Bayesovské učení

Jak složitě odvodit známé věci
nové pojmy: nejpravděpodobnější hypotéza (MAP), maximálně věrohodná hypotéza (ML), bayesovsky optimální predikce

Jednotlivé predikce budeme uvádět na následujícím příkladu (který se bude postupně zesložiťovat).

Předpokládejme, že výrobce vyrábí velké pytle bombónů. Bombóny jsou dvou druhů – buď plněné (l), které mi nechutnají, nebo celé (c), které jsou výborné. Výrobce ale chce provokovat zákazníka, tak balí oba druhy do stejného obalu a navíc do velkého pytle namíchá oba druhy v jednom z následujících poměrů (procent celých bombónů):

hypotéza	h_1	h_2	h_3	h_4	h_5
procent celých b.	100%	75%	50%	25%	0%
apriorní pravděp. hyp.	10%	20%	40%	20%	10%

První bombón je celý (neplněný). Otázka zní, jaký bude další bombón, vytažený ze stejného pytle?

nejpravděpodobnější hypotéza

MAP – maximum aposteriory probability

Předpokládáme, že pytel bonbónů je velký a můžeme zanedbat změnu pravděpodobnosti způsobenou tím, že už v pytli jeden bonbón chybí. Vzpomeneme si na bayesův vzorec:

$$P(h_i|B=c) = \frac{P(B=c|h_i) \cdot P(h_i)}{\sum_{j=1,\ldots,5} P(B=c|h_j) \cdot P(h_j)} = \frac{P(B=c|h_i) \cdot P(h_i)}{P(B=c)}$$

Hledáme nejpravděpodobnější hypotézu $argmax_i P(h_i|B=c) = argmax_i P(B=c|h_i) \cdot P(h_i)$. Pro jednotlivá i spočteme hodnotu v tabulce:

i	$P(h_i)$	$P(B = c \mid h_i)$	$P(B = c \mid h_i) \cdot P(h_i)$	$P(h_i \mid B = c)$
1	0,1	1	0,1	0,2
2	0,2	0,75	0,15	0,3
3	0,4	0,5	0,2	**0,4**
4	0,2	0,25	0,05	0,1
5	0,1	0	0	0

Nejpravděpodobnější hypotéza je tedy h_3, která nám předpoví další bonbón padesát na padesát.

$h_{MAP} = argmax_i P(data|h_i) \cdot P(h_i)$

Pozn: nejpravděpodobnější odhad koresponduje s MDL– minimal description length principem.

$$\begin{aligned} h_{MAP} &= argmax_h P(data|h)P(h) \\ &= argmin_h[-log_2 P(data|h) - log_2 P(h)] \end{aligned}$$

kde první logaritmus odpovídá kódování dat při známé hypotéze a druhý minimálnímu počtu bitů potřebných k zakódování hypotézy.

Bayesovsky optimální odhad

Bayesovsky optimální odhad váží předpovědi jednotlivých hypotéz pravděpodobností těchto hypotéz, tj.

$$\begin{aligned} P(N = c|data) &= \sum_{j=1,\ldots,5} P(N = c|h_j, data) \cdot P(h_j|data) \\ &= \sum_{j=1,\ldots,5} P(N = c|h_j) \cdot P(h_j|data) \end{aligned}$$

Druhý přechod můžeme udělat korektně pouze za předpokladu podmíněné nezávislosti nových pozorování a starých pozorování při znalosti hypotézy h_j. Tato nezávislost je splněna např. pokud jsou jednotlivá pozorování identicky nezávisle rozložená, zkratka iid - independently identically distibuted.

i	$P(h_i \mid B = c)$	$P(N = c \mid h_i)$	$P(N = c \mid h_i) \cdot P(h_i \mid B = c)$
1	0,2	1	0,2
2	0,3	0,75	0,225
3	0,4	0,5	0,2
4	0,1	0,25	0,02
5	0	0	0
Σ			0,645

Maximálně věrohodná hypotéza

ML - maximum likelihood

Často se stává, že **neznám apriorní pravděpodobnost jednotlivých hypotéz**. V tom případě se bere často rovnoměrná distribuce, tj. považujeme všechny hypotézy za stejně pravděpodobné, a nejpravděpodobnější odhad za tohoto přidaného předpokladu nazýváme **maximálně věrohodný odhad**, tj.

$$h_{ML} = argmax_i P(data|h_i)$$

Příklad na odhad parametrů

ML - maximum likelihood

Předpokládejme, že se na trhu objevil nový výrobce a my neznáme jeho apriorní pravděpodobnosti jednotlivých poměrů, ani poměry, ve kterých bonbóny míchá.

