Statistická teorie učení

Petr Havel   Marek Myslivec

přednáška z 9. týdne

1 Úvod

Představme si situaci výrobce a zákazníka, který si u výrobce objednal algoritmus rozpoznávání. Zákazník dodal experimentální materiál, např. obrázky $x_1, \dots, x_l$ a k nim příslušné ohodnocení $k_1, \dots, k_l$, které by měly být výstupem požadovaného algoritmu. Předpokládejme, že experimentální materiál je dost rozsáhlý, např. $l = 10000$. Výsledkem, který výrobce předvádí zákazníkovi, je strategie rozpoznávání. Zákazníka zajímá, jakou kvalitu rozpoznávání má vytvořená strategie. Za měřítko kvality budeme považovat pravděpodobnost chybného rozhodnutí. Jelikož se tato pravděpodobnost nedá přímo změřit, zákazník se s výrobcem dohodl, že náhradou této kvality bude počet chyb, kterých se strategie dopustí na experimentálním materiálu. Zákazník si myslí, že při dostatečně velkém experimentálním materiálu bude mít nalezená strategie velmi nízkou pravděpodobnost chybného rozhodnutí. Tento problém je však daleko složitější. Otázkou odhadu chyb strategií se budeme zabývat v následujících odstavcích.

2 Základní pojmy

Zavedeme následující označení

$x_1, \dots, x_l$ ... příznaky, $x_i \in X$
$k_1, \dots, k_l$ ... ohodnocení, $k_i \in K$
$T = ((x_1, k_1), (x_2, k_2), \dots, (x_l, k_l))$ ... trénovací množina
T^* ... množina všech trénovacích množin, $T \in T^*$

$Q : X \to K$ ... strategie (rozhodovací algoritmus)
$\overline{Q}$ ... množina všech strategií, $Q \in \overline{Q}$
$V : T^* \to \overline{Q}$ ... algoritmus učení, určuje strategii Q jako funkci $Q = V(T)$

Pro každou trénovací množinu T můžeme dostat různou strategii Q. Strategie Q je tedy náhodná veličina.

$R(Q)$ ... riziko strategie Q, tj. pravděpodobnost chybného rozhodnutí strategie Q

$$R(Q) = \int W(Q(x), k)p(x, k)dxdk$$

kde $W(Q(x), k)$ je ztrátová funkce a $p(x, k)$ jsou sdružené pravděpodobnosti, které však neznáme, a tudíž nemůžeme $R(Q)$ takto vypočítat. Dokážeme však určit

$\widehat{R}(T, Q)$ ... relativní četnost chyby strategie Q na trénovací množině T

$\widehat{R}(T, V(T))$ ... relativní četnost chyb strategie naučené na T na datech z T

Příklad (špatný algoritmus učení)
Klasifikujeme do dvou tříd, $K = \{1, 2\}$. Máme k dispozici trénovací množinu $T = ((x_1, k_1), \ldots, (x_l, k_l)),\ x_i \in R,\ k_i \in K$. Máme strategii $Q_B(x) = V(T)$ takovou, která si zapamatuje všechny vzorky v trénovací množině T a klasifikuje následovně

$$Q_B(x) = \begin{cases} k_i & \text{pro } x = x_i \\ 1 & \text{pro } x \notin T \end{cases}$$

Pak $\widehat{R}(T_l, Q_B(x)) = 0$ (na trénovací množině neudělá chybu). Výsledkem klasifikace však bude stále třída $k = 1$, protože náhodné $x \in R$ bude přesně rovno některému $x_i \in T$ s nulovou pravděpodobností. Skutečná chyba je apriorní pravděpodobnost druhého stavu, tedy $R(Q_B(x)) = P(k = 2)$. Q_B je zřejmě hodně špatná strategie, protože například pro $P(k = 2) = 0,5$ bude mít stejnou chybu jako náhodné hádání, tj. $0,5$.

Lze říci, že žádný ani libovolně velký počet trénovacích dat nemůže být zárukou, že při praktickém použití algoritmu rozpoznávání bude dosaženo stejné kvality rozpoznávání, jaká byla zjištěna na trénovacích datech.

3 Odhad chyby jedné strategie

Budeme uvažovat situaci, kdy se neučíme, máme jednu pevnou strategii Q a odhadujeme pravděpodobnost chyby $R(Q)$. Platí, že $\widehat{R}(T, Q)$ konverguje k

