Výběr modelu pomocí metody cross-validation

Cvičení z RPZ

Vojtěch Franc, Ondřej Drbohlav

10. března 2005

1 Úvod a popis problému

V tomto cvičení byste si měli osvojit používání metody cross-validation pro výběr modelu. Metodu použijeme na konkrétním problému aproximace experimentálně naměřených dat polynomem. Stupeň polynomu, který by se pro aproximaci měl použít, není předem známý. Je třeba ho určit tak, aby výsledná aproximace co nejlépe odpovídala skutečnému průběhu měřené funkce.

Problém je následující. Máme k dispozici K vzorků $(x_1, y_1), ..., (x_K, y_K)$, kde y_i je změřená hodnota neznámé reálné funkce $f(x)$ v bodě x_i.

Naměřené vzorky y_i jsou porušeny šumem, který má v našem případě Gaussovské rozdělení $N(0, \sigma)$ s nulovou střední hodnotou a variancí σ. Uvažujeme tedy model měření

$$y = f(x) + N(0, \sigma). \tag{1}$$

Naším cílem je aproximovat neznámou funkci $f(x)$ polynomem $p(x, \mathbf{a}, d)$

$$p(x, \mathbf{a}, d) = a_0 + a_1 \cdot x + ... + a_d \cdot x^d \,, \tag{2}$$

kde d je stupeň polynomu a vektor $\mathbf{a} = [a_0, a_1, ..., a_d]$ je vektor koeficientů tohoto polynomu. Chceme zvolit takový stupeň polynomu d a takový vektor $\mathbf{a}^*$, aby tato aproximace byla co nejlepší.

Jak to však provést? Především je třeba si rozmyslet, co znamená "nejlepší aproximace funkce". Tradiční volbou je hodnocení chyby aproximace jako součtu odchylek aproximace od naměřených dat umocněných na druhou a hledání takových parametrů aproximačního modelu, které tento součet minimalizují (= metoda nejmenších čtverců). Ukažme si takový přístup na příkladě, který se nám v dalším bude hodit. Nechť zvolený řád polynomu je d a máme dánu *podmnožinu* naměřených vzorků. Ta nechť je zadána množinou T, ve které jsou indexy vybraných vzorků, tj. $T \subset \{1, 2, \ldots, K\}$. Vektor $\mathbf{a}^*$ se ve smyslu nejmenších čtverců

(Least Squares, LS) spočítá jako

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} \sum_{i \in T} (y_i - p(x_i, \mathbf{a}, d))^2 \,. \tag{3}$$

Vyřešení této úlohy je popsáno v oddíle 2. Teď se však zamysleme, jestli bychom pomocí metody nejmenších čtverců dokázali najít jak $\mathbf{a}^*$, tak i stupeň polynomu d. Odpověď zní že ano, ale řešení by se nám nelíbilo. Pro vysoké stupně polynomu d totiž dochází k tzv. "přefitování". To znamená, že s rostoucím řádem polynomu chyba na trénovacích datech klesá (nebo je dokonce nulová), ale polynom s vysokým stupněm nemusí vůbec dobře odpovídat skutečné funkci $f(x)$.

Nyní se dostáváme k principu metody cross-validation. V tomto případě se určí stupeň polynomu d tak, aby byla nejmenší tzv. cross-validační chyba E_{CROSS}. Cross-validační chyba se spočítá následujícím způsobem. Náhodně rozdělíme naměřenou množinu dat $(x_1, y_1), ..., (x_K, y_K)$ na n skupin. Tyto skupiny nechť jsou opět zadány pomocí množin indexů vzorků, které označme $S_1, S_2, ..., S_n$. Tyto množiny nechť se nepřekrývají, jsou stejně velké a jejich sjednocení dá celou množinu, takže $\bigcup_{i=1}^{n} S_i = \{1, 2, \ldots, K\}$. Těmto množinám budeme říkat "testovací".

Výpočet cross-validační chyby popisuje následující pseudokód:

- for $i = 1 : n$,
  - Vytvoř "trénovací" množinu vzorků T_i jako doplněk testovací množiny S_i. Tj. v T_i jsou indexy všech vzorků, které nejsou v S_i.
  - Odhadni vektor $\mathbf{a}^*$ na trénovací množině T_i na základě nejmenších čtverců (3).
  - Spočti chybu na E_i na testovací množině S_i jako součet kvadrátů odchylek aproximace od vzorků.
- end.
- Spočti cross-validační chybu E_{CROSS} jako průměr z testovacích chyb $E_{CROSS} = \frac{1}{n} \sum_{i=1}^{n} E_i$.

Podle počtu podmnožin n mluvíme o tzv. n-fold cross-validační chybě. Oblíbená je volba $n = 10$. Pozn: limitní případ, kdy n je rovno počtu naměřených dat se nazýva leave-one-out.

2 Metoda nejmenších čtverců

Naším cílem je určit parametry polynomu tak, aby střední kvadratická chyba na trénovací množině byla minimální. Tento problém lze jednoduše zapsat v mati-

covém vyjadření. Zaveďme nésledující značení

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^d \\ 1 & x_2 & x_2^2 & \dots & x_2^d \\ & & \vdots & & \\ 1 & x_K & x_K^2 & \dots & x_K^d \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{bmatrix} \qquad \mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{bmatrix}. \tag{4}$$

Problém odhadu vektoru $\mathbf{a}^*$ pomocí nejmenších čtverců lze maticově zapsat jako

$$\mathbf{a}^* = \underset{\mathbf{a}}{\operatorname{argmin}} (\mathbf{X} \cdot \mathbf{a} - \mathbf{y})^T \cdot (\mathbf{X} \cdot \mathbf{a} - \mathbf{y}) \tag{5}$$

Tento problém má analytické řešení

$$\mathbf{a}^* = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y} \tag{6}$$

3 Zadání cvičení

1. Implementujte metodu nejmenších čtverců pro odhad parametrů polynomu podle popisu v sekci 2.

2. Implementujte metodu cross-validation pro nalezení stupně polynomu s nejmenší cross-validační chybou. Stupně polynomu volte v rozmezí 1 až 10. Trénovací množinu uloženou v MAT-souboru si stáhněte zde. Pro rozdělení dat na trénovací a testovací množinu můžete použít funkci crossval.m.

3. Zobrazte závislost cross-validační chyby a trénovací chyby na stupni polynomu.