Domáca úloha č. 1

2-INF-150: Strojové učenie, Zima 2012

Termín: 15.10.2012, M163 (pod dvere)

1. Maticový počet. Uvažujme problém lineárnej regresie, pričom optimalizujeme normu L_2. Nech X je dizajnová matica, y je cieľový vektor trénovacej množiny a θ je vektor parametrov lineárnej funkcie, ktorú v procese trénovania hľadáme. Na prednáške sme odvodili, že pre vektor θ musí platiť:

$$X^T X\theta = X^T y$$

Ak matica $(X^T X)$ je regulárna, môžeme túto rovnicu prenásobiť zľava inverznou maticou $(X^T X)^{-1}$ a ďalším odvodením dostaneme:

$$\theta = (X^T X)^{-1} X^T y = X^{-1} X^{T^{-1}} X^T y = X^{-1} y$$

$$\theta = X^{-1} y$$

Kde je v tomto odvodení problém?

2. Pevnosť betónu v tlaku. V tejto úlohe sa budeme zaoberať predpovedaním *pevnosti betónu v tlaku* (compressive strength). Pevnosť v tlaku je schopnosť materiálu vydržať príslušnú tlakovú silu; pri aplikácií vyšších tlakových síl sa začne drobiť. Dáta s 1030 príkladmi, 9 atribútmi a 1 cieľovou hodnotou nájdete v UCI machine learning repository na adrese `http://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength`.

Ku každej časti odovzdajte krátky popis a zdrojový kód programov alebo funkcií, ktoré ste napísali. Podľa vášho popisu by malo byť jednoduché zrekonštruovať výsledky vašej práce.

Programy môžete písať v ľubovoľnom programovacom jazyku (ale obzvlášť ľahké je to v jazyku Octave alebo Matlab). Ak váš jazyk priamo nemá reprezentáciu matíc a vektorov, zvoľte si reprezentáciu, s ktorou sa vám dobre pracuje.

a) Z množiny dát náhodne vyberte 800 trénovacích príkladov (ďalej len trénovacia množina). Zvyšných 230 príkladov bude slúžiť na testovanie (ďalej len testovacia množina).

b) Napíšte funkciu *linregd*, ktorá na vstupe dostane trénovaciu dizajnovú maticu X, cieľový vektor y a celé číslo d. Funkcia najskôr rozvinie množinu atribútov pomocou všetkých bázových funkcií stupňa nanajvýš d, potom vypočíta koeficienty ich lineárnej kombinácie tak, aby bol minimalizovaný súčet štvorcov chýb a vráti vektor týchto koeficientov na výstupe.

   Taktiež napíšte zodpovedajúcu funkciu h, ktorá na vstupe dostane maticu príkladov X, celé číslo d a vektor koeficientov vypočítaných pomocou vašej funkcie *linregd* a ktorá vráti vektor, ktorý pre každý riadok matice X obsahuje predpovedanú hodnotu pomocou regresnej funkcie vypočítanej funkciou *linregd*.

   Napríklad pre $d = 2$ a 3 atribúty (x_1, x_2, x_3) použijete bázové funkcie $1, x_1, x_2, x_3, x_1^2, x_1x_2, x_1x_3, x_2^2, x_2x_3, x_3^2$.

c) Nech $h_{d,t}$ je regresná funkcia natrénovaná vašou funkciou *linregd* s parametrom d pre prvých t trénovacích príkladov z časti a). Pre všetky kombinácie $d \in \{1, 2, 3\}$ a $t \in \{100, 200, 300, 400, 500, 600, 700, 800\}$ natrénujte funkciu $h_{d,t}$ a spočítajte tzv. *trénovaciu chybu na jeden príklad*, t.j. použite $h_{d,t}$ na tú istú trénovaciu množinu t príkladov a spočítajte sumu štvorcov chýb predelenú počtom príkladov t. Prezentujte výsledky formou tabuľky a prehľadného grafu.

d) Každú funkