



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 srovnajte 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

## Generativní modely dat popsaných stromovou strukturou

*Obor:* Matematické inženýrství

*Druh práce:* Bakalářská práce

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

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

## Generative models of tree structured data

[illegible]

**Key words:** keywords in alphabetical order separated by commas

# Obsah

<b>Úvod</b>	<b>7</b>
0.1 Příklad . . . . .	7
<b>1 Teorie pravděpodobnosti</b>	<b>8</b>
1.1 Definice pravděpodobnosti . . . . .	8
1.2 Hustoty pravděpodobnosti . . . . .	9
1.2.1 Normální rozdělení . . . . .	9
1.2.2 Gamma rozdělení . . . . .	10
1.2.3 Inverzní gamma rozdělení . . . . .	10
<b>2 Optimalizace</b>	<b>11</b>
2.1 Metoda nejmenších čtverců . . . . .	11
2.2 Bayesovská lineární regrese . . . . .	11
2.3 Gradient Descent . . . . .	13
2.3.1 ADAM . . . . .	13
2.4 Divergence . . . . .	13
2.4.1 f-divergence . . . . .	13
<b>Závěr</b>	<b>17</b>

# Úvod

## 0.1 Příklad

Ze začátku uvažujme jednoduchý příklad, který naznačí následující problematiku. Předpokládejme že máme trénovací množinu obsahující  $n$  pozorování  $x$ , nebo-li  $\mathbf{X} = (x_1, \dots, x_n)$ . Dále máme ke každému  $x$  právě jedno pozorování  $t$ , psáno  $\mathbf{t} = (t_1, \dots, t_n)$ . Celé to můžeme zapsat jako  $(x_1, \dots, x_n) \mapsto (t_1, \dots, t_n)$ . Naším cílem je využít tuto trénovací množinu k predikci hodnot  $\hat{t}$  a tedy k určení nové hodnoty  $\hat{x}$ , jakožto výstupní proměnné. Pozorované hodnoty  $(t_1, \dots, t_n)$  jsou ale zatíženy nepřesnostmi a přestože závislost  $(t_1, \dots, t_n)$  může být na  $(x_1, \dots, x_n)$  kvadratická, nemusí se podařit najít kvadratickou funkci tak, aby procházela všemi body. Naším cílem je tedy nafitovat data pomocí polynomické funkce řádu  $n$  ve tvaru

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \dots + w_n x^n = \sum_{i=0}^n w_i x^i \quad (1)$$

Tato funkce je lineární v neznámých parametrech  $\mathbf{w}$ . Takové modely nazýváme lineární a jejich vlastnosti budeme nadále využívat.

Abychom našli ten nejlepší možný fit, je nutno pomocí derivace minimalizovat tzv. chybovou funkci  $E(\mathbf{w})$ , která je tvaru

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N [y(x_n, \mathbf{w}) - t_n]^2 \quad (2)$$

Pomocí minimalizace získáme parametry  $\mathbf{w}$  a jsme tudíž schopni sestavit předpis polynomu, který nejlépe daná data proloží. Je zde několik problémů, např.: jaký řád polynomu zvolit, více popsáno v ...

Tento jednoduchý příklad lze modifikovat mnoha způsoby, které budeme postupně rozebírat. Nejprve budeme ale potřebovat základy z teorie pravděpodobnosti.

# Kapitola 1

## Teorie pravděpodobnosti

### 1.1 Definice pravděpodobnosti

**Definice 1.1.1.** (Kolmogorova definice pravděpodobnosti). Mějme množinu  $\Omega$  vybavenou  $\sigma$ -algebrou  $\mathcal{A}$ , tedy souborem podmnožin obsahujícím  $\Omega$  a uzavřeným na doplňky a spočetná sjednocení. Pak libovolnou funkci  $P : \mathcal{A} \rightarrow \mathbb{R}$ , 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 1.1.1.** (Vlastnosti  $P$ ). Mějme pravděpodobnostní prostor  $(\Omega, \mathcal{A}, P)$  a necht  $(\forall j \in \mathbb{N})(A_j \in \mathcal{A})$  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:  $(\forall A \in \mathcal{A})(P(A) \leq 1)$ ,
6. Komplementarita:  $A \in \mathcal{A} \Rightarrow P(A^C) = 1 - P(A)$

**Definice 1.1.2.** (Podmíněná pravděpodobnost). Necht'  $A, B \in \mathcal{A}$  a  $P(B) > 0$ . Pak definujeme podmíněnou pravděpodobnost:

$$P(A|B) = \frac{P(A, B)}{P(B)} \quad (1.1)$$

**Věta 1.1.2.** (Součinové pravidlo). Necht'  $A_1, \dots, A_n \in \mathcal{A}$  a dále necht' také  $P(A_1, \dots, A_n) > 0$ . Potom platí:

$$P(A_1, \dots, A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_2, A_1) \cdot \dots \cdot P(A_n|A_1, \dots, A_{n-1}) \quad (1.2)$$

**Věta 1.1.3.** (Bayseova věta). Necht'  $A \in \mathcal{A}$  a  $P(B) \neq 0$ . Potom platí:

$$P(A, B) = \frac{P(B, A)P(A)}{P(B)} \quad (1.3)$$



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

**Věta 1.1.4.** (Nezávislost jevů). Necht'  $A_j \in \mathcal{A} (\forall j \in \mathbb{N})$ . Potom jevy nazveme nezávislé právě tehdy když platí podmínka

$$P(A_1, \dots, A_k) = \prod_{i=1}^k P(A_i) \quad (1.4)$$

## 1.2 Hustoty pravděpodobnosti

**Poznámka.** Budeme uvažovat pouze spojitá rozdělení pravděpodobnosti.

**Definice 1.2.1.** (Hustota pravděpodobnosti). Hustotou pravděpodobnosti rozumíme spojitou funkci  $f(x)$ , která splňuje následující dvě podmínky:

1.  $f(x) \geq 0$
2.  $\int_{-\infty}^{\infty} f(x) = 1$

**Poznámka.** Hustotu pravděpodobnosti lze definovat také pro vícerozměrné funkce  $f(\mathbf{x})$ . Podmínky, které musí vícerozměrná hustota splňovat jsou analogické:

1.  $f(\mathbf{x}) \geq 0$
2.  $\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(\mathbf{x}) = 1$

### 1.2.1 Normální rozdělení

Nejdůležitější hustota pravděpodobnosti pro spojitě 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

$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \quad (1.5)$$

Uved'me zde dvě důležité charakteristiky Gaussova rozdělení a to jsou střední hodnota (někdy také očekávaná hodnota)  $\mathbb{E}(X)$  alternativně značeno  $\langle X \rangle$  rozptyl (případně variance)  $\mathbb{D}(X)$  alternativně značeno  $\text{var}(X)$ .

- $\mathbb{E}(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu$
- $\mathbb{D}(X) = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sigma^2$

Můžeme definovat také d-rozměrnou hustotu a to vztahem

$$\mathcal{N}(\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right\} \quad (1.6)$$

kde  $\Sigma$  je d×d kovariační matice a  $\mu$  je vektor středních hodnot.

### 1.2.2 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:

$$\Gamma(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \exp\{-\beta x\} x^{\alpha-1} \quad (1.7)$$

Stejně jako u Gaussova rozdělení uvedeme některé důležité charakteristiky.

- $\mathbb{E}(X) = \int_0^\infty x \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta x} x^{\alpha-1} dx = \frac{\alpha}{\beta}$
- $\mathbb{D}(X) = \int_0^\infty x^2 \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta x} x^{\alpha-1} dx = \frac{\alpha}{\beta^2}$

### 1.2.3 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:

$$i\Gamma(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \exp\left\{-\frac{\beta}{x}\right\} x^{-\alpha-1} \quad (1.8)$$

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

- $\mathbb{E}(X) = \int_0^\infty x \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\frac{\beta}{x}} x^{-\alpha-1} dx = \frac{\beta}{\alpha-1}$ , pro  $\alpha > 1$
- $\mathbb{D}(X) = \int_0^\infty x^2 \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\frac{\beta}{x}} x^{-\alpha-1} dx = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)^2}$ , pro  $\alpha > 2$

