ, díky čemuž dostaneme $\nabla_\theta L(\theta)$. Symbol ∇_θ značí gradient funkce $L(\theta)$ přes všechny hodnoty θ. Použijeme bod θ_0 funkce $L(\theta)$ jako výchozí bod, ze kterého se pohybujeme ve směru záporného gradientu s krokem $h \in \mathbb{R}_+$. Matematickou interpretaci toho postupu můžeme vyjádřit následujícím zápisem:

$$\theta_{n+1} = \theta_n - h \cdot \nabla_\theta L(\theta_n) \tag{1.2}$$

Tento postup provádíme, dokud nejsme v minimu funkce a získáme tak vektor parametrů $\hat{\theta}$, jak je popsáno v (1.1).

ADAM

Předchozí metoda není při větším množství dat tak rychlá, jak bychom pro výpočet minima ztrátové funkce potřebovali. Používáme proto adaptivní iterační gradientní metodu ADAM (Adaptive Moment Estimation), která navíc používá druhý moment gradientu. Zatímco Gradient descent má krok stále stejný, u metody ADAM je krok h adaptivní. Popřípadě můžeme ladit i zapomínací koeficienty, což už je ale mimo rámec této práce a my nebudeme při výpočtech využívat.

1.1.2 Metoda nejmenších čtverců

Předpokládejme že máme množinu $x = \{x_i\}_{i=1}^n$ ke každému x_i máme právě jedno pozorování y_i, komplexně zapsáno zobrazením jako $(x_1, \ldots, x_n) \longmapsto (y_1, \ldots, y_n)$. Naším cílem je najít nejlepší proložení

dat, čili fit, pomocí polynomické funkce řádu $p \leq n$ ve tvaru

$$\hat{y}(x, \theta) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_p x^p = \sum_{i=0}^{p} \theta_i x^i,$$

která je lineární v neznámých parametrech $\theta = \left(\theta_0, \theta_1, \ldots, \theta_p\right)$. Takové modely nazýváme lineární a jejich vlastnosti budeme nadále využívat.
Abychom našli ten nejlepší možný fit, je nutno minimalizovat ztrátovou funkci $L(\theta)$:

$$L(\theta) = \sum_{i=1}^{n} \left[\hat{y}(x_i, \theta) - y_i\right]^2 = (\mathbb{X} \cdot \theta - y)^{\mathsf{T}} (\mathbb{X} \cdot \theta - y) \tag{1.3}$$

Tato funkce znázorňuje čtverec vzdálenosti pozorovaní y k hledané funkci $\hat{y}(x, \theta)$, jenž chceme mít co nejmenší - proto metoda nejmenších čtverců. Matice $\mathbb{X}$ je tvaru

$$\mathbb{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^p \\ 1 & x_2 & x_2^2 & \ldots & x_2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^p \end{pmatrix} \tag{1.4}$$

Odhadovat parametry θ můžeme pomocí gradientní metody a to způsobem, který je popsán rovnicí (1.2). Najdeme gradient ztrátové funkce

$$\nabla_\theta \, L(\theta) = 2\mathbb{X}^{\mathsf{T}}(\mathbb{X} \cdot \theta - y) \tag{1.5}$$

a postupujeme pomocí (1.2), dokud nezískáme $\hat{\theta} = \arg\min_\theta L(\theta)$.
Metoda nejmenších čtverců má ovšem analytické řešení. Systém rovnic můžeme zapsat formou matice a vektorů

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^p \\ 1 & x_2 & x_2^2 & \ldots & x_2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^p \end{pmatrix} \cdot \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{pmatrix}$$

Pro jednoduchost budeme tímto zápisem rozumět následující rovnici

$$y = \mathbb{X} \cdot \theta + \epsilon \tag{1.6}$$

Naším cílem je získání parametrů θ. Jelikož je $y - \hat{y} = \epsilon$, přepíšeme rovnici pomocí $\hat{y}$, čímž získáme

$$\hat{y} = \mathbb{X} \cdot \theta \tag{1.7}$$

a obě strany rovnice vynásobíme zleva $\mathbb{X}^{\mathsf{T}}$. Tím nám rovnice přejde do tvaru

$$\mathbb{X}^{\mathsf{T}} \cdot y = \mathbb{X}^{\mathsf{T}} \cdot \mathbb{X} \cdot \theta$$

Teď už stačí rovnici zleva vynásobit inverzní maticí $(\mathbb{X}^{\mathsf{T}} \cdot \mathbb{X})^{-1}$. Dostaneme tak konečné řešení

$$\hat{\theta} = (\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathsf{T}} y \tag{1.8}$$

Tento postup zahrnuje i lineární regresi pro hodnotu $p = 1$.

1.2 Úvod do pravděpodobnosti a Bayesovská statistika

1.2.1 Pravděpodobnostní míra

Definice 1.2.1. *(Kolmogorova definice pravděpodobnosti). Mějme množinu Ω vybavenou σ-algebrou $\mathcal{A}$, tedy souborem podmnožin obsahujícím Ω a uzavřeným na doplňky a spočetná sjednocení. Pak libovolnou funkci $P : \mathcal{A} \to \mathbb{R}$, která splňuje :*

1. $(\forall A \in \mathcal{A})(P(A) \geq 0)$.
2. $P(\Omega) = 1$
3. $\forall A_j$ *disjunktní platí* $P(\sum_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} P(A_j)$

Věta 1.2.1. *(Vlastnosti P). Mějme pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$ a něcht $(\forall j \in \mathbb{N})(A_j \in \mathcal{A})$ a $B \in \mathcal{A}$. Pak platí:*

1. $P(\emptyset) = 0$,
2. *Aditivita:* $P(\sum_{j=1}^{n} A_j) = \sum_{j=1}^{n} P(A_j)$,
3. *Monotonie:* $A \subset B \Rightarrow P(A) \leq P(B)$,
4. *Subtraktivita:* $A \subset B \Rightarrow P(B \setminus A) = P(B) - P(A)$,
5. *Omezenost:* $(\forall A \in \mathcal{A})(P(A) \leqslant 1)$,
6. *Komplementarita:* $A \in \mathcal{A} \Rightarrow P(A^C) = 1 - P(A)$

Definice 1.2.2. *(Podmíněná pravděpodobnost). Nechť $A, B \in \mathcal{A}$ a $P(B) > 0$. Pak definujeme podmíněnou pravděpodobnost:*

$$P(A|B) = \frac{P(A, B)}{P(B)} \tag{1.9}$$

Věta 1.2.2. *(Součinové pravidlo). Nechť $A_1, \ldots, A_n \in \mathcal{A}$ a dále nechť také $P(A_1, \ldots, A_n) > 0$. Potom platí:*

$$P(A_1, \ldots, A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_2, A_1) \cdot \ldots \cdot P(A_n|A_1, \ldots, A_{n-1}) \tag{1.10}$$

Věta 1.2.3. *(Bayseova věta). Nechť $A \in \mathcal{A}$ a $P(B) \neq 0$. Potom platí:*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.11}$$

Poznámka. *$P(A)$ nazýváme prior a $P(A|B)$ nazýváme posterior.*

Věta 1.2.4. *(Nezávislost jevů). Nechť $A_j \in \mathcal{A}(\forall j \in \mathbb{N})$. Potom jevy nazveme nezávislé právě tehdy když platí podmínka*

$$P(A_1, \ldots, A_k) = \prod_{i=1}^{k} P(A_i) \tag{1.12}$$

1.2.2 Hustoty pravděpodobnosti

Primárním cílem generativního modelování je hledání distribuce nebo-li hustoty pravděpodobnosti daných dat. Výhodou je, že pro hustotu pravděpodobnosti můžeme využívat stejně pravidlo podmíněnosti (1.9), součinové pravidlo (1.10) a Bayeseovo pravidlo (1.11). Toto se pro nás ukáže jako naprosto klíčové. Budeme uvažovat pouze spojitá rozdělení pravděpodobnosti náhodné veličiny X.

Definice 1.2.3. *(Náhodná veličina) Máme prostor $(\Omega, \mathcal{A})$, potom funkci $X : (\Omega, \mathcal{A}) \to (\mathbb{R}^n, \mathcal{B}_n)$, kde $\mathcal{B}_n$ značí borelovskou σ-algrebru v $\mathbb{R}^n$, nazveme náhodnou veličinou.*

Definice 1.2.4. *(Hustota pravděpodobnosti). Hustotou pravděpodobnosti náhodné veličiny X rozumíme spojitou funkci $p_X(x)$, která splňuje následující dvě podmínky:*

1. $p_X(x) \geq 0$
2. $\int_\Omega p_X(x)\mathrm{d}x = 1$

Náhodná veličina X stejně tak i hustota $p_X(x)$ může být vícerozměrná. Index budeme vynechávat, protože bude jasné, ke které náhodné veličině hustota patří. Podívejme se nyní, jak určit hustotu transformované veličiny.

Věta 1.2.5. *(Transformace náhodné veličiny) Nechť $X \sim p_X(x)$ a nechť $h : \mathbb{R}^n \to \mathbb{R}^n$ je regulární a prosté zobrazení na množině H, takové že $\int_H p_X(x)\mathrm{d}x = 1$. Potom je $Y = h(X)$ náhodná veličina a její hustota $\forall y \in h(H)$ má následující tvar:*

$$p_Y(y) = p_X\left(g^{-1}(y)\right) \mid \det \mathbb{J}_{g^{-1}}(y) \mid \tag{1.13}$$

Nyní ukážeme jak určit ty nejdůležitější charakteristiky náhodné veličiny.

Definice 1.2.5. *(Střední hodnota náhodné veličiny) Má-li náhodná veličina X spojitou hustotu pravděpodobnosti $p(x)$, definujeme její střední (očekávanou) hodnotu $\mathbb{E}[X]$, alternativně značeno $\langle X \rangle$, vztahem*

$$\mathbb{E}[X] = \int_\Omega x p(x)\mathrm{d}x \tag{1.14}$$

Definice 1.2.6. *(Rozptyl náhodné veličiny) Má-li náhodná veličina X spojitou hustotu pravděpodobnosti $p(x)$, definujeme rozptyl (varianci) $\mathbb{D}[X]$, alternativně značeno $var(X)$, vztahem*

$$\mathbb{D}[X] = \int_\Omega x^2 p(x)\mathrm{d}x = \int_\Omega (x - \mathbb{E}[X])^2 \,\mathrm{d}x \tag{1.15}$$

Definice 1.2.7. *(Entropie) Má-li náhodná veličina X spojitou hustotu pravděpodobnosti $p(x)$, definujeme entropii náhodné veličiny $\mathbb{H}[X]$ vztahem*

$$\mathbb{H}[X] = -\int_\Omega p(x) \ln p(x)\mathrm{d}x \tag{1.16}$$

Poznámka. *Kvůli zjednodušení zápisu nebudeme později uvádět integrační množinu, automaticky tak budeme předpokládat,že se integruje přes celý nosič hustoty.*

V dalším textu uvedeme jednotlivá rozdělení a pro přehlednost jejich výše zmíněné charakteristiky, jelikož je této práci budeme využívat.

1.2.2.1 Rovnoměrné rozdělení

Začněme jedním z nejjednodušších rozdělení. Rovnoměrné rozdělení, někdy také uniformní, přiřazuje všem hodnotám stejnou pravděpodobnost. Je definováno na intervalu (a, b) a můžeme ji vyjádřit následujícím způsobem.

$$U(a,b) = \begin{cases} \frac{1}{b-a}, & \text{pro } x \in (a,b) \\ 0, & \text{jinak} \end{cases} \tag{1.17}$$

- $\mathbb{E}[X] = \frac{1}{2}(a+b)$
- $\mathbb{D}[X] = \frac{1}{12}(b-a)^2$
- $\mathbb{H}[X] = \ln(b-a)$

1.2.2.2 Normální rozdělení

Nejdůležitější hustota pravděpodobnosti pro spojité proměnné se nazývá normální nebo také Gaussovo rozdělení. Jeho hustota je definována $\forall x \in \mathbb{R}$ pomocí dvou parametrů $\mu \in \mathbb{R}$ a $\sigma^2 > 0$ jako

$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{1.18}$$

- $\mathbb{E}[X] = \mu$
- $\mathbb{D}[X] = \sigma^2$
- $\mathbb{H}[X] = \frac{1}{2}\ln 2\pi e\sigma^2$

Budeme využívat i d-rozměrnou variantu Gaussova rozdělení, které je definováno vztahem

$$\mathcal{N}(\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \tag{1.19}$$

kde Σ je d×d matice, kterou nazveme kovarianční a $\boldsymbol{\mu}$ je vektor středních hodnot.

