Nedoučené a přeučené klasifikátory

Návod k devátému cvičení

ZS 2011/2012

Cílem tohoto cvičení je ilustrovat teoretické pojmy z přednášky na jednoduchých příkladech. Konkrétně se budeme zabývat problémem přeučení a nedoučení klasifikátorů, souvislostí tohoto problému s křivkou učení.

Zadání

V tomto cvičení budeme pracovat s k-konjunkcemi (k-konjunkce jsou konjunkce obsahující nejvýše k literálů). Na přednášce jste se seznámili s algoritmem pro učení konjunkcí. Tento algoritmus zde nicméně není použitelný. Jednak totiž předpokládá příklady bez šumu a zároveň není tento algoritmus použitelný pro hledání k-konjunkcí, ale pouze konjunkcí s neomezenou délkou (omezenou pouze počtem výrokových proměnných v datasetu). Z toho důvodu používáme v tomto cvičení algoritmus založený na metodě větví a mezí (branch and bound), který hledá takovou konjunkci délky nejvýše k, která minimalizuje chybu na trénovacích datech.

V obou úkolech si budete pouze hrát s již připraveným kódem a budete se snažit interpretovat získané výsledky.

Nedoučení a přeučení

Matlabovský skript `cviceni8.m` se skládá ze tří hlavních částí: generování příkladů, konstrukce křivky *bias-variance-trade-off* a konstrukce křivek učení. Prostudujte kód pro konstrukci těchto křivek. Všimněte si, že ve funkci `conj_bb` je proměnná `prefer_long`, která určuje, zda mají být preferovány delší konjunkce. Pokud je tato proměnná nastavena na 1, vybere algoritmus ze všech konjunkcí s minimální chybou na trénovacích datech tu, která má nejvíce literálů. **Zkonstruujte křivku *bias-variance-trade-off* s daty obsahující šum i bez něj a pro verzi algoritmu preferující dlouhé konjunkce a pro verzi s `prefer_long = 0`. Diskutujte rozdíly. Kde se více projeví přeučení a proč?**

Křivky učení

Třetí část skriptu `cviceni8.m` je určena pro konstrukci křivek učení. Prostudujte tuto část. **Vysvětlete, jak to, co zobrazují křivky učení, souvisí s křivkou *bias-variance-trade-off*. Otestujte vliv nastavení šumu trénovacích dat na průběhy křivek učení.**