# Kapitola 2

## Optimalizace

### 2.1 Metoda nejmenších čtverců

Mějme soubor bodů o  $n$  prvcích, tedy  $(x_i, y_i) \forall i \in \hat{n}$ . Chceme najít polynom předem daného stupně tak, aby co nejlépe prokládal dané body. Jinými slovy se pokusíme najít koeficienty  $\theta_n$  daného polynomu. Mějme tedy sadu rovnic, kterou již zapíšeme ve formě matic následujícím způsobem:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^k \\ 1 & x_2 & x_2^2 & \dots & x_2^k \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^k \end{pmatrix} \cdot \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{pmatrix} \quad (2.1)$$

Pro jednoduchost budeme tímto maticovým zápisem rozumět následující rovnici

$$\mathbf{y} = \mathbb{X} \cdot \Theta \quad (2.2)$$

Naší cílem je získání parametrů  $\Theta$ , proto obě strany rovnice vynásobíme zleva  $\mathbb{X}^T$ . Tím nám rovnice přejde do tvaru

$$\mathbb{X}^T \cdot \mathbf{y} = \mathbb{X}^T \cdot \mathbb{X} \cdot \Theta \quad (2.3)$$

Ted' už stačí rovnici zleva vynásobit inverzní maticí  $(\mathbb{X}^T \cdot \mathbb{X})^{-1}$ . Dostaneme tak konečné řešení

$$\Theta = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{y} \quad (2.4)$$

Vidíme že pokud máme zadán soubor bodů  $(x_i, y_i) \forall i \in \hat{n}$ , není problém kýžené parametry získat.

### 2.2 Bayesovská lineární regrese

Uvažujme standardní problém na lineární regresi, avšak více specifikujme chyby  $\varepsilon$ , kterými je zatížen každý bod  $y_i$  pro  $\forall i \in \hat{n}$ .

$$\mathbf{y} = \mathbb{X} \cdot \Theta + \epsilon \quad (2.5)$$

kde  $\varepsilon_i \sim \mathcal{N}(0, 1)$  a pro jeho pravděpodobnost tudíž platí  $P(\varepsilon_i) \propto \exp(-\frac{1}{2}\varepsilon_i^2)$ . Z rovnice (2.5) jednoduchou úpravou dostaneme

$$\epsilon = \mathbf{y} - \mathbb{X} \cdot \Theta \quad (2.6)$$

Pokusme se tuto rovnost přepsat pomocí pravděpodobností. Využijeme vícerozměrné Gaussovo rozdělení (1.6).

**Poznámka.** Zanedbáváme normalizační konstantu, proto využíváme znak  $\propto$ .

$$P(\epsilon) \propto P(\mathbf{y}|\mathbb{X}, \Theta) \propto \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbb{X}\Theta)^T(\mathbf{y} - \mathbb{X}\Theta)\right) \quad (2.7)$$

Snažíme se získat pravděpodobnost  $P(\Theta|\mathbf{y}, \mathbb{X})$ , kterou získáme pomocí Bayesovy věty (1.3).

$$P(\Theta|\mathbf{y}, \mathbb{X}) = \frac{P(\mathbf{y}|\mathbb{X}, \Theta)P(\Theta|\mathbb{X})}{P(\mathbf{y}|\mathbb{X})} \propto P(\mathbf{y}|\mathbb{X}, \Theta)P(\Theta|\mathbb{X}). \quad (2.8)$$

K tomu abychom mohli pokračovat ve výpočtu  $P(\Theta|\mathbf{y}, \mathbb{X})$ , potřebujeme určit  $P(\Theta|\mathbb{X})$ . Jelikož je  $\Theta$  nezávislé na  $\mathbb{X}$ , můžeme psát pouze  $P(\Theta)$ .

Pro pravděpodobnost  $P(\Theta)$  předpokládáme následující vztah:

$$P(\Theta) = \mathcal{N}(0, \alpha^{-1}\mathbb{I}) \propto \exp\left(-\frac{1}{2}\Theta^T\Theta\alpha\right) \quad (2.9)$$

