Domáca úloha č. 3

2-AIN-150, Zima 2018

Termín: 12.11.2018, 23:59, moodle.uniba.sk/fmph

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!

Výber podstatných atribútov

Poriadne čítajte, čo treba odovzdať.

Uvažujme lineárnu regresiu, kde máme m atribútov, n trénovacích príkladov, ale výstup reálne závisí iba od niektorých atribútov.

Vašou úlohou bude pre zadané dáta (t.j. sadu vstupov $\vec{x}^{(i)}$ a očakávaných výstupov $y^{(i)}$) zistiť, na ktorých atribútoch y naozaj závisí (resp. vyplýva to z pozorovaných dát).

Nemusíte naprogramovať všeobecnú metódu. Stačí, že použitím metód zo strojového účenia poloručne-poloautomaticky zistíte výsledok. Do moodlu potom odovzdajte nasledovné veci:

- zistený výsledok (t.j. zoznam podstatných atribútov)
- stručný komentár vášho postupu a zdrojové kódy, ktoré vám pomohli k získaniu výsledku

V balíku k úlohe sú dva vstupné súbory. Zo súboru `input.txt` zistite požadovaný výsledok. Súbor `sample.txt` je pre vás na overenie kvality vašich postupov. Na ňom by ste mali zistiť, že podstatných je prvých 10 atribútov. Ale pokojne môžete tento súbor ignorovať.

Vo svojich programoch môžete používať knižnice, ktoré robia základnú matiku, maticové operácie, rátajú inverzné matice, sústavy lineárnych rovníc, numericky/symbolicky derivujú. Takisto môžete používať knižnice, ktoré počítajú lineárnu regresiu (a jej regularizované varianty), robia cross validáciu, t.j. napr. príslušné časti scikit-learn.

Nepoužívajte knižnicné funkcie, ktoré explicitne robia výber atribútov (urobia za vás domácu na jeden riadok).

Hinty a poznámky

- L1 regularizácia (spomínaná na prednáške) tlačí váhy do nuly pre atribúty, ktoré sa jej zdajú byť nepodstatné. Treba jej ale dobre nastaviť parameter α.
- Cross validácia vie byť tiež dobrý pomocník (môžete napríklad porovnávať úspešnosť pre rôzne vybrané množiny atribútov).

Pokyny pre Python V balíku je súbor `template.py`, v ktorom môžete doprogramovať funkciu `select_features(X, y)`. Máte v ňom naprogramované načítanie dát a ukážky ako pustiť $L1$ regresiu a cross validáciu. Program sa spúšťa príkazom `python template.py <vstupný súbor>`. Na spustenie programu potrebujete mať nainštalovaný scikit-learn.

Pokyny pre iné jazyky Napíšte podobnú funkciu ako v Pythone a vhodne ju okomentujte a otestujte. Vstupný súbor obsahuje na prvom riadku čísla n, m. Nasleduje n riadkov s trénovacími príkladmi. Každý trénovací príklad je na jednom riadku a má nasledovný formát: Obsahuje najprv m čísel – atribúty vstupu a potom jedno číslo – očakávaný výstup.