- $\mathbb{E}[X] = \boldsymbol{\mu}$
- $\mathbb{D}[X] = \Sigma$
- $\mathbb{H}[X] = \frac{1}{2}\ln\det(2\pi e\Sigma)$

1.2.2.3 Gamma rozdělení

Gamma rozdělení je definováno stejně jako normální rozdělení pomocí dvou parametrů $\alpha > 0$ a $\beta > 0$. Jeho hustota pravděpodobnosti má smysl pro $\forall x > 0$ a můžeme ji najít v několika možných tvarech. My uvedeme tento:

$$\Gamma(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-\beta x\} \tag{1.20}$$

Stejně jako u Gaussova rozdělení uvedeme některé důležité charakteristiky.

- $\mathbb{E}[X] = \frac{\alpha}{\beta}$
- $\mathbb{D}[X] = \frac{\alpha}{\beta^2}$
- $\mathbb{H}[X] = \alpha - \ln\beta + \ln\Gamma(\alpha) + (1-\alpha)\psi(\alpha)$

1.2.2.4 Inverzní gamma rozdělení

Inverzní gamma rozdělení je gamma rozdělení akorát pro převrácenou hodnotu x, je tedy opět popsáno dvěma parametry $\alpha > 0$ a $\beta > 0$ a definováno pro $\forall x > 0$. Jeho hustotu můžeme zapsat následovně:

$$i\Gamma(\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left\{-\frac{\beta}{x}\right\} \tag{1.21}$$

Střední hodnota a rozptyl $\mathrm{i}\Gamma(\alpha,\beta)$ nejsou ale definována pro $\alpha > 0$, platí:

- $\mathbb{E}[X] = \frac{\beta}{\alpha-1}$, pro $\alpha > 1$
- $\mathbb{D}[X] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)^2}$, pro $\alpha > 2$
- $\mathbb{H}[X] = \alpha + \ln\beta + \ln\Gamma(\alpha) - (1+\alpha)\psi(\alpha)$

1.2.3 Bayesovská metoda nejmenších čtverců

Uvažujme standardní problém na nejmenší čtverce (1.7),

$$y = \mathbb{X} \cdot \theta + \epsilon,$$

jednoduchou úpravou dostaneme

$$\epsilon = y - \mathbb{X} \cdot \theta \tag{1.22}$$

Jelikož je pro jednu složku šumu platí $\varepsilon_i \sim \mathcal{N}(0, 1)$ a jsou iid, pro hustotu celého vektoru ϵ platí

$$p(\epsilon) \propto \exp\left\{-\frac{1}{2}\epsilon^\mathsf{T}\epsilon\right\} \tag{1.23}$$

Poznámka. *Normalizační konstantu hustot není nutno neustále psát, proto využíváme znak úměrnosti* $\propto$.

Po transformaci ϵ podle vztahu (1.22) a věty (1.13) dostaneme

$$p(\epsilon) = p(y|\mathbb{X}, \theta) \propto \exp\left\{-\frac{1}{2}(y - \mathbb{X}\theta)^\mathsf{T}(y - \mathbb{X}\theta)\right\} \tag{1.24}$$

Snažíme se získat hustotu $p(\theta|\mathbf{y}, \mathbb{X})$, kterou získáme pomocí Bayesovy věty (1.11).

$$p(\theta|y, \mathbb{X}) = \frac{p(y|\mathbb{X}, \theta)p(\theta|\mathbb{X})}{p(y|\mathbb{X})} \propto p(y|\mathbb{X}, \theta)p(\theta|\mathbb{X}). \tag{1.25}$$

K tomu abychom mohli pokračovat ve výpočtu $p(\theta|y, \mathbb{X})$, potřebujeme určit $p(\theta|\mathbb{X})$. Jelikož je θ nezávislé na $\mathbb{X}$, můžeme psát pouze $p(\theta)$.
Pro hustotu $p(\theta)$ předpokládáme následující vztah:

$$p(\theta) = \mathcal{N}\left(0, \alpha^{-1}\mathbb{I}\right) \propto \exp\left\{-\frac{1}{2}\theta^\mathsf{T}\theta\alpha\right\} \tag{1.26}$$

Nyní můžeme pokračovat dosazením do (1.25)

$$\begin{aligned} p(y|\mathbb{X}, \theta)p(\theta|\mathbb{X}) &\propto \exp\left\{-\frac{1}{2}(y - \mathbb{X}\theta)^\mathsf{T}(y - \mathbb{X}\theta)\right\} \exp\left\{-\frac{1}{2}\theta^\mathsf{T}\theta\alpha\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(y^\mathsf{T}y - \theta^\mathsf{T}\mathbb{X}^\mathsf{T}y - y^\mathsf{T}\mathbb{X}\theta + \theta^\mathsf{T}\mathbb{X}^\mathsf{T}\mathbb{X}\theta + \theta^\mathsf{T}\theta\alpha\right)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[y^\mathsf{T}y - \theta^\mathsf{T}\mathbb{X}^\mathsf{T}y - y^\mathsf{T}\mathbb{X}\theta + \theta^\mathsf{T}\left(\mathbb{X}^\mathsf{T}\mathbb{X} + \alpha\mathbb{I}\right)\theta\right]\right\} \end{aligned} \tag{1.27}$$

