13 VZD – Naivný Bayesov klasifikátor, modely podmienených pravdepodobností

Matej Choma

June 3, 2019

Pri klasifikácií na základe podmienenej pravdepodobnosti uvažujeme klasifikačnú úlohu, kde na základe p diskrétnych príznakov reprezentovaných $\mathbf{X} = (X_1, \ldots, X_p)^T, \mathbf{X} \in \mathcal{X}$ predikujeme diskrétnu predikovanú veličinu $Y \in \mathcal{Y}$. Z dát v trénovacej množine odhadneme pravdepodobnosti $\mathrm{P}(Y = y|\mathbf{X} = x)$ pre všetky $y \in \mathcal{Y}$ a $x \in \mathcal{X}$. Samotná predikcia na základe napozorovaných príznakov $\mathbf{x}$ sa nazýva **MAP odhad** (z angl. *maximum a posteriori*):

$$\hat{Y} = \arg\max_{y \in \mathcal{Y}} \mathrm{P}(Y = y|\mathbf{X} = \mathbf{x}).$$

Za predpokladu, že sa nám podarí odhadnúť $\mathrm{P}(\mathbf{X} = \mathbf{x}|Y = y)$, môžeme pri odhadovaní $\mathrm{P}(Y = y|\mathbf{X} = \mathbf{x})$ využiť Bayesovu vetu:

$$\mathrm{P}(Y = y|\mathbf{X} = \mathbf{x}) = \frac{\mathrm{P}(\mathbf{X} = \mathbf{x}|Y = y) \cdot \mathrm{P}(Y = y)}{\mathrm{P}(\mathbf{X} = \mathbf{x})},$$

kde

$$\mathrm{P}(\mathbf{X} = \mathbf{x}) = \sum_{y \in \mathcal{Y}} \mathrm{P}(\mathbf{X} = \mathbf{x}|Y = y) \cdot \mathrm{P}(Y = y).$$

Odhadnúť pravdepodobnosť $\mathrm{P}(Y = y)$, ktorú potrebujeme, je triviálne. Pre všetky hodnoty y ostáva $\mathrm{P}(\mathbf{X} = \mathbf{x})$ rovnaké a teda finálne pre predikciu dostávame:

$$\hat{Y} = \arg\max_{y \in \mathcal{Y}} \mathrm{P}(\mathbf{X} = \mathbf{x}|Y = y) \cdot \mathrm{P}(Y = y).$$

1 Naivný Bayesov klasifikátor

Naivný Bayesov klasifikátor (angl. *Naive Bayes*) popisuje spôsob ako odhadovať pravdepodobnosti $\mathrm{P}(\mathbf{X} = \mathbf{x}|Y = y)$.

Naivita v naivnom Bayesovi predpokladá, že **pre zafixované** $Y = y$ **sú všetky príznaky nezávislé**. Formálne, pre všetky $y \in \mathcal{Y}$ a $\mathbf{x} = (x_1, \ldots, x_p)^T \in \mathcal{X}$ platí:

$$\mathrm{P}(\mathbf{X} = \mathbf{x}|Y = y) = \mathrm{P}(X_1 = x_1|Y = y) \cdot \ldots \cdot \mathrm{P}(X_p = x_p|Y = y).$$

Výsledný **MAP odhad** pre naivného Bayesa teda vyzerá nasledovne:

$$\hat{Y} = \arg\max_{y \in \mathcal{Y}} \prod_{i=1}^{p} \mathrm{P}(X_i = x_i|Y = y) \cdot \mathrm{P}(Y = y).$$

1.1 Vlastnosti naivného Bayesa

Napriek tomu, že predpoklad nezávislosti príznakov je väčšinou nesprávny, má model niekoľho dobrých vlastností a občas dáva až prekvapivo dobré výsledky.

Vďaka rozkladu združenej podmienenej pravdepodobnosti $\mathrm{P}(\mathbf{X} = \mathbf{x}|Y = y)$ na súčin marginálnych ostávajú príznaky separované – odhad podmienej pravdepodobnosti každého príznaku prebieha nezávisle od ostatných. To značne pomáha proti problémom s dimenzionalitou (angl. *curse of dimensionality*). Množstvo dát potrebné na odhad $\mathrm{P}(X_i = x_i|Y = y)$ nenarastá s množstvom príznakov.

Z dôvodu nepresného predpokladu býva odhad $\mathrm{P}(\mathbf{X} = \mathbf{x}|Y = y)$ nie dobrý. Nás ale zaujíma MAP odhad a teda pokiaľ je odhadnutá pravdepodobnosť pre y väčšia, ostáva tento odhad správny. Emipricky sa ukazuje, že je toto častá situácia.

2 Modely naivného Bayesa

V tejto časti sa budeme zaoberať problematikou odhadu $\mathrm{P}(X = x|Y = y)$, kde X je jeden z príznakov.

2.1 Modely podmienených pravdepodobností - Bernoulliho rozdelenie

V najjednoduchšom prípade nadobúda X hodnoty 0, 1. Jedná sa o *Bernoulliho rozdelenie* s parametrom $p_y = \mathrm{P}(X = 1|Y = y)$. Značí sa $(X|Y = y) \sim \mathrm{Be}(p_y)$.

