Odhady parametrů
Metoda maximální věrohodnosti
Bayesovské odhadování

Jan Kolorenč, Petr Kujan

27. května 2002

1 Motivace

Cílem rozpoznávání je výběr nejlepší rozhodovací strategie. Jedním z nejčastějších kritérií výběru strategie je minimalizace střední ztráty při použití zvolené strategie.

$$R(d) = \sum_{k \in K} \sum_{\mathbf{x} \in X} P(\mathbf{x}, k) W(k, d(\mathbf{x})) \tag{1}$$

Kde X je množina příznaků x, K množina skrytých parametrů k, D množina rozhodnutí d, $W : K \times D \to R$ označuje ztrátu způsobenou rozhodnutím d, pokud je objekt ve stavu k. Kritérium většinou nelze vypočítat, neboť není známa funkce $P(\mathbf{x}, k)$, proto se nahrazuje jinými kritérii, která, za jistých předpokladů, vyčíslit lze a blíží se původnímu. Náhradní kritéria získávají dodatečnou informaci učením z trénovací množiny.

2 Učení

Funkci $P(\mathbf{x}, k)$ lze napsat jako součin $P(\mathbf{x}|k)P(k)$, kde odhad apriorní pravděpodobnosti zjistíme relativně snadno, jako relativní četnost na trénovací množině, pokud není trénovací množina vychýlena vlivem nepozorovaných parametrů. Odhadnout aposteriorní pravděpodobnost je již složitější. Ve většině případů je známa třída funkcí odpovídající aposteriorní pravděpodobnosti a jsou k dispozici soubory vzorků s ohodnocením (trénovací množina). Jednou z možností odhadu aposteriorní pravděpodobnosti je parametrizace třídy funkcí a odhad parametrů z trénovací množiny. Následující dvě metody odhadu $P(\mathbf{x}|k)$ dávají téměř identické výsledky, ale přístupy se značně liší. Metoda maximální věrohodnosti pohlíží na parametry jako na fixní neznámé hodnoty a maximalizuje pravděpodobnost získání trénovací množiny. Bayesova metoda bere parametry jako náhodnou veličinu se známou apriorní pravděpodobností $P(\mathbf{x})$, výskyt vzorků trénovací množiny převádí na aposteriorní pravděpodobnost. Výsledné číslo z vypočítané hustoty pravděpodobnosti je nutno vybrat dle zvoleného kritéria. Bayesův přístup s každým vzorkem zostřuje hledané parametry. V obou případech se jedná o učení s učitelem, neboť trénovací množina obsahuje pro každý její prvek hodnotu skrytého stavu. Další možnosti odhadu parametrů jsou: metoda momentů - srovnávají se výběrové momenty s momenty příslušného rozdělení, intervalové odhady - hledají parametr tak, aby se s danou pravděpodobností vyskytoval v daném intervalu, jako náhodné veličiny se berou meze intervalu.

2.1 Předpoklady

Dále uvedené metody předpokládají znalost třídy funkcí rozdělení pravděpodobnosti až na určitý počet parametrů, které odhadují z trénovací množiny nebo multimnožiny $T = ((x_1, k_1), (x_2, k_2) \ldots, (x_n, k_n)),\ x_i \in X,\ k_i \in K$. obsahující vzájemně nezávislé vzorky. Například pro normální rozdělení jsou parametry μ a σ.

2.2 Metoda maximální věrohodnosti (ML odhad)

Nechť trénovací množina obsahuje vzájemně nezávislé jevy s rozdělením pravděpodobnosti $P(\mathbf{x}|k)$ známého tvaru, až na konečný počet parametrů $\mathbf{q}$. $P(\mathbf{x}|\mathbf{q})$ je podmíněná pravděpodobnost pozorování x za podmínky $\mathbf{q}$, $\mathbf{q}$ jsou hledané parametry vyjadřující charakteristiky skrytých stavů, a $P(\mathbf{x}, \mathbf{q}) = P(\mathbf{x}|\mathbf{q})P(\mathbf{q})$. Pak pravděpodobnost realizace trénovací množiny za předpokladu parametrů je

$$L = P(T|\mathbf{q}) = P(x_1, x_2 \ldots, x_n|\mathbf{q}) = \prod_{i=1}^{N} P(x_i|\mathbf{q}) \tag{2}$$

