Výběr parametrů pomocí křížové validace

Návod ke třináctému cvičení

ZS 2011/2012

Na poslední přednášce se seznámíte s několika metodami odhadu chyb klasifikátoru. Jednou z těchto metod je tzv. křížová validace, která je užitečná především tehdy, nemáme-li dostatek dat na to, abychom je mohli rozdělit na dostatečně velkou trénovací a testovací množinu. V tomto cvičení zkusíte použít metodu křížové validace pro výběr optimálních parametrů klasifikačního modelu.

Nastudujte si...

1. Z přednáškových slajdů si nastudujte metodu stratifikované k-fold křížové validace. **Stručně:** Máme trénovací množinu $\mathcal{T}$ obsahující příklady, které v našem případě patří do dvou tříd. Rozdělíme $\mathcal{T}$ na k pokud možno stejně velkých disjunktních podmnožin $\mathcal{T}_i$. Všechny tyto podmnožiny by měly obsahovat *přibližně* stejné poměry pozitivních a negativních příkladů (tj. máme-li v trénovací množině 90% pozitivních a 10% negativních příkladů, pak by i v každé $\mathcal{T}_i$ mělo být přibližně 90% pozitivních a 10% negativních příkladů). Poté opakujeme k-krát následující postup (pro $i = 1, \ldots, k$). Vezmeme i-tou podmnožinu trénovacích příkladů (tj. $\mathcal{T}_i$) a označíme ji jako $Test_i$ a sjednocení všech zbývajících podmnožin $Train_i = \bigcup_{j=1,\ldots,k,j\neq i} \mathcal{T}_j$. Naučíme klasifikátor na množině $Train_i$ a použijeme jej pro klasifikaci příkladů z množiny $Test_i$. Porovnáním se skutečným zařazením příkladů z množiny $Test_i$ do tříd dostaneme chybu e_i. Chyba odhadnutá pomocí metody křížové validace je průměrem těchto chyb e_i.

Data a použité klasifikátory...

2. Trénovací data budeme generovat ze směsi k jednorozměrných normálních rozdělení se stejným rozptylem (pomocí funkce `generate_examples`). Pozitivní příklady jsou generovány ze směsi, jejíž složky mají střední hodnoty 2, 4, ..., $2q$. Podobně negativní příklady jsou generovány ze směsi, jejíž složky mají střední hodnoty 1, 3, ..., $2q - 1$ (kde q je volitelný parametr).

3. Pro klasifikování příkladů použijeme bayesovské rozhodování. K tomu potřebujeme znát pravděpodobnostní model pro pozitivní a negativní příklady. Tyto

modely budou směsi normálních rozdělení, jejichž parametry budeme odhadovat pomocí EM algoritmu. Jak víte z první části předmětu, EM algoritmus potřebuje znát počet složek směsi. Tento parametr se budeme snažit optimálně nastavit s použitím křížové validace (tj. vybereme takový počet složek směsi, který bude minimalizovat odhad chyby daný křížovou validací).

Implementujte...

4. **Prostudujte zdrojové kódy ke třináctému cvičení. Doimplementujte funkci `cross_validation`.**

Otestujte...

5. Otestujte na vygenerovaných datech.