Domáca úloha č. 1

2-AIN-150, Zima 2020

Termín: 22.10.2020, 23:59, boza@fmph.uniba.sk, predmet mailu: Strojove ucenie DU 1

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!

Model pre dáta s málo nenulovými atribútmi

Uvažujme lineárnu regresiu, kde máme m atribútov, n trénovacích príkladov, ale väčšinu matice X tvoria nuly a len niekoľko prvkov v každom riadku je nenulových.

V takejto situácii je obvykle vhodné siahnúť po (stochastic) gradient descente (matica $X^T X$ by bola príliš veľká). Vašou úlohou je vymyslieť a naprogramovať ako spočítať gradient rýchlejšie ako v čase $O(mn)$ (resp. ako v čase $O(m)$ pre jeden príklad). Následne tento výpočítaný gradient použite na natrénovanie lineárnej regresie.

Inými slovami: Máte zadané n, m, X, $\vec{y}$, kde X obsahuje skoro samé nuly, natrénujte lineárnu regresiu na riedkych dátach dostatočne efektívne.

V balíku k úlohe je jeden vstup. Mali by ste na ňom dosiahnúť strednú kvadratickú trénovaciu chybu približne 1.05 (pokiaľ máte chybu okolo 2330, tak vám chýba jedna dôležitá vec).

Vo svojich programoch môžete používať knižnice, ktoré robia základnú matiku, maticové operácie, rátajú inverzné matice, sústavy lineárnych rovníc, numericky/symbolicky derivujú. Explicitne máte zakázané použiť knižnicu scikit-learn a jej obdoby v iných jazykoch (čokolvek, čo urobí domácu za vás na jeden riadok :).

Pokyny pre Python V balíku je súbor `template.py`, v ktorom doprogramujte funkciu `fit(n, m, X, y)`. Program sa spušťa príkazom `python template.py <vstupný súbor>`.

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: Najprv obsahuje číslo k – počet nenulových prvkov vo vstupným dátach. Nasleduje k dvojíc tvaru p, v, kde p je pozícia nenulového prvku a v je jeho hodnota. Potom nasleduje číslo y – požadovaný výsledok.

Všeobecné pokyny

Úlohy odovzdávajte mailom na mail s predmetom uvedeným v nadpise. Svoj kód vložte do prílohy mailu.

Ideálne odovzdávajte domáce úlohy v Pythone (doprogramujte požadované funkcionality zo zadania). Pokiaľ chcete použiť iný jazyk, môžete, ale musíte zároveň naprogramovať aj réžiu okolo (načítanie, výpis, ...). Bolo by ale vhodné, aby som váš program vedel rozbehať pod linuxom bez nutnosti inštalácie komerčných programov (t.j. overte si, či váš Matlabový kód ide spustiť v Octave, C# či funguje pod Monom, ...).