Generativní modely dat popsané stromovou strukturou

Jakub Bureš

Katedra matematiky

Vedoucí práce: doc. Ing. Václav Šmídl, Ph.D.

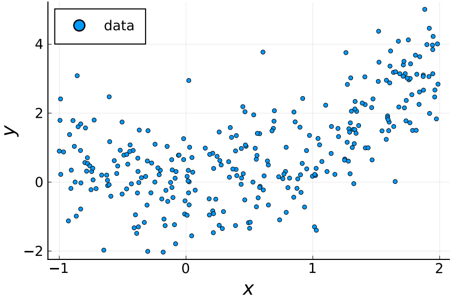
27. srpna 2020

Přehled

1. Motivace
2. Generativní modely
   - Variační autoencdoer
3. Stromové struktury
4. Otázky oponenta

Motivace

Máme k dispozici následující data a chceme predikovat nová, hledáme tedy hustotu $p(y|x)$.



Obrázek: Zadaná data, ze kterých chceme predikovat nová.

Nic těžkého $\Rightarrow$ problém vede na úlohu nejmenších čtverců.

Motivace

Model

$$\mathbf{y} = \mathbb{X} \cdot \theta + \epsilon. \tag{1}$$

Předpokládáme $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$ a položíme $X^\mathsf{T} = \left(1\ x\ x^2 \dots x^s\right)$, kde s je stupeň polynomu, jakým data prokládáme. Obdržíme tvar hustoty

$$p(y|x) = \mathcal{N}\left(X^\mathsf{T} \cdot \theta, \sigma^2\right) \tag{2}$$

Problém nastane v okamžiku, kdy budeme chtít znát například $p(y|x = 20)$, čili extrapolace mimo interval daných dat. Odpověď nemusí být přesná.
Řešení?

- $\int p(y|x, \theta)p(\theta)\mathrm{d}\theta = p(y|x)$
- **Generativní model**

Generativní modely

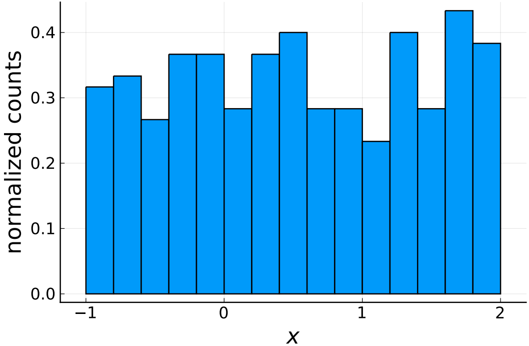
Generativní model

Mějme nějakou množinu datových záznamů $\mathbf{x} = \{x_1, \ldots, x_n\}$, představující nezávislé proměnné a nějakou množinu $\mathbf{y} = \{y_1, \ldots, y_n\}$, jakožto závislé proměnné. Generativní model je potom takový model, který se učí sdruženou hustotu pravděpodobnosti $p(x, y)$.

- Odhad hustoty pravděpodobnosti $p(x, y)$
- Součinové pravidlo $p(y, x) = p(y|x) \cdot p(x)$
- Pokusíme se tedy v dalším kroku najít $p(x)$.

Generativní model

Hustotu $p(x)$ můžeme určit například pomocí histogramu x–ových složek



Obrázek: Histogram x–ových složek zadaných dat.

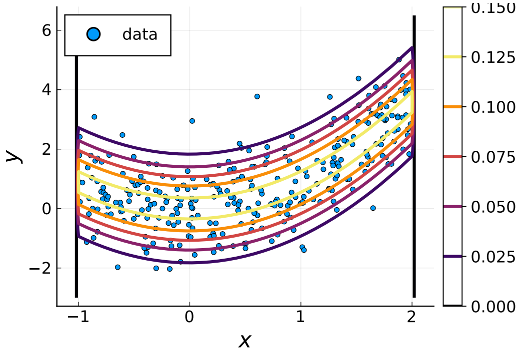
Histogram odpovídá uniformnímu rozdělení

$$p(x) = \mathrm{U}(-1, 2) \tag{3}$$

Generativní model

Pro sdruženou hustotu $p(y, x)$ pak dostaneme vztah

$$p(y, x) = \mathrm{U}(-1, 2) \cdot \mathcal{N}\left(X^{\mathsf{T}} \cdot \theta, \sigma^2\right) \tag{4}$$



Obrázek: Contour plot sdružené hustoty $p(y, x)$.