Ako **odhad parametru** p_y sa najčastejšie používa

$$\hat{p}_y = \frac{N_{1,y}}{N_{0,y} + N_{1,y}},$$

kde $N_{j,y}$ značí počet dát pre $X = j$ a $Y = y$. Z pohľadu matematickej štatistiky sa jedná o odhad maximálnej vierohodnosti **MLE** parametru Bernoulliho rozdelenia.

2.1.1 Bayesovský prístup k odhadom

V prípade veľkého (alebo malého) p_y sa môže stať, že jedna z hodnôt príznaku nie je v trénovacích dátach zastúpená. Odhadnutá pravdepodobnosť $\mathrm{P}(X = x|Y = y)$ je potom rovná 0, čím sa vylúči možnosť modelu predikovať dané y napriek tomu, že na základe ostatných príznakov by to bolo možné. Riešením je do odhadu parametru zakomponovať našu expertnú znalosť ako apriorné rozdelenie[1].

Pre rovnomerné rozdelenie parametru dostávame upravený odhad

$$\hat{p}_y = \frac{N_{1,y} + 1}{N_{0,y} + N_{1,y} + 2}.$$

Tento odhad sa nazýva *add-one smoothing* a netrpí na kolaps, pokiaľ sa jedna z hodnôt nenachádza v trénovacej množine.

2.2 Modely podmienených pravdepodobností - Kategorické rozdelenie

Pokiaľ príznak X nadobúda k rôznych hodnôt $c_1, \ldots, c_k$, hovoríme o kategorickom rozdelení $(X|Y = y) \sim \mathrm{Cat}(\mathbf{p}_y)$ s paramterom $\mathbf{p}_y = (p_{1,y}, \ldots, p_{k,y})^T$ a $p_{j,y} = \mathrm{P}(X = c_j|Y = y)$.

Ako **odhad k-rozmerného parametru** $\mathbf{p}_y$ sa najčastejšie používa $\hat{\mathbf{p}}_y = (\hat{p}_{1,y}, \ldots, \hat{p}_{k,y})_T$, kde

$$\hat{p}_{j,y} = \frac{N_{j,y}}{N_{1,y} + \ldots + N_{k,y}}$$

a $N_{j,y}$ značí počet dát pre $X = c_j$ a $Y = y$.

$\hat{p}_{j,y}$ môže byť opäť 0 a s Bayesovským prístupom sa dá získať robustnejší odhad

$$\hat{p}_{j,y} = \frac{N_{j,y} + 1}{N_{1,y} + \ldots + N_{k,y} + k}$$

[1] Presný postup je popísaný v handoute BI-VZD ku 6. prednáške na 57. strane. Verím, že to nie je pre túto otázku potrebné.

2.3 Modely podmienených pravdepodobností - Gaussovo rozdelenie

V prípade, že príznak X je spojitou náhodnou veličinou, berieme do úvahy namiesto podmienenej pravdepodobnosti $\mathrm{P}(X = x|Y = y)$ radšej podmienenú hustotu pravdepodobnosti $f_{X|y}(x)$. Táto hustota zodpovedá distribučnej funkcii

$$F_{X|y}(x) = \mathrm{P}(X \leq x|Y = y),$$

a **MAP odhad** sa uskutoční ako

$$\hat{Y} = \arg\max_{y \in \mathcal{Y}} \prod_{i=1}^{l} \mathrm{P}(X_i = x_i|Y = y) \cdot \prod_{i=l+1}^{p} f_{X_i|y}(x_i) \cdot \mathrm{P}(Y = y),$$

kde príznaky $X_1, \ldots, X_l$ sú diskrétne a príznaky $X_l + 1, \ldots, X_p$ spojité.

Častým modelom je $(X|Y = y) \sim \mathrm{N}(\mu_y, \sigma_y^2)$. Podmienená hustota je potom pre každé $x \in \mathbb{R}$ určená vzťahom

$$f_{X|y}(x) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \mathrm{e}^{-\frac{1}{2\sigma_y^2}(x-\mu_y)^2}.$$

Pre odhad parametrov sa obvykle používajú MLE odhady

$$\hat{\mu}_y = \frac{1}{N_y} \sum_{i=1}^{N_y} x_i \qquad \text{a} \qquad \hat{\sigma}_y^2 = \frac{1}{N_y} \sum_{i=1}^{N_y} (x_i - \hat{\mu}_y)^2,$$

kde $x_1, \ldots, x_{N_y}$ sú hodnotý príznaku X pre ktoré platí $Y = y$.

3 Poznámky

Naivný Bayes nám dáva model pre združenú pravdepodobnosť

$$\mathrm{P}(\mathbf{X} = \mathbf{x}, Y = y) = \mathrm{P}(\mathbf{X} = \mathbf{x}|Y = y)\mathrm{P}(Y = y).$$

Takýto prístup sa nazýva **generatívny** a dáva nám úplnú informáciu o rozdelení, z ktorého boli dáta “generované”. **Generatívny model** môže byť teda použítý na vygenerovanie nových pozorovaní.

Opakom je **diskriminatívny** prístup, ktorý odhaduje $\mathrm{P}(Y = y|\mathbf{X} = \mathbf{x})$ z trénovacích dát priamo. Tento prístup je veľmi častý, napríklad v logistickej reresií alebo v neurónových sieťach.