Jedná se o součin dvou vícerozměrných gaussovských distribucí, proto můžeme předpokládat tvar řešení pomocí kvadratické formy, která také odpovídá vícerozměrnému Gaussovu rozdělení. Tento tvar navíc obsahuje zbytek z po nejmenších čtvercích, ten ovšem není nutné psát. Platí:

$$p(\theta|y, \mathbb{X}) \exp\left\{-\frac{1}{2}\left(\theta - \hat{\theta}\right)\Sigma^{-1}\left(\theta - \hat{\theta}\right) + z\right\} \propto \exp\left\{-\frac{1}{2}\left(\theta - \hat{\theta}\right)\Sigma^{-1}\left(\theta - \hat{\theta}\right)\right\}$$

a upravujeme dále tak, abychom dokázali určit $\hat{\theta}$ a Σ. Roznásobením dostaneme

$$p(\theta|y, \mathbb{X}) \propto \exp\left\{-\frac{1}{2}\left(\theta^\mathsf{T}\Sigma^{-1}\theta - \hat{\theta}^\mathsf{T}\Sigma\theta - \theta^\mathsf{T}\Sigma^{-1}\hat{\theta} + \hat{\theta}^\mathsf{T}\Sigma^{-1}\hat{\theta}\right)\right\} \tag{1.28}$$

z čehož už při porovnání výrazu $\theta^{\mathsf{T}}\left(\mathbb{X}^{\mathsf{T}}\mathbb{X}+\alpha\mathbb{I}\right)\theta$ v konečném tvaru rovnice (1.27) s výrazem $\theta^{\mathsf{T}}\Sigma^{-1}\theta$ v předchozí rovnici (1.28), plyne předpis pro

$$\Sigma^{-1} = \mathbb{X}^{\mathsf{T}}\mathbb{X} + \alpha\mathbb{I} \tag{1.29}$$

Tento je výsledek je pro nás velmi důležitý a budeme jej i nadále využívat.
Přímo porovnávejme další dva výrazy z těchto rovnic

$$-y^{\mathsf{T}}\mathbb{X}\theta = -\hat{\theta}^{\mathsf{T}}\Sigma^{-1}\theta$$

Nyní z této rovnice jednoduchou úpravou a dosazením za Σ dostaneme další velmi důležitý předpis pro $\hat{\theta}$, a to

$$\hat{\theta} = \Sigma\mathbb{X}^{\mathsf{T}}y = \left(\mathbb{X}^{\mathsf{T}}\mathbb{X}+\alpha\mathbb{I}\right)^{-1}\mathbb{X}^{\mathsf{T}}y \tag{1.30}$$

1.2.4 Divergence

Divergence je funkce $D(.\|.) : S \times S \rightarrow \mathcal{R}$, kde je S je prostor pravděpodobnostních rozdělení a které splňuje následující dvě podmínky:

1. $D(q\|p) \geq 0$
2. $D(q\|p) = 0$ pro $p = q$

Divergence do jisté popisuje vzdálenost nebo rozdíl mezi dvěma distribucemi. Jelikož divergence nemusí splňovat podmínku symetrie a trojúhelníkové nerovnosti, nejedná se tedy o metriku, nýbrž o semimetriku.

f-divergence

Nejdůležitější skupinou divergencí jsou takzvané f-divergence. Jsou definovány pomocí konvexní funkce $f(x)$, kde $x > 0$ a takové že $f(1) = 0$. Jsou tvaru

$$D_f(q\|p) = \int q(x) f\left(\frac{q(x)}{p(x)}\right)\mathrm{d}x \tag{1.31}$$

kde $supp(q)$ značí nosič funkce $q(x)$.

Kullback-Leiblerova divergence

Pro nás bude užitečná tzv. Kullback-Leiblerova divergence, kde za funkci f bereme přirozený logaritmus, značeno log. To je rozhodně konvexní funkce pro kterou platí $\log 1 = 0$. Tvar KL-divergence je následující:

$$D_{KL}(q\|p) = \int q(x)\log\frac{q(x)}{p(x)}\mathrm{d}x \tag{1.32}$$

1.2.5 ELBO

Předpokládejme že máme pozorování x a z jsou skryté (latentní) proměnné. Toto je zcela obecná definice a z může obsahovat i parametry. Posteriorní distribuci latentní proměnné Z můžeme napsat pomocí Bayesova pravidla (1.11), jehož jmenovatel se někdy také nazývá evidence, takto:

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} = \frac{p(x|z)p(z)}{\int p(x,z)\mathrm{d}z} \tag{1.33}$$

Dále zadefinujeme nový objekt

$$\log p(x) = \log \int p(x, z)\mathrm{d}z \tag{1.34}$$

a abychom mohli pokračovat, využijeme pomocnou funkci $q(z|\theta)$

$$\log p(x) = \log \int p(x, z)\mathrm{d}z = \log \int q\left(z|w\right) \frac{p(x, z)}{q\left(z|w\right)}\mathrm{d}z = \log \mathbb{E}_q \left[\frac{p(x, z)}{q\left(z|w\right)}\right] \tag{1.35}$$

Dále využijeme Jensenovu nerovnost, díky které získáme spodní hranici (lower bound), odtud Evidence Lower Bound, čili ELBO.

$$\log \mathbb{E}_q \left[\frac{p(x, z)}{q\left(z|w\right)}\right] \geq \mathbb{E}_q \left[\log \frac{p(x, z)}{q\left(z|w\right)}\right] = \mathcal{L}\left(w\right) \tag{1.36}$$

ELBO můžeme rozepsat pomocí součinového pravidla (1.10), využít vlastností logaritmů a dle definice KL-divergence (1.32), přepsat do tvaru

$$\mathcal{L}\left(w\right) = \mathbb{E}_q \left[\log \frac{p(x, z)}{q\left(z|w\right)}\right] = \mathbb{E}_q \left[\log \frac{p(x|z)p(z)}{q\left(z|w\right)}\right] = \mathbb{E}_q \left[\log p(x|z)\right] - \mathbb{E}_q \left[\log \frac{p(z)}{q\left(z|w\right)}\right] \tag{1.37}$$

$$= \mathbb{E}_q \left[\log p(x|z)\right] - D_{KL}\left(q\left(z|w\right) \| p(z)\right) \tag{1.38}$$

Budeme-li maximalizovat ELBO přes všechny variační parametry w, získáme nejbližší možnou hodnotu k $\log p(x)$. Navíc je maximalizace ELBO ekvivalentní k minimalizaci KL-divergence mezi $q(z|w)$ a $p(z|w)$, jelikož platí

$$\begin{aligned} D_{KL}\left(q\left(z|w\right) \| p(z|x)\right) &= \mathbb{E}_q \left[\log \frac{q(z|w)}{p\left(z|x\right)}\right] \\ &= \mathbb{E}_q \left[\log \frac{q(z|w)p(x)}{p\left(x|z\right)p(z)}\right] \\ &= -\mathbb{E}_q \left[\log p(x|z)\right] + \mathbb{E}_q \left[\log \frac{q(z|w)}{p\left(z|x\right)}\right] + \mathbb{E}_q \left[\log p(x)\right] \\ &= -\mathbb{E}_q \left[\log p(x|z)\right] + D_{KL}\left(q\left(z|w\right) \| p(z)\right) + \log p(x) \end{aligned} \tag{1.39}$$