$R(Q)$, když délka trénovací množiny $l \to \infty$.[1] Postupně přikládáme vzorky z T, tj. opakujeme nezávislé pokusy s pravděpodobností chyby $p = R(Q)$. Počet chyb má tedy binomické rozdělení $B(l, p)$. Pro velké l konverguje $B(l, p)$ k normálnímu rozdělení $N(lp, \sqrt{lp(1-p)})$ se střední hodnotou lp a směrodatnou odchylkou menší než $\frac{1}{4\sqrt{l}}$. Lze dokázat, že pravděpodobnost toho, že odchylka $\widehat{R}(T_l, Q)$ od $R(Q)$ bude větší než ε je

$$\boxed{P\{|\widehat{R}(T_l, Q) - R(Q)| > \varepsilon\} \leq e^{-2\varepsilon^2 l} = \eta} \tag{1}$$

kde

$\widehat{R}(T_l, Q)$	...	relativní četnost chyby strategie Q na trénovací množině délky l
$R(Q)$	...	skutečná pravděpodobnost chyby
ε	...	přesnost odhadu chyby
η	...	spolehlivost tvrzení, že skutečná chyba je v intervalu $\widehat{R}(T_l, Q) \pm \varepsilon$ (jinými slovy pravděpodobnost, že skutečná chyba je mimo uvedený interval)

Příklad 1:
Známe relativní četnost chyb strategie $\widehat{R}(T_l, Q) = 0,1$. Jaká je pravděpodobnost, že pro dané l je skutečná chyba $R(Q)$ v rozmezí $0,09 \div 0,11$?

Ze zadání plyne $\varepsilon = 0,01$. Pro hledanou pravděpodobnost potom s využitím (1) platí $P \leq e^{-2.0,01^2 l} = \eta$.

Je vidět, že pro fixní délku trénovací množiny l nelze mít zároveň malé η a ε. To znamená, že při krátkém experimentu (malé l) nemůže být dosaženo takového hodnocení $R(Q)$, které by bylo zároveň přesné a spolehlivé. Lze ale naplánovat experiment s takovou délkou l, že bude dosažena jakákoliv předem určená přesnost $\varepsilon > 0$ a jakákoliv předem určená spolehlivost $\eta > 0$. Délka takového experimentu je pak podle (1)určena vzorcem

$$l \geq \frac{-\ln \eta}{2\varepsilon^2} \tag{2}$$

[1] Připadá-li vám předchozí věta v rozporu se závěrem uvedeného příkladu, pak je třeba si uvědomit, že předchozí věta neříká nic o vztahu dvou náhodných čísel $\widehat{R}(T, V(T))$ a $R(V(T))$ (v příkladu označených jako $\widehat{R}(T, Q_B)$ a $R(Q_B)$).

Příklad 2:
Jakou minimální délku trénovací množiny potřebujeme, abychom dosáhli přesnosti 2% se spolehlivostí 0,1%?

Dosazením do (2) dostáváme

$$l = \frac{-\ln 0,001}{2.0,02^2} = 8635$$

Musíme mít k dispozici trénovací množinu alespoň o délce $l = 8635$.

4 Odhad chyby při učení z více strategií

Při výběru strategie rozpoznávání je třeba brát zřetel na velikost množiny $\overline{Q}$, ze které se vybírá. Náš výběr bude špatný, pokud alespoň jedna ze špatných strategií vydrží zkoušku, tedy vztah (1) není schopen tuto špatnou strategii odhalit. Při učení proto nestačí splnit

$$P\{|\widehat{R}(T_l, Q) - R(Q)| > \varepsilon\} \leq \eta \tag{3}$$

ale je třeba splnit přísnější podmínku

$$P\{\max_{Q\in\overline{Q}} |\widehat{R}(T_l, Q) - R(Q)| > \varepsilon\} \leq \eta \tag{4}$$

Potřebujeme tedy vyřadit z učení ty algoritmy, které mají velký rozdíl $|\widehat{R}(T_l, Q) - R(Q)|$.

Základní rozdíl mezi (3) a (4) je v tom, že při $l \to \infty$ konverguje pravděpodobnost v (3) k nule při jakékoliv startegii Q a jakémkoliv $\varepsilon > 0$, kdežto konvergence (4) záleží na složitosti množiny $\overline{Q}$.

5 Případ N strategií

Pro případ, kdy se množina $\overline{Q}$ skládá z konečného počtu N strategií, musí platit

$$P\left\{\max_{Q\in\overline{Q}} |\widehat{R}(T_l, Q) - R(Q)| > \varepsilon\right\} \leq \sum_{Q\in\overline{Q}} P\left\{|\widehat{R}(T_l, Q) - R(Q)|\right\} \leq Ne^{-2\varepsilon^2 l} = \eta \tag{5}$$

neboli součet je zřejmě větší než maximum. Opět můžeme zaručit přesnost ε se spolehlivostí η. Pro délku trénovací množiny l pak plyne

$$l = \frac{\ln N - \ln \eta}{2\varepsilon^2} \quad . \tag{6}$$

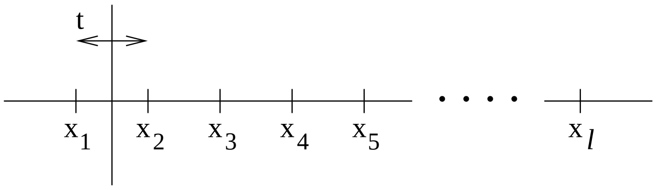
Z praktického hlediska definuje vztah (6) požadavky na délku učení příliš hrubě, tj. s velkou rezervou.

