Teoretická domáca úloha č. 1

2-AIN-150, Zima 2018

Termín: 5.11.2018, 23:59, cez moodle.uniba.sk/fmfi

Skôr ako sa pustíte do riešenia domácej úlohy, oboznámte sa so všeobecnými pokynmi, ktoré sú priložené na konci tohto dokumentu. Riešenia, ktoré odovzdáte, musia byť vaše vlastné. Neopisujte a nesnažte sa nájsť riešenia v literatúre alebo na internete!

Teória strojového učenia

Uvažujme problém regresie nad množinou hypotéz $H = \{h_b(x) = 2x + b\}$.

a) Popíšte algoritmus, ktoré pre dané trénovacie dáta $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$, vyberie hypotézu, ktorá minimalizuje chybu danú funkciou: $E(b) = \sum_{i=1}^{n} (h_b(x^{(i)}) - y^{(i)})^2$.

b) Uvažujme, že dáta sú generované distribúciou $P_{x,y}$ definovanou nasledovne:

- rozdelenie x-ov je rovnomerné na intervale $[0, 100]$.
- pre dané x je $\Pr(y = 2x + 7|x) = 0.4$ a $\Pr(y = 2x - 5|x) = 0.6$ (iné hodnoty y sa v kombinácii s x nevyskytujú).

Aká je optimálna testovacia chyba pre množinu hypotéz H (t.j. testovacia chyba pre najlepšiu hypotézu z H) ak predpokladáme, že dáta sú nezávislé vzorky z distribúcie $P_{x,y}$?

c) Pre distribúciu $P_{x,y}$ z časti b) a $n = 1$ spočítajte očakávanú trénovaciu a testovaciu chybu (t.j. vygerujeme jeden príklad, natrénujeme a z tohoto procesu rátame očakávané chyby).

d) Pre distribúciu $P_{x,y}$ z časti b) a všeobecné n spočítajte očakávanú trénovaciu a testovaciu chybu. V akom vzťahu sú tieto chyby ku optimálne testovacej chybe keď $n \to \infty$? Vedeli by ste na základe vašich výsledkov zvoliť vhodnú veľkosť trénovacej množiny?

Všeobecné pokyny

Riešenie odovzdajte vo formáte PDF cez moodle.