Z toho jednoduchou úpravou dostaneme konečný vztah

$$D_{KL}\left(q\left(z|w\right) \| p(z|x)\right) = -\mathcal{L}\left(w\right) + \log p(x) \tag{1.40}$$

Příklad

Předvedeme příklad, jak ELBO využít v praxi.
Uvažujme pouze sadu dvou pozorování y_1 a y_2 s normálním rozdělením $\mathcal{N}_{\mathrm{i}}(\theta, 1)$ pro i $\in$1,2. Dále uvažujme jeden parametr $\theta \sim \mathcal{N}(0, \alpha)$ a nechť α má inverzní gamma rozdělení, tedy $\alpha \sim i\Gamma(0, 0)$. Snažíme se získat sdruženou distribuci parametrů θ a α, tedy $p(\theta, \alpha|y_1, y_2)$. Tuto distribuci můžeme přepsat pomocí definice podmíněné pravděpodobnosti (1.9) a řetězového pravidla (1.10) jako

$$p(\theta, \alpha|y_1, y_2) = \frac{p(\theta, \alpha, y_1, y_2)}{p(y_1, y_2)} = \frac{p(y_1|\theta)p(y_2|\theta)p(\theta)p(\alpha)}{p(y_1, y_2)} \tag{1.41}$$

Dosazením předpokladů do čitatele dostaneme:

$$p(y_1|\theta)p(y_2|\theta)p(\theta)p(\alpha) \propto \exp\left\{-\frac{1}{2}(y_1 - \theta)^2\right\} \cdot \exp\left\{-\frac{1}{2}(y_2 - \theta)^2\right\} \cdot \frac{1}{\sqrt{\alpha}} \exp\left\{-\frac{\theta^2}{2\alpha}\right\} \cdot \frac{1}{\alpha} \tag{1.42}$$

Zdánlivě se nám může zdát určení jmenovatele jako jednoduché, protože pravděbodnost $p(y_1, y_2)$ lze získát tzv. marginalizací, nebo-li vyintegrováním přes θ a α.

$$\begin{aligned} p(y_1, y_2) &= \int p(\theta, \alpha, y_1, y_2)\mathrm{d}\theta\mathrm{d}\alpha \\ &= \int p(y_1|\theta)p(y_2|\theta)p(\theta)p(\alpha)\mathrm{d}\theta\mathrm{d}\alpha \\ &= \int \exp\left\{-\frac{1}{2}(y_1-\theta)^2\right\} \cdot \exp\left\{-\frac{1}{2}(y_2-\theta)^2\right\} \cdot \frac{1}{\sqrt{\alpha}} \exp\left\{-\frac{\theta^2}{2\alpha}\right\} \cdot \frac{1}{\alpha}\, \mathrm{d}\theta\mathrm{d}\alpha \end{aligned} \tag{1.43}$$

Po bližším přezkoumání (1.43) zjistíme, že nelze přes α vyintegrovat. Proto použijeme ELBO. Dle definice KL-divergence a za předpokladu $q(\theta, \alpha) = q(\theta)q(\alpha)$ můžeme psát:

$$D_{KL}\left(q\left(\theta, \alpha|\mu, \sigma, \gamma, \delta\right) || p(\theta, \alpha|y_1, y_2)\right) = \int q(\alpha)q(\theta) \ln\left\{\frac{q(\alpha)q(\theta)}{p(\theta, \alpha|y_1, y_2)}\right\} \mathrm{d}\theta\mathrm{d}\alpha = \blacklozenge \tag{1.44}$$

Nezapomínejme, že $q(\theta)$ a $q(\alpha)$ jsou distribuce, pro které zvolíme tvar

$$\begin{aligned} q(\theta) &= \mathcal{N}(\mu, \sigma) \\ q(\alpha) &= i\Gamma(\gamma, \delta) \end{aligned}$$

Dle (1.2.4) navíc víme, že platí $\int q(\alpha)q(\theta)\mathrm{d}\theta\mathrm{d}\alpha = 1$. Výraz budeme rozepisovat pomocí pravidel pro logaritmy a postupně upravovat. Výraz $p(y_1, y_2)$ v integrálu je konstanta, kterou můžeme pro jednoduchost zanedbat. Výsledek budeme na konci maximalizovat a konstanta polohu maxima nemění.

$$\begin{aligned} \blacklozenge &\propto \int q(\alpha)q(\theta) \ln \frac{q(\alpha)q(\theta)}{p(y_1|\theta)p(y_2|\theta)p(\theta)p(\alpha)} \mathrm{d}\theta\mathrm{d}\alpha \\ &= \int q(\theta)q(\alpha)\left(-\ln p(y_1|\theta) - \ln p(y_2|\theta) - \ln p(\theta) - \ln p(\alpha) + \ln q(\theta) + \ln q(\alpha)\right) \mathrm{d}\alpha\mathrm{d}\theta \end{aligned} \tag{1.45}$$

Poslední dva výrazy jsou tzv. entropie pro Gaussovo rozdělení, resp. inverzní gamma rozdělení. Můžeme využít již známých výsledků:

$$\begin{aligned} \int q(\theta) \ln q(\theta)\, \mathrm{d}\theta &\propto -\frac{1}{2} \ln \sigma \\ \int q(\alpha) \ln q(\alpha)\, \mathrm{d}\alpha &= -\gamma - \ln \delta\Gamma(\gamma) + (1+\gamma)\psi(\gamma) \end{aligned}$$