Vztah (2) se nazývá věrohodnost. Kritérium maximální věrohodnosti se snaží nalézt takové parametry $\mathbf{q}^*$, které maximalizují pravděpodobnost (2)

$$\mathbf{q}^* = \operatorname*{argmax}_{\mathbf{q}^*} \prod_{i=1}^{N} P(x_i|\mathbf{q}) \tag{3}$$

Příklad 2.1 Odhad parametrů pro normální rozdělení jednoho skrytého stavu
Normální rozložení se střední hodnotou μ a rozptylem σ^2 je dáno hustotou pravděpodobnosti

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Pro danou trénovací množinu hledáme parametry μ, σ, aby věrohodnost

$$P(T|\mu,\sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

byla maximální. Věrohodnost můžeme transformovat libovolnou rostoucí funkcí, aniž by se změnila poloha maxima. Výhodný, nejen pro tento příklad, je logaritmus, který změní součiny v součty.

$$L = \ln P(T|\mu,\sigma) = \sum_{i=1}^{N} -\ln\sqrt{2\pi} - \ln\sigma - \frac{(x_i-\mu)^2}{2\sigma^2}$$

Hledáním maxima zjistíme požadované parametry

$$\frac{\partial L}{\partial \mu} = 0 \Rightarrow -\frac{\sum_{i=1}^{N} 2(x_i-\mu)}{2\sigma^2} = 0 \Rightarrow \mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\frac{\partial L}{\partial \sigma} = 0 \Rightarrow -\frac{N}{\sigma} + \sum_{i=1}^{N} 2\frac{(x_i-\mu)}{2\sigma^3} = 0 \Rightarrow \sigma^2 = \frac{\sum_{i=1}^{N} (x_i-\mu)^2}{N}$$

Příklad 2.2 Odhad parametrů pro normální rozdělení v n-rozměrném prostoru
Normální rozdělení je dáno hustotou

$$P(x) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det \mathbf{C}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{C}^{-1}(\mathbf{x}-\mu)}$$

μ je střední hodnota normálního rozdělení, $\mathbf{C}$ je disperzní matice normálního rozdělení, symetrická, pozitivně definitní. Jsou-li vstupem vektory $\mathbf{x}_1 \ldots \mathbf{x}_l$, pak podmíněná hustota realizace trénovací množiny τje

$$P(\tau|\mu,\mathbf{C}) = \prod_{i=1}^{l} \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det \mathbf{C}}} e^{-\frac{1}{2}(\mathbf{x}_i-\mu)^T \mathbf{C}^{-1}(\mathbf{x}_i-\mu)}$$

Transformací logaritmem ze stejného důvodu jako v předcházejícím případě dostaneme

$$L = \ln(P(\tau|\mu, \mathbf{C})) = -\frac{nl}{2}\ln(2\pi) - \frac{l}{2}\ln(\det \mathbf{C}) - \frac{1}{2}\sum_{i=1}^{l}(\mathbf{x}_i - \mu)^T \mathbf{C}^{-1}(\mathbf{x}_i - \mu) \tag{4}$$

Hledáme parametry μ *a* $\mathbf{C}$ *maximalizující následující věrohodnost (4). Extrémy (4) zjistíme parciálními derivacemi podle složek parametrů.*

$$\frac{\partial L}{\partial \mu_s} = \Big[\mathbf{C}^{-1}(\sum_{i=1}^{l}\mathbf{x}_i - l\mu)\Big]_s$$

Tedy

$$\mathbf{C}^{-1}\Big(\frac{1}{l}\sum_{i=1}^{l}\mathbf{x}_i - \mu\Big) = 0$$

Protože $\mathbf{C}$ *je pozitivně definitní, pak*

$$\mu = \frac{1}{l}\sum_{i=1}^{l}\mathbf{x}_i$$

$$\frac{\partial L}{\partial C^{-1}_{s,t}} = \frac{l}{2\det C^{-1}}\frac{\partial \det C^{-1}}{\partial C^{-1}_{s,t}} - \frac{1}{2}\sum_{i=1}^{l}[(\mathbf{x}_i - \mu)^T]_s[(\mathbf{x}_i - \mu)^T]_t$$