Chci odhadnout poměr celých bonbónů, tj. můj prostor hypotéz bude h_θ, kde $\theta \in \langle 0; 1 \rangle$. Každá h_θ odpovídá doměnce že podíl celých bonbónů je právě θ.

Neznám pravděpodobnost jednotlivých hypotéz, proto hledám maximálně věrohodný odhad.

Při dané hypotéze h_θ je pravděpodobnost vytažení dané posloupnosti c celých a l plněných bonbónů:

$$P(data|h_\theta) = \theta^c \cdot (1 - \theta)^l$$

Častým trikem je rovnici zlogaritmovat:

$$L(data|h_\theta) = c \cdot \log_2 \theta + l \cdot \log_2 (1-\theta)$$

Funkce L je logaritmus věrohodnosti hypotézy, česky **log likelihood**.

Hledáme maximum přes všechna θ, tj. místo, kde se derivace rovná nule (či kraj..):

$$\begin{aligned} \frac{\partial L(data|h_\theta)}{\partial \theta} &= \frac{c}{\theta} - \frac{l}{1-\theta} \\ \frac{c}{\theta} &= \frac{l}{1-\theta} \\ \theta &= \frac{c}{c+l} \end{aligned}$$

Druhý příklad

Výrobce zavedl různobarevné obaly O (zelený a modrý) a do obou balí oba druhy bonbónů, ale pro každý druh bonbónu má jinou pravděpodobnost volby zeleného obalu.

Pro tento případ potřebujeme v modelu tři parametry:

$P(B = c)$	$P(O = z \mid B = c)$	$P(O = z \mid B = l)$
θ_0	θ_1	θ_2

a označíme si pozorované četnosti viz následující tabulka:

obal$\backslash typ$	celý	plněný
zelený	z_c	z_l
modrý	m_c	m_l

Pravděpodobnost dat za dané hypotézy $h_{\theta_0,\theta_1,\theta_2}$ je:

$$
\begin{aligned}
P(data|h_{\theta_0,\theta_1,\theta_2}) &= \theta_1^{z_c} \cdot (1-\theta_1)^{m_c} \cdot \theta_0^{z_c+m_c} \cdot \theta_2^{z_l} \cdot (1-\theta_2)^{m_l} \cdot (1-\theta_0)^{z_l+m_l} \\
L(data|h_{\theta_0,\theta_1,\theta_2}) &= z_c \log_2 \theta_1 + m_c \log_2 (1-\theta_1) + (z_c+m_c) \log_2 \theta_0 + z_l \log_2 \theta_2 \\
&\quad + m_l \log_2 (1-\theta_2) + (z_l+m_l) \log_2 (1-\theta_0) \\
\frac{\partial L(data|h_{\theta_0,\theta_1,\theta_2})}{\partial \theta_0} &= \frac{z_c+m_c}{\theta_0} - \frac{z_l+m_l}{1-\theta_0} \\
\theta_0 &= \frac{(z_c+m_c)}{z_c+m_c+z_l+m_l} \\
\frac{\partial L(data|h_{\theta_0,\theta_1,\theta_2})}{\partial \theta_2} &= \frac{z_l}{\theta_2} - \frac{m_l}{1-\theta_2} \\
\theta_2 &= \frac{z_l}{z_l+m_l}
\end{aligned}
$$

Odhad parametrů Bayesovské sítě

při úplných datech se redukuje na spočtení podílu odpovídajících frekvencí.

Naive Bayes model

Pokud nechceme učit složitě strukturu, tak se často používá naivní bayesovský model, který je velmi snadné učit a často funguje překvapivě dobře.