Vypočítejme zbývající výrazy, kde pro jednoduchost budeme pro střední hodnoty využívat značení pomocí špičatých závorek:

$$\begin{aligned} \int q(\theta)q(\alpha) \ln p(y_1|\theta)\, \mathrm{d}\alpha\, \mathrm{d}\theta &= \left\langle -\frac{1}{2}(y_1-\theta)^2 \right\rangle = -\frac{1}{2}\left(y_1^2 - 2y_1\mu + \mu^2 + \sigma\right) \\ \int q(\theta)q(\alpha) \ln p(y_2|\theta)\, \mathrm{d}\alpha\, \mathrm{d}\theta &= \left\langle -\frac{1}{2}(y_2-\theta)^2 \right\rangle = -\frac{1}{2}\left(y_2^2 - 2y_2\mu + \mu^2 + \sigma\right) \\ \int q(\theta)q(\alpha) \ln p(\theta)\, \mathrm{d}\alpha\, \mathrm{d}\theta &= \left\langle -\frac{\theta^2}{2\alpha} - \frac{1}{2}\ln\alpha \right\rangle = -\frac{1}{2}\left(\left(\mu^2 + \sigma\right)\frac{\gamma}{\delta} + \ln\delta - \psi(\gamma)\right) \\ \int q(\theta)q(\alpha) \ln p(\alpha)\, \mathrm{d}\alpha\, \mathrm{d}\theta &= \langle -\ln\alpha \rangle = \psi(\gamma) - \ln\delta \end{aligned}$$

Nyní máme všechny výrazy pro výpočet distribuce $q(\theta, \alpha)$ numericky a to pomocí minimalizace KL-divergence přes parametry $\mu, \sigma, \gamma, \delta$.

1.3 Teorie grafů

Poslední teoretickou kapitolou bude vsuvka do toerie grafů. Tato kapitola poslouží k výhodnému popsání složitějších datových struktur.

Definice 1.3.1. *(Graf) Grafem G se rozumí dvojice (V, H), kde V je množina vrcholů grafu G, H je množina hran tohoto grafu a tyto množiny jsou vzájemně disjunktní.*

Definice 1.3.2. *(Cesta v grafu) Cestou v grafu rozumíme posloupnost vrcholů a hran $(v_0, h_1, v_1, \ldots, h_t, v_t)$, kde vrcholy $v_0, \ldots, v_t$ jsou navzájem různé vrcholy grafu G a pro každé $i = 1, 2, \ldots, t$ je $e_i = \{v_{i-1}, v_i\} \in H$*

Definice 1.3.3. *(Souvislost grafu) Řekneme že graf G je souvislý, jestliže pro každé dva vrcholy v_0 a v_1 existuje v G cesta z v_0 do v_1.*

Definice 1.3.4. *(Cyklus v grafu) Cyklem v grafu G rozumíme posloupnost vrcholů a hran $(v_0, h_1, v_1, \ldots, h_t, v_t = v_0)$, kde vrcholy $v_0, \ldots, v_{t-1}$ jsou navzájem různé vrcholy grafu Ga pro každé $i = 1, 2, \ldots, t$ je $e_i = \{v_{i-1}, v_i\} \in H$*

Definice 1.3.5. *(Strom) Strom je souvislý graf neobsahující cyklus.*

Primárním cílem je výhodně zadefinovat stromovou strukturu. Jak na tyto definice napasovat generativní model, se pokusíme osvětlit až ve třetí kapitole, toto je pouze nezbytný aparát. Nejprve je nutno generativní model zadefinovat.

Kapitola 2

Generativní modely

Ve strojovém učení se setkáváme s dvěma hlavními typy modelů a to jsou generativní modely a diskriminativní modely. Jak už napovídá název této práce, budeme se zde zabývat výhradně generativními modely.

Definice 2.0.1. *(Generativní model) Mějme nějakou množinu datových záznamů x, představující nezávislé proměnné a nějakou množinu y, jakožto závislé proměnné. Generativní model je potom takový model, který se učí sdruženou distribuci* $p(x, y)$.

Příklad

Jeden ze způsobů jak odhadnout distribuci $p(y, x)$ je využití součinového pravidla (1.10), pomocí kterého získáme

$$p(y, x) = p(y|x)p(x) \tag{2.1}$$

Problém je tedy převeden na hledání distribucí $p(y|x)$ a $p(x)$. Pro ilustraci uvažujme následující množinu datových záznamů.

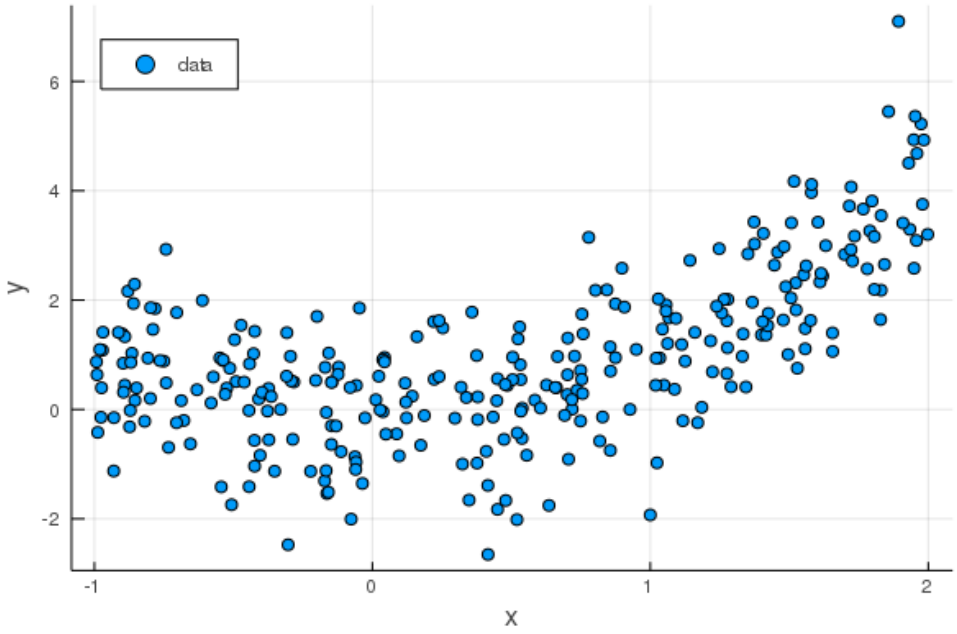
Určit distribuci $p(x)$ není nic těžkého, jelikož jsou tato data na ose x rozděleny rovnoměrně a to přesněji na intervalu $(-1, 2)$. To můžeme určit například z histogramu x-ových souřadnic jednotlivých bodů.

$$p(x) = U(-1, 2) \tag{2.2}$$

Nyní přejdeme k hledání distribuce $p\,(y|x)$. Tu můžeme určit pomocí metody nejmenších čtverců (1.8), protože víme že pro takovou distribuci platí

$$p(y|x) = \mathcal{N}\left(\mathbb{X}\theta, \sigma^2 I\right) \tag{2.3}$$

kde σ^2 je rozptyl jedné složky šumu ε_i a matice $\mathbb{X}$ je ve tvaru definovaném pomocí (1.7). Nyní máme obě složky k určení $p(y, x)$, zobrazme proto tzv. contour plot, abychom si udělali představu, jak tato distribuce vypadá.