Aplikací $\frac{\partial L}{\partial C^{-1}_{s,t}} = 0$ *dostáváme*

$$C_{t,s} = \frac{1}{l}\sum_{i=1}^{l}[(\mathbf{x}_i - \mu)^T]_s[(\mathbf{x}_i - \mu)^T]_t$$

Odhad disperzní matice je dán

$$\mathbf{C} = \frac{1}{l}\sum_{i=1}^{l}(\mathbf{x}_i - \mu)^T(\mathbf{x}_i - \mu)^T$$

Příklad 2.3 Zjištění parametrů neznámé mince
$p = P(orel) = 1 - P(pana)$, *z* N *měření padl orel* k *krát. Předpokládáme rovnoměrné rozložení s parametrem* p.

$$L = P(T|p) = p^k(1-p)^{N-k}$$

$$\frac{\partial L}{\partial p} = 0,\ kp^{k-1}(1-p)^{N-k} - (N-k)p^k(1-p)^{N-k-1} = 0 \Rightarrow p = \frac{k}{N}$$

Výsledek souhlasí s frekvenční definicí pravděpodobnosti.

2.3 Bayesovské odhadování

Odhad vychází z definice rizika klasifikace popsaného vztahem (1), který modifikuje pro konkrétní $\mathbf{x}$, tedy trénovací množinu. Nechť $P(\mathbf{x}|\mathbf{q})$ je podmíněná pravděpodobnost pozorování $\mathbf{x}$ za podmínky $\mathbf{q}$, $\mathbf{q}$ jsou hledané parametry, a trénovací množina obsahuje vzájemně nezávislé jevy s rozdělením pravděpodobnosti $P(\mathbf{x}, q) = P(\mathbf{x}|\mathbf{q})P(\mathbf{q})$. Střední riziko klasifikace pro trénovací množinu je

$$R = \sum_T \sum_{\mathbf{q}} P(T, \mathbf{q}) W(\mathbf{q}, \mathbf{q}^*(T)) \tag{5}$$

Cílem je najít takové rozhodnutí $\mathbf{q}^*$, které minimalizuje (5). Známe-li všechny potřebné pravděpodobnosti, pak minimalizující řešení je

$$\mathbf{q}^* = \operatorname*{argmin}_{\mathbf{q}^*} \sum_{\mathbf{q}} P(T, \mathbf{q}) W(\mathbf{q}, \mathbf{q}^*) \tag{6}$$

Vhodnou ztrátovou funkcí může být ve smyslu nejmenších čtverců

$$W(\mathbf{q}, \mathbf{q}^*) = (\mathbf{q} - \mathbf{q}^*)^2 \tag{7}$$

Potom lze vyjádřit 6

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{q}^*}(\sum_{\mathbf{q}} P(T, \mathbf{q})(\mathbf{q} - \mathbf{q}^*)^2) &= 0 \\
\sum_{\mathbf{q}} 2\, P(T, \mathbf{q})\, (\mathbf{q} - \mathbf{q}^*) &= 0 \\
\sum_{\mathbf{q}} P(T, \mathbf{q})\, \mathbf{q} &= \mathbf{q}^* \sum_{\mathbf{q}} P(T, \mathbf{q})
\end{aligned}$$

$$\mathbf{q}^* = \frac{\sum_{\mathbf{q}} \mathbf{q}\, P(T, \mathbf{q})}{\sum_{\mathbf{q}} P(T, \mathbf{q})} \tag{8}$$

Protože $\sum_{\mathbf{q}} P(T, \mathbf{q}) = P(T)$ můžeme výsledek přepsat jako

$$\mathbf{q}^* = \frac{\sum_{\mathbf{q}} \mathbf{q}\, P(T, \mathbf{q})}{P(T)} = \sum_{\mathbf{q}} \mathbf{q}\, P(\mathbf{q}|T) = \frac{\sum_{\mathbf{q}} \mathbf{q}\, P(T|\mathbf{q})\, P(\mathbf{q})}{P(T)} \tag{9}$$