Spojité veličiny

Mějme hypotézu, že je naše x normálně rozložené, jen neznám μ a σ, tj. $h_{\mu,\sigma} = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$.

Pozorovali jsme $x_1, \ldots, x_n$, maximálně věrohodná hypotéza je:

$$\mathrm{L}=\sum_{j=1}^{N} log \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_j-\mu)^2}{2\sigma^2}}$$

$$=\mathrm{N}\cdot\left(log \frac{1}{\sqrt{2\pi}\sigma}\right) - \sum_{j=1}^{N} \frac{(x_j-\mu)^2}{2\sigma^2}$$

Položíme derivace podle μ, σ rovny nule a dostaneme maximálně věrohodné odhady.

Spočtěte.

Lineární gausovské rozložení

- Mějme veličinu X, normálně rozloženou (na tom ale nezáleží)
- mějme veličinu Y, normálně rozloženou, s pevným rozptylem σ, jejíž střední hodnota μ závisí lineárně na X, tj. $\mu = a \cdot x + b$;

$$P(Y|X = x) = N(ax + b, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(y-(ax+b))^2}{2\sigma^2}}$$

- máme parametry a, b, σ. Můžeme derivací určit maximálně věrohodný odhad.
- Hádanka: Považujme σ za pevné, parametry nechme jen a, b.

  Pak maximalizujeme $e^{\frac{-(y-(ax+b))^2}{2\sigma^2}}$ součin přes všechny příklady.

  Zlogaritmujte, součin se změní na součet, změňte znaménko a minimalizujte - co minimalizujete?

BIC kriterium

- Mám hypotézy (např. struktury BN) h_m, $m = 1, \ldots, M$
- v každé musím doladit parametry θ_m (parametry BN, ale i např. spočítat průměr, regresní β, atd.)
- chci nejpravděpodobnější model při znalosti pozorovaných dat $DATA$, tj.

$$argmax_{h_m} P(h_m|DATA)$$

$$
\begin{aligned}
P(h_m|DATA) &= \frac{1}{P(DATA)} \cdot P(h_m) \cdot P(DATA|h_m) \\
&= \frac{P(h_m)}{P(DATA)} \int P(DATA|\theta_m, h_m) P(\theta_m|h_m) d\theta_m
\end{aligned}
$$

Integrál aproximujeme (Laplace aprox.) a trochu zjednodušíme a dostaneme:

$$
logP(DATA|h_m) = logP(DATA|\hat{\theta}_m, h_m) - \frac{d_m}{2} \cdot logN + O(1)
$$

- $\hat{\theta}_m$ maximálně věrohodný odhad θ_m,
  tj. maximalizující $P(DATA|\theta_m, h_m)$ pro pevná $DATA$ a h_m
- d_m počet parametrů modelu h_m

BIC kriterium

- místo maximalizace $P(DATA|h_m)$ budeme minimalizovat $-2 \cdot logP(DATA|h_m)$, tj.

$$BIC = -2 \cdot logP(DATA|\hat{\theta}_m, h_m) + d \cdot logN$$

- značíme $loglik = \sum_{i=1}^{N} logP_{\hat{\theta}}(y_i)$, dostaneme:
- $BIC = -2 \cdot loglik + d \cdot logN$
- $AIC = -2 \cdot loglik + d \cdot 2$
- MDL odpovídá BIC, jinak odvozeno

Bayesovské učení parametrů

Pokud máme málo dat, můžeme místo odhadování parametrů držet celé pravděpodobnostní rozložení na parametrech, tj. odhadovat hyperparametry.

Pravděpodobnostní rozložení na parametru θ má beta rozložení $beta[a, b]$, kde a je počet pozitivních příkladů a b počet negativních příkladů (snižujících θ).

$$beta[a, b](\theta) = \alpha\theta^{a-1}(1-\theta)^{b-1}$$

Při odhadu více parametrů předpokládáme jejich nezávislost.

Parametry θ lze dokreslit do BN, namnožit uzly pro jednotlivé příklady dat a propagací spočítat rozložení na θ.