2.1 Variační autoencoder

Cílem je najít hustotu $p(x)$ vzorků $\{x_i\}_{i=1}^N$, jehož empirická hustota se dá zapsat pomocí Diracovy delta funkce

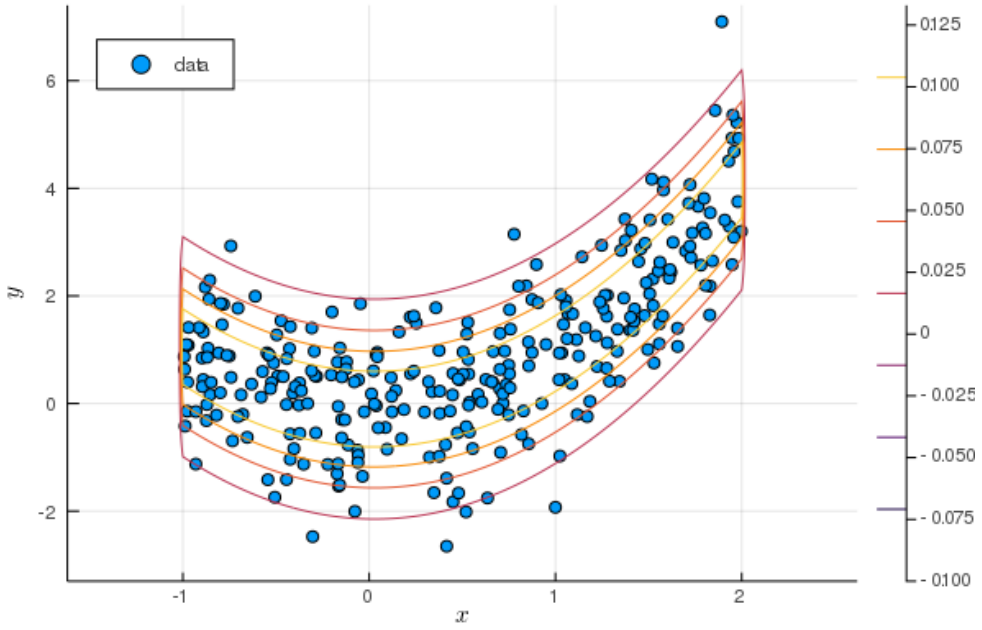
$$p_{\text{emp}}(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i) \tag{2.4}$$

Předpokládáme následující vztahy $x = f_\theta(z) + \epsilon$, $\epsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$ a vzájemnou nezávislost x_i. Z toho můžeme určit distribuce:

$$\begin{aligned} p(x|z) &= \mathcal{N}\left(f_\theta(z), \sigma^2 I\right), \\ p(z) &= \mathcal{N}(0, I), \end{aligned} \tag{2.5}$$

a proto má smysl využít následující formu apromaximace

$$p(x) = \int p(x|z)p(z)\mathrm{d}z \tag{2.6}$$



2.1.1 Naivní přístup

K nalezení $p(x)$ je třeba najít parametry θ transformace $f_\theta(z)$, proto zkusme využít KL-divergence a hledat tak θ minimalizací $D_{KL}\left(p_{\text{emp}}(x)\|p_\theta(x)\right)$

$$\begin{aligned}\hat{\theta} &= \arg\min \sum_{i=1}^{n} \log p\left(x_i\right) \\ &= \arg\min \sum_{i=1}^{n} \log \int \mathcal{N}\left(f_\theta(z), \sigma\right) \mathcal{N}\left(0, 1\right) \mathrm{d}z \\ &= \arg\min \sum_{i=1}^{N} \log \sum_{j=1}^{N} \exp\left\{-\frac{1}{2\sigma^2}\left(x - f_\theta(z)\right)^2\right\}\end{aligned} \tag{2.7}$$

Integrace přes z je nahrazena vzorkováním. Tento postup ovšem nemusí konvergovat ke správným výsledkům.

2.1.2 Variační Bayseova metoda

Lepší metodou se ukazuje vzorkovat z podmíněné distribuce $q(z|x)$ a využít ELBO:

$$\begin{aligned}D_{KL}\left(q\left(z|x\right)\|p(z|x)\right) &= \mathbb{E}_q\left[\log q(z|x) - \log p\left(z|x\right)\right] \\ &= \mathbb{E}_q\left[\log q(z|x) - \log p(x|z) - \log p(z) + \log p(x)\right]\end{aligned} \tag{2.8}$$

Tuto rovnici můžeme přepsat pomocí KL-divergence

$$\log p(x) - D_{KL}\left(q\left(z|x\right)\|p(z|x)\right) = \mathbb{E}_q\left[\log p(x|z)\right] - D_{KL}\left(q\left(z|x\right)\|p(z)\right) \tag{2.9}$$

kde pravá strana této rovnice je lower bound $\log p(x)$. Jestliže vybereme parametrickou formu distribuce

$$q\left(z|x\right) = \mathcal{N}\left(\mu_\phi(x), \sigma^2_\phi(x)\right) \tag{2.10}$$

můžeme parametry θ a ϕ minimalizovat zároveň a to následovně:

$$\begin{aligned}\hat{\theta}, \hat{\phi} &= \arg\min \sum_{i=1}^{n} \log p\left(x_i\right) \\ &= \arg\min \left\{\mathbb{E}_q\left[\log p(x|z)\right] - D_{KL}\left(q\left(z|x\right) \| p(z)\right)\right\}\end{aligned} \tag{2.11}$$

V metodě variačního autoencoderu jsou důležité následující dvě věci. První je trik v reparametrizaci

$$z = \mu(x) + \sigma^2(x) \odot \epsilon \tag{2.12}$$

a druhý je analytické řešení KL-diveregence dvou gaussovských distribucí

$$\begin{aligned}D_{KL}\left(q\left(z|x\right) \| p(z)\right) &= \frac{1}{2}\left[\mathrm{tr}(\sigma^2(x)) - \mu^{\mathsf{T}}(x)\mu(x) - k - \log\det\left(\sigma^2(x)\right)\right] \\ &= \frac{1}{2}\left[\sum_{l=1}^{k}(\sigma^2(x)) - \mu^{\mathsf{T}}(x)\mu(x) - k - \sum_{l=1}^{k}\log\sigma^2(x)\right]\end{aligned} \tag{2.13}$$