$P(\mathbf{q})$ vystihuje naši znalost o pravděpodobnosti výskytu jednotlivých hodnot q_i ještě před provedením experimentu. Z (9) je vidět, že $(q)^*$ se určí jako těžiště aposteriorní podmíněné pravděpodobnosti $\sum_{\mathbf{q}} \mathbf{q}\, P(\mathbf{q}|T)$

Příklad 2.4 Zjištění parametrů neznámé mince
Zadání je stejné jako v příkladu 2.3. Parametr $p = P(orel)$, $W = (p - p^*)^2$. *Jelikož nemáme žádnou informaci o pravděpodobnosti jednotlivých hodnot* q_i, *zvolíme* $P(q_i) = 1$, *na intervalu* $< 0,1 >$, $P(q_i) = 0$ *jinak, tedy všechny hodnoty jsou stejně pravděpodobné.*

$$p^* = \frac{\int p\,P(T,p)\,dp}{\int P(T,p)\,dp}$$
$$P(T,p) = p^k(1-p)^{N-k}$$
$$p^* = \frac{\int_0^1 p^{k+1}(1-p)^{N-k}\,dp}{\int_0^1 p^k(1-p)^{N-k}\,dp}$$

Integrací per partes

$$f_N(k) = \int_0^1 x^k(1-x)^{N-k}\,dx =$$
$$= \left[\frac{x^{k+1}}{k+1}(1-x)^{N-k}\right]_0^1 - \int_0^1 \frac{x^{k+1}}{k+1}(k-N)(1-x)^{N-k-1}\,dp =$$
$$= 0 + \frac{N-k}{k+1} f_N(k)$$

dostáváme rekurzivní vztah. Pro $f_N(0) = \int_0^1 (1-x)^N\,dx = \frac{1}{N+1}$, *tedy* $f_N(k) = \frac{1}{N+1}\frac{1}{N}\frac{2}{N-1}\cdots\frac{k}{N-k+1} = \frac{(N-k)!k!}{(N+1)!}$

$$p^* = \frac{f_{N+1}(k+1)}{f_N(k)} = \frac{k+1}{N+2}$$

Výsledek, na rozdíl ML, potlačuje extremality pro malé hodnoty N a k, pro velké je téměř shodný.

2.4 Rozdíly ML a Bayesovského odhadu

Odhady se liší vždy kromě limitního případu, pro N rostoucí do nekonečna, avšak pro vysoké N je rozdíl malý. Pro vícedimenzionální případy se snadněji používá ML, díky gradientu oproti multidimenzionálnímu integrálu používaného Bayesovskou metodou. Bayesovský odhad lze počítat rekurzivně s neustálým vylepšováním odhadu.
Obě metody dávají odhad parametrů konzistentní ($E(\mathbf{q}^*) \to \mathbf{q}$ pro $n \to \infty$ a $\mathrm{var}(\mathbf{q}^*) \to 0$ pro $n \to \infty$), eficientní (minimální rozptyl), odhad nemusí být nevychýlený.

3 Učení klasifikátoru

Pro nastavení klasifikátoru z trénovací množiny existují 2 základní metody:

- Odhadování parametrů pravděpodobností s následným požitím výsledku v Bayesovském přístupu (naivní Bayes. přístup)
- Přímý výběr strategie

	Odhadování parametrů	Výběr strategie
pro	jednoduchý výpočet, bayesovská teorie	Známý odhad rozdílu skutečnosti a přiblížení
proti	Není vtah mezi skutečným a odhadnutým rizikem	výpočetně náročné

Reference

[1] M. I. Schlesinger, V. Hlaváč: *Deset přednášek z teorie statistického a strukturního rozpoznávání*, ČVUT 1999

[2] Ondřej Drbohlav *Odhad metodou maximální věrohodnosti a Bayesův odhad*, http://cmp.felk.cvut.cz/cmp/courses/recognition/Lab_archive/RPZ_01-02w/MLnew/bay.html