Variační autoencoder

Cílem je najít hustotu $p(\mathbf{x})$ vzorků $\{x_i\}_{i=1}^n$ za pomoci latentní proměnné $\mathbf{z}$, u které předpokládáme

$$p(\mathbf{z}) = \mathcal{N}(0, \mathbb{I}) \tag{5}$$

Jelikož lze každou náhodnou veličinu transformovat pomocí věty o transformaci náhodné veličiny, budeme hledat můžeme f_θ takovou, že

$$\mathbf{x} = f_\theta(\mathbf{z}). \tag{6}$$

Takovou funkci je obecně velmi těžké najít, proto hledáme f_θ následujícího modelu

$$\mathbf{x} = f_\theta(\mathbf{z}) + \epsilon. \tag{7}$$

kde $\epsilon \sim \mathcal{N}\left(0, \sigma^2 \cdot \mathbb{I}\right)$ a tudíž volíme $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}\left(f_\theta(\mathbf{z}), \sigma^2 \cdot \mathbb{I}\right)$. Využijeme následující formu aproximace

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z}. \tag{8}$$

Variační autoencoder

Víme-li, že

$$D_{KL}(q\|p) = \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \mathrm{d}\mathbf{x}, \tag{9}$$

můžeme použít ELBO.

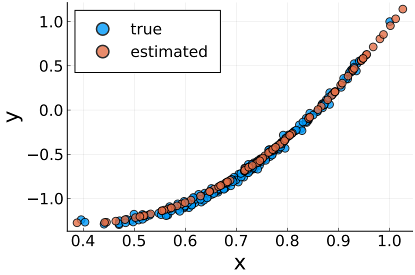
$$\begin{aligned} D_{KL}\left(q\left(\mathbf{z}|\mathbf{x}\right)\|p(\mathbf{z}|\mathbf{x})\right) &= \mathbb{E}_q\left[\log q(\mathbf{z}|\mathbf{x}) - \log p\left(\mathbf{z}|\mathbf{x}\right)\right] \\ &= \mathbb{E}_q\left[\log q(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{x}|\mathbf{z}) - \log p(\mathbf{z}) + \log p(\mathbf{x})\right]. \end{aligned} \tag{10}$$

Tuto rovnici můžeme přepsat pomocí další KL–divergence

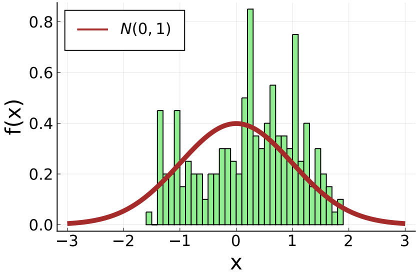
$$\log p(\mathbf{x}) - D_{KL}\left(q\left(\mathbf{z}|\mathbf{x}\right)\|p(\mathbf{z}|\mathbf{x})\right) = \mathbb{E}_q\left[\log p(\mathbf{x}|\mathbf{z})\right] - D_{KL}\left(q\left(\mathbf{z}|\mathbf{x}\right)\|p(\mathbf{z})\right), \tag{11}$$

kde pravá strana této rovnice je lower bound objektu $\log p(\mathbf{x})$, tedy ELBO.

Variační autoencoder



(a) Skutečné vzorky $\{\mathbf{x}, \mathbf{y}\}$ (modře) a jejich odhad pomocí VAE (červeně).



(b) Transformace odhadu vzorků $\{\mathbf{x}, \mathbf{y}\}$ zpět na Histogram vzorků $\mathbf{z}$

- Červené vzorky byly vygenerovány pomocí $f_\theta(z) : \mathbb{R}^1 \to \mathbb{R}^2$
- Histogram byl vygenerován pomocí $g(x, y) : \mathbb{R}^2 \to \mathbb{R}^1$

Stromové struktury

Příklad inspirovaný finanční aplikací.

- Máme k dispozici i-tého klienta, u které víme, zda–li splácel půjčku $y_i \in \{0, 1\}$, 0 značí ano, 1 značí ne.
- Dále máme k dispozici seznam jeho transakcí $X = \{x_{i,j}\}_{j=1}^{N_x}$, kde N_x počet těchto transakcí a u jednotlivých klientů se liší.
- Hodnoty a počty těchto transakcí se liší navíc, jestli daný klient splácel nebo nesplácel půjčku.
- Uvažujeme dvě gaussovské směsi (GM), $m \in \hat{a}, n \in \hat{b}$

$$\begin{aligned} p(x_m|y = 1) &= w_1 \cdot \mathcal{N}\left(\mu_1, \sigma_1^2\right) + (1 - w_1) \cdot \mathcal{N}\left(\mu_2, \sigma_2^2\right), \\ p(x_n|y = 0) &= w_2 \cdot \mathcal{N}\left(\mu_3, \sigma_3^2\right) + (1 - w_2) \cdot \mathcal{N}\left(\mu_4, \sigma_4^2\right), \end{aligned} \tag{12}$$