6 Charakterizace $\overline{Q}$

V předchozím odstavci nás zajímal pouze počet strategií. To je však nepoužitelné pro nekonečný počet strategií a nevhodné pro velké N. Lze ukázat, že existují popisy (charakterizace) množiny $\overline{Q}$, které hrají roli N ve vztahu (5).

Uvažujme strategie $Q_1, Q_2 \in \overline{Q}$ a vzorky v trénovací množině $X = (x_1, \dots, x_l)$. Řekneme, že Q_1 a Q_2 jsou ekvivalentní vzhledem k X, pokud $Q_1(x_i) = Q_2(x_i)$ pro všechna i. Každá posloupnost X vytváří na $\overline{Q}$ třídy ekvivalence, jejich počet označíme $\Delta\left(\overline{Q}, x_1, \dots, x_l\right)$.

Příklad :
Množina všech strategií $\overline{Q}$ je prahování v R s prahem t.



Pozorování s prahem t v oblasti vlevo od x_1 patří do jedné třídy ekvivalence (rozhoduji stále stejně), podobně oblast mezi x_1 a x_2 patří do jedné třídy ekvivalence atd. Počet tříd ekvivalence (počet strategií, které se nějak liší) je $l + 1$.

6.1 Entropie množiny strategií

Matematické očekávání $\log \Delta\left(\overline{Q}, x_1, \ldots, x_l\right)$ nazveme entropií množiny strategií $\overline{Q}$ na posloupnostech délky l

$$H_l(\overline{Q}) = \sum_{x_1 \in X} \sum_{x_2 \in X} \cdots \sum_{x_l \in X} \sum_{k_1 \in K} \sum_{k_2 \in K} \cdots \sum_{k_l \in K} \prod_{i=1}^{l} p(x_i, k_i) \log \Delta\left(\overline{Q}, x_1, \ldots, x_l\right) \tag{7}$$

což vyjadřuje průměr $\log \Delta\left(\overline{Q}, x_1, \ldots, x_l\right)$ s váhami $p(x_i, k_i)$. $H_l(\overline{Q})$ neumíme spočítat, už jen proto, že neznáme $p(x_i, k_i)$.
Naším cílem je opět ukázat, jaká má být délka učení l, aby výsledek učení byl dost přesný a spolehlivý. Lze využít následujícího tvrzení

Platí že, pokud $\frac{H_l(\overline{Q})}{l} \to 0$ pro $l \to \infty$, potom $P\left\{|\widehat{R}(T_l, Q) - R(Q)|\right\} \to 0.$

Stále však není snadné určit $\frac{H_l(\overline{Q})}{l}$, proto vytvoříme hrubější, ale konstruktivnější podmínky.

6.2 Funkce růstu

Zavedeme funkci růstu jako řadu čísel

$$m_l(\overline{Q}) = \max_{x_1, \ldots, x_l} \Delta_l\left(\overline{Q}, x_1, \ldots, x_l\right)$$

pro $l = 1, 2, \ldots, \infty$. Protože maximum je zřejmě větší než průměr, bude platit

$$\log m_l(\overline{Q}) \geq H_l(Q)$$

Aby bylo splněno

$$\lim_{l \to \infty} \frac{H_l(Q)}{l} = 0$$

jistě stačí splnit

$$\lim_{l \to \infty} \frac{\log m_l(\overline{Q})}{l} = 0 \quad ,$$

což se bude počítat snadněji a nepotřebujeme znát $p(x_i, k_i)$ jako v případě (7).

Dále platí

$$P\left\{\max_{Q \in \overline{Q}} |\widehat{R}(T_l, Q) - R(Q)| > \varepsilon\right\} \leq 3 m_{2l}(\overline{Q}) e^{\frac{-\varepsilon^2 (l-1)}{4}} = \eta \tag{8}$$

Příklad :
Pro prahování z předcházejícího příkladu platí $m_{2l}(\overline{Q}) = 2l + 1$. Po dosazení do (8) dostáváme

$$\eta = 3(2l+1)e^{\frac{-\varepsilon^2(l-1)}{4}}$$

z čehož lze opět vypočítat délku trénovací množiny l při zvolené přesnosti ε a spolehlivosti η.

Uvedený výpočet je obecně poměrně složitý. V praktických situacích se využívá charakterizace $\overline{Q}$ zvané kapacita neboli Vapnik-Červoněnkisova dimenze (VC dimenze), o které je pojednáno v následující přednášce.

Reference

[1] Michail I. Schlesinger and Václav Hlaváč. *Deset přednášek z teorie statistického a strukturního rozpoznávání.* ČVUT, Prague, Czech Republic, 1999.