Nyní můžeme pokračovat dosazením do (2.8):

$$\begin{aligned} P(\mathbf{y}|\mathbb{X}, \Theta)P(\Theta|\mathbb{X}) &\propto \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbb{X}\Theta)^T(\mathbf{y} - \mathbb{X}\Theta)\right) \exp\left(-\frac{1}{2}\Theta^T\Theta\alpha\right) \\ &\propto \exp\left(-\frac{1}{2}(\mathbf{y}^T\mathbb{Y} - \Theta^T\mathbb{X}^T\mathbf{y} - \mathbb{Y}^T\mathbb{X}\Theta + \Theta^T\mathbb{X}^T\mathbb{X}\Theta + \Theta^T\Theta\alpha)\right) \\ &\propto \exp\left(-\frac{1}{2}[\mathbf{y}^T\mathbf{y} - \Theta^T\mathbb{X}^T\mathbf{y} - \mathbf{y}^T\mathbb{X}\Theta + \Theta^T(\mathbb{X}^T\mathbb{X} + \alpha\mathbb{I})\Theta]\right) \end{aligned} \quad (2.10)$$

Pro dokončení je důležitý předpoklad tvaru řešení a to:

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

který dále pomocí prvního tvaru upravíme tak, abychom dokázali určit  $\hat{\Theta}$ ,  $\Sigma$  a  $z$ . Roznásobením dostaneme

$$\exp\left(-\frac{1}{2}[\Theta^T\Sigma^{-1}\Theta - \hat{\Theta}^T\Sigma\Theta - \Theta^T\Sigma^{-1}\hat{\Theta} + \hat{\Theta}^T\Sigma^{-1}\hat{\Theta}] + z\right)$$

z čehož už okamžitě plyne předpis pro

$$\Sigma^{-1} = \mathbb{X}^T\mathbb{X} + \alpha\mathbb{I} \quad (2.11)$$

Tento je výsledek je pro nás velmi důležitý a budeme jej i nadále využívat.

Přímo porovnáním také můžeme vidět, že

$$-\mathbf{y}\mathbb{X}\Theta = -\hat{\Theta}^T\Sigma^{-1}\Theta$$

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

$$\hat{\Theta} = \Sigma\mathbb{X}^T\mathbf{y} = (\mathbb{X}^T\mathbb{X} + \alpha\mathbb{I})^{-1}\mathbb{X}^T\mathbf{y} \quad (2.12)$$

Pro  $z$  nám zbývá

$$z = \mathbf{y}^T\mathbf{y} - \hat{\Theta}^T\Sigma^{-1}\hat{\Theta}.$$

Používáme ale znak úměrnosti a  $\exp(z)$  je pouze konstanta, můžeme ji tedy vynechat a dostaneme

$$P(\Theta|\mathbf{y}, \mathbb{X}) \propto \exp\left(-\frac{1}{2}(\Theta - \hat{\Theta})\Sigma^{-1}(\Theta - \hat{\Theta})\right)$$

## 2.3 Gradient Descent

Jedná se iterativní optimalizační metodu pomocí které hledáme minimum dané funkce. My se snažíme zminimalizovat funkci  $L = (\mathbb{Y} - \mathbb{X}\Theta)^T(\mathbb{Y} - \mathbb{X}\Theta)$  (loss function), neboli funkci (2), pokud nebudeme provádět maticový zápis. Minimalizujeme  $L$ , tedy derivujeme dle vektoru  $\Theta$  a dostaneme

$$\nabla_{\Theta} L = 2\mathbb{X}^T(\mathbb{X}\Theta - \mathbb{Y}). \quad (2.13)$$

Použijeme bod  $\mathbf{a}$  funkce  $L(\Theta)$  jako vchozí bod, ze kterého se pohybujeme ve směru záporného gradientu s krokem  $\gamma \in \mathbb{R}_+$ , který můžeme s každou iterací měnit. Toto provádíme, dokud nejsme v minimu funkce. Tento postup můžeme zapsat jako

$$a_{n+1} = a_n - \gamma \nabla_{\Theta} L(a_n) \quad (2.14)$$

### 2.3.1 ADAM

Předchozí metoda není tak rychlá, jak bychom pro výpočet minima funkce potřebovali. Používáme proto iterační gradientní metodu ADAM, která navíc používá druhý moment gradientu, popřípadě můžeme ladit i záporné koeficienty.

## 2.4 Divergence

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

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

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

### 2.4.1 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

$$D_f(P||q) = \int_{\text{supp}(q)} P(x) f\left(\frac{q(x)}{P(x)}\right) dx \quad (2.15)$$

kde  $\text{supp}(q)$  značí nosič funkce  $q(x)$ .