Stromové struktury

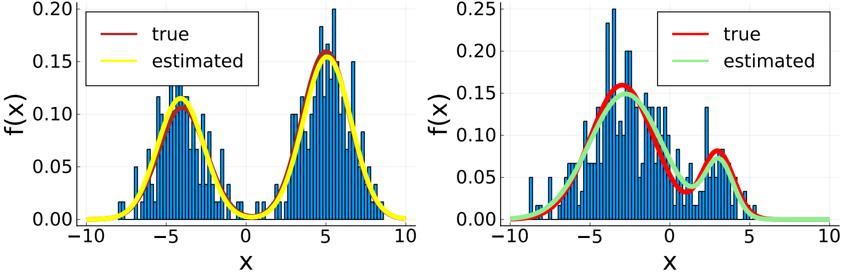
- Dále uvažujeme

$$\begin{aligned} p(N_{\mathrm{kx}}^{(1)}|y=1) &= \mathrm{Po}\left(\lambda_1\right) \quad k \in \hat{c}, \\ p(N_{\mathrm{lx}}^{(0)}|y=0) &= \mathrm{Po}\left(\lambda_2\right) \quad l \in \hat{d}. \end{aligned} \tag{13}$$

- Jak nyní rozhodnout, do které třídy (schopný či neschopný splácet) patří nový klient?
- Nejprve odhadneme všechny parametry všech rozdělení pomocí MLE.

Stromové struktury

$$
\begin{aligned}
\hat{\mu_1}, \hat{\mu_2}, \hat{\sigma_1}^2, \hat{\sigma_2}^2, \hat{w}_1 &= \underset{\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,w_1}{\arg\max} \log\left(\prod_{m=1}^{a} p(x_m|y=1)\right) \\
\hat{\mu_3}, \hat{\mu_4}, \hat{\sigma_3}^2, \hat{\sigma_4}^2, \hat{w}_2 &= \underset{\mu_3,\mu_4,\sigma_3^2,\sigma_4^2,w_2}{\arg\max} \log\left(\prod_{n=1}^{b} p(x_n|y=0)\right), \\
\hat{\lambda_1} &= \frac{1}{c}\sum_{k=1}^{c} N_{\mathrm{kx}}^{(1)}, \\
\hat{\lambda_0} &= \frac{1}{d}\sum_{l=1}^{d} N_{\mathrm{lx}}^{(0)}
\end{aligned}
\tag{14}
$$



(c) $p(x|y = 1)$ (d) $p(x|y = 0)$

Obrázek: Dvě GM, kde červenou a hnědou barvou jsou nakresleny skutečné distribuce, zeleně a žlutě jsou jejich MLE.

- MLE zde funguje i přesto, že MLE gaussovské směsi nemá analytické řešení.

Stromové struktury

- V dalším kroku sestavíme

$$
\begin{aligned}
p\left(\mathbf{x}, N_x^{(1)} | y = 1\right) &= \left(\prod_{i=1}^{N_x^{(1)}} p(x_i | y = 1) \right) \cdot p\left(N_x^{(1)} | y = 1\right), \\
p\left(\mathbf{x}, N_x^{(0)} | y = 0\right) &= \left(\prod_{j=1}^{N_x^{(0)}} p(x_j | y = 0) \right) \cdot p\left(N_x^{(0)} | y = 0\right)
\end{aligned}
\tag{15}
$$

- S jejich pomocí provedeme věrohodnostní poměry

$$
\begin{aligned}
\Lambda_0(\mathbf{x}) &= \frac{p\left(\mathbf{x}, N_x^{(0)} | y = 0\right)}{p\left(\mathbf{x}, N_x^{(1)} | y = 1\right) + p\left(\mathbf{x}, N_x^{(0)} | y = 0\right)} \in \langle 0, 1 \rangle \\
\Lambda_1(\mathbf{x}) &= \frac{p\left(\mathbf{x}, N_x^{(1)} | y = 1\right)}{p\left(\mathbf{x}, N_x^{(1)} | y = 1\right) + p\left(\mathbf{x}, N_x^{(0)} | y = 0\right)} \in \langle 0, 1 \rangle.
\end{aligned}
\tag{16}
$$

Děkuji za pozornost.

Co je to regulární zobrazení ve větě o transformaci náhodné veličiny?

Jaký je rozdíl mezi metrikou a semimetrikou?

Proč je jasné, že

$$\bar{x}_i = \frac{1}{N_x^{(i)}} \sum_{l=1}^{N_x^{(i)}} x_l^{(i)} \tag{17}$$

bude mít normální rozdělení? Platí to obecně nebo za nějakých předpokladů?