Kdybychom totiž nevybrali aproximační distribuci gaussovskou, nemohli bychom tímto způsobem $\hat{\theta}, \hat{\phi}$ určit. Toto řešení KL-divergence vede na konečný tvar

$$\hat{\theta}, \hat{\phi} = \arg\min\left[\sum_{i=1}^{n}\sum_{j=1}^{p}\left[x_i - f_\theta\left(\mu(x_i) + \sigma^2(x_i)\cdot\epsilon_{i,j}\right)\right] - \frac{1}{2}\left[\sum_{l=1}^{k}(\sigma^2_\phi(x_i)) - \mu^{\mathsf{T}}_\phi(x_i)\mu_\phi(x_i) - k - \sum_{l=1}^{k}\log\sigma^2_\phi(x_i)\right]\right] \tag{2.14}$$

Příklad

Kapitola 3

Stromové struktury

Stromovou strukturou dat rozumíme množinu datových záznamů popsaných pomocí množiny vrcholů a hran. Vrcholy dané stromové struktury představují jednotlivé body x a y. V podstatě si to můžeme představit opravdu jako strom - má jeden kořen, v první úrovni se dělí na k_1 větví, každá další větev se v druhé úrovni dělí na k_{2_i} a tak dále. My se v této práci budeme zabývat pouze kořenem a první úrovní větví. Ilustrujeme to na jednoduchém příkladu.

Příklad

Předpokládejme takový model který ke každému y_i, $i \in \hat{n}$ přiřazuje určitý počet $x_j^{(k)}$, přičemž celkový počet x_j je $j \in \hat{m}$, $m > n$. Tento počet může být generován například Poissonovým rozdělením s danou střední hodnotou λ, navíc celkový součet těchto počtů musí být m. Jednotlivá y_i potom můžou být určena nějakou kvadratickou závislostí na průměrech k nim přiřazených x_j^k. Všechna x_j jsou přitom generována pomocí uniformního rozdělení. Pokusme se nalézt sdruženou distribuci $p(y, \bar{x})$, kde $\bar{x}_j$ značí výběrový průměr x_j . Použijeme opět součinové pravidlo

$$p(y, \bar{x}_j) = p(y|\bar{x}_j)p(\bar{x}_j) \tag{3.1}$$

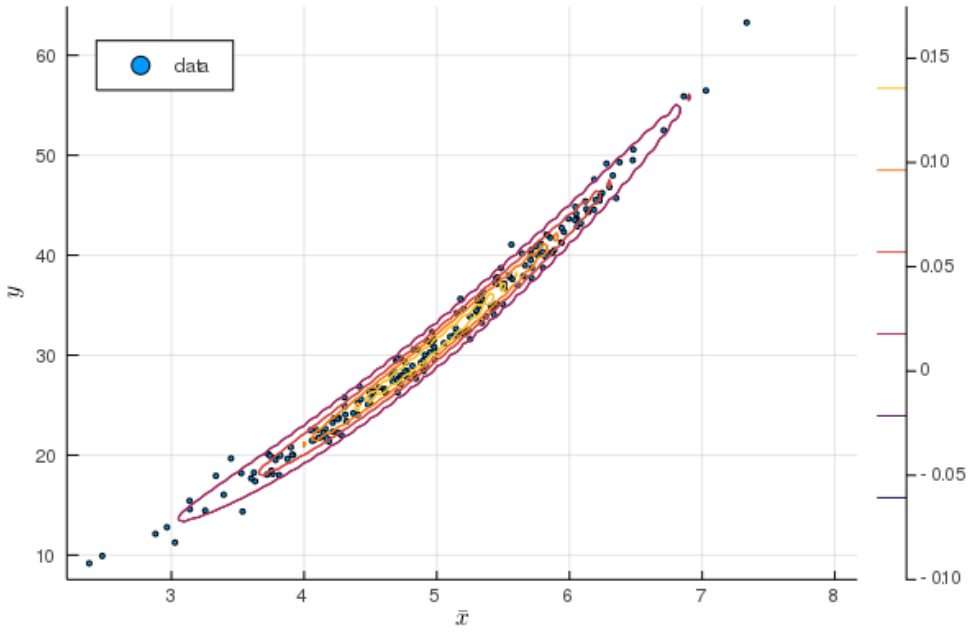
a pokusme se nalézt tyto dvě distribuce. Tento postup je nám známý už z kapitoly (2). Pro určení podmíněné distribuce $p\left(y|\bar{x}_j\right)$ použijeme opět metodu nejmenších čtverců a dostaneme

$$p\left(y|\bar{x}\right) = \mathcal{N}\left(\mathbb{X}\theta, \sigma^2 I\right) \tag{3.2}$$

Ovšem zde jsou prvky matice $\mathbb{X}$ jednotlivé průměry $\bar{x}_i$. Určit distribuci $p(\bar{x}_j)$ lze určit obdobně pomocí histogramu. Navíc víme-li, že se jedná o výběrové průměry, je z centrální limitní věty jasné, že se bude jednat o normální rozdělení.

$$p(\bar{x}_j) = \mathcal{N}\left(\frac{1}{j}\sum_{j=1}^{m}\bar{x}_j, std(\bar{x}_j)\right) \tag{3.3}$$

Pro vizualizaci sdružené distribuce $p(y, \bar{x}_j)$ využijeme opět contour plot.


data
60
50
40
30
20
10
3
4
5
6
7
8
y
$\bar{x}$
0.15
0.10
0.05
0
- 0.05
- 0.10

Závěr

Text závěru....

Literatura

[1] S. Allen, J. W. Cahn: *A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening*. Acta Metall., 27:1084-1095, 1979.

[2] G. Ballabio et al.: *High Performance Systems User Guide*. High Performance Systems Department, CINECA, Bologna, 2005. `www.cineca.it`

[3] J. Becker, T. Preusser, M. Rumpf: *PDE methods in flow simulation post processing*. Computing and Visualization in Science, 3(3):159-167, 2000.