#### 2.4.1.1 Kullback-Leiblerova divergence

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

$$D_{KL}(p||q) = \int_{\text{supp}(q)} p(x) \ln\left(\frac{q(x)}{p(x)}\right) dx \quad (2.16)$$

Předvedeme příklad, kde je Kullback-Leiblerova divergence velmi užitečná. Pro začátek uvažujme pouze sadu dvou souřadnic  $y_1$  a  $y_2$  s normálním rozdělením  $\mathcal{N}_i(\theta, 1)$  pro  $i \in 1, 2$ . Dále uvažujme jeden parametr  $\theta \sim \mathcal{N}(0, \alpha)$  a necht'  $\alpha$  má inverzní gamma rozdělení, tedy  $\alpha \sim i\Gamma(0, 0)$ . Snažíme se získat pravděpodobnosti parametrů  $\theta$  a  $\alpha$ , tedy  $P(\theta, \alpha|y_1, y_2)$ . Tuto pravděpodobnost můžeme přepsat pomocí definice podmíněné pravděpodobnosti a řetězového pravidla jako

$$P(\theta, \alpha|y_1, y_2) = \frac{P(\theta, \alpha, y_1, y_2)}{P(y_1, y_2)} = \frac{P(y_1|\theta)P(y_2|\theta)P(\theta)P(\alpha)}{P(y_1, y_2)} \quad (2.17)$$

Dosažením předpokladů do čitatele dostaneme:

$$P(y_1|\theta)P(y_2|\theta)P(\theta)P(\alpha) \propto \exp\left\{-\frac{1}{2}(y_1 - \theta)^2\right\} \cdot \exp\left\{-\frac{1}{2}(y_2 - \theta)^2\right\} \cdot \frac{1}{\sqrt{\alpha}} \exp\left\{-\frac{\theta^2}{2\alpha}\right\} \cdot \frac{1}{\alpha} d\theta d\alpha \quad (2.18)$$

Zdánlivě se nám může zdát určení jmenovatele jako jednoduché. Standardním způsobem bychom pravděpodobnost  $P(y_1, y_2)$  získali tzv. marginalizací, nebo-li vyintegrováním přes  $\theta$  a  $\alpha$ .

$$\begin{aligned} P(y_1, y_2) &= \int P(\theta, \alpha, y_1, y_2) d\theta d\alpha \\ &= \int P(y_1|\theta)P(y_2|\theta)P(\theta)P(\alpha) d\theta d\alpha \\ &= \int \exp\left\{-\frac{1}{2}(y_1 - \theta)^2\right\} \cdot \exp\left\{-\frac{1}{2}(y_2 - \theta)^2\right\} \cdot \frac{1}{\sqrt{\alpha}} \exp\left\{-\frac{\theta^2}{2\alpha}\right\} \cdot \frac{1}{\alpha} d\theta d\alpha \end{aligned} \quad (2.19)$$

Po bližším přezkoumání (2.17) zjistíme, že nelze přes  $\alpha$  vyintegrovat. Proto použijeme KL-divergenci. Dle definice KL-divergence můžeme psát:

$$P(y_1, y_2) = \int P(\theta, \alpha, y_1, y_2) d\theta d\alpha \stackrel{\text{KLD}}{\approx} \int_G q(\alpha)q(\theta) \ln \frac{P(\theta, \alpha, y_1, y_2)}{q(\alpha)q(\theta)} d\theta d\alpha = \diamond \quad (2.20)$$

kde  $G = \text{supp}(q(\theta)) \times \text{supp}(q(\alpha))$ . Nezapomínejme, že  $q(\theta)$  a  $q(\alpha)$  jsou distribuce, pro které si apriori zvolíme

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

Dle (1.2.1) navíc víme, že platí  $\int_G q(\alpha)q(\theta) d\theta d\alpha = 1$ . Výraz budeme rozepisovat pomocí pravidel pro logaritmy a postupně upravovat.

$$\begin{aligned} \diamond &\stackrel{2.15.}{=} \int_G q(\alpha)q(\theta) \ln \frac{P(y_1|\theta)P(y_2|\theta)P(\theta)P(\alpha)}{q(\alpha)q(\theta)} d\theta d\alpha \\ &= \int_G q(\theta)q(\alpha) (\ln P(y_1|\theta) + \ln P(y_2|\theta) + \ln P(\theta) + \ln P(\alpha) - \ln q(\theta) - \ln q(\alpha)) d\alpha d\theta \end{aligned} \quad (2.21)$$

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

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

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

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

Nyní máme všechny výrazy pro výpočet  $P(y_1, y_2)$  numericky.

Pokusme se nyní využít KL - divergenci v poněkud složitějším případě. Nalezneme koeficienty polynomu pátého stupně, který nejlépe proloží dva body.

Uvažujme následující pravděpodobnostní model

$$P(\mathbf{y}, \theta | X, \alpha) = P(\mathbf{y} | \theta, X) P(\theta | \alpha) P(\alpha) = \mathcal{N}(X\theta, I) \mathcal{N}(0, \alpha^{-1} I) \Gamma(0, 0)$$

Dále k hledání minima využijme KL divergenci a aproximační distribuce

$$q(\theta) = \mathcal{N}(\hat{\theta}, \Sigma)$$

$$q(\alpha) = \Gamma(\gamma, \delta)$$

KL divergence je tedy tvaru

$$\begin{aligned} KL(q||p) &= \int q(\theta)q(\alpha) \ln \frac{q(\theta)q(\alpha)}{P(\mathbf{y}|\theta, X)P(\theta|\alpha)P(\alpha)} d\alpha d\theta \\ &= \int q(\theta)q(\alpha) (\ln q(\theta) + \ln q(\alpha) - \ln P(\mathbf{y}|\theta, X) - \ln P(\theta|\alpha) - \ln P(\alpha)) d\alpha d\theta \end{aligned}$$

Následující členy jsou opět záporně vzaté entropie jednotlivých distribucí, víme tedy že:

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

Ostatní členy budeme nyní řešit zároveň:

$$\begin{aligned} \star &= \int q(\theta)q(\alpha) (-\ln P(\mathbf{y}|\theta, X) - \ln P(\theta|\alpha) - \ln P(\alpha)) d\alpha d\theta \\ &= \left\langle \frac{1}{2} ((\mathbf{y} - X\theta)^\top (\mathbf{y} - X\theta) + \alpha\theta^\top \theta - 5 \ln \alpha) - \ln \alpha \right\rangle \\ &= \left\langle \frac{1}{2} (\mathbf{y}^\top \mathbf{y} - \theta^\top X^\top \mathbf{y} - \mathbf{y}^\top X\theta + \theta^\top X^\top X\theta + \alpha\theta^\top \theta - 7 \ln \alpha) \right\rangle \\ &= \left\langle \frac{1}{2} (\mathbf{y}^\top \mathbf{y} - \theta^\top X^\top \mathbf{y} - \mathbf{y}^\top X\theta + \text{tr}(\theta^\top X^\top X\theta + \alpha\theta^\top \theta) - 7 \ln \alpha) \right\rangle \\ &= \left\langle \frac{1}{2} (\mathbf{y}^\top \mathbf{y} - \theta^\top X^\top \mathbf{y} - \mathbf{y}^\top X\theta + \text{tr}(X^\top X\theta\theta^\top + \alpha\theta\theta^\top) - 7 \ln \alpha) \right\rangle \\ &= \left\langle \frac{1}{2} (\mathbf{y}^\top \mathbf{y} - \theta^\top X^\top \mathbf{y} - \mathbf{y}^\top X\theta + \text{tr}((X^\top X + \alpha I)\theta\theta^\top) - 7 \ln \alpha) \right\rangle \end{aligned}$$

Po výpočtu středních hodnot dostaneme konečný výsledek, který je tvaru:

$$\star = \frac{1}{2} \left( \mathbf{y}^\top \mathbf{y} - \hat{\theta}^\top X^\top \mathbf{y} - \mathbf{y}^\top X\hat{\theta} + \text{tr} \left( \left( X^\top X + \frac{\gamma}{\delta} I \right) (\hat{\theta}\hat{\theta}^\top + \Sigma) \right) - 7(\psi(\gamma) - \ln \delta) \right)$$

Tímto máme vypočteny všechny výrazy pro optimalizaci.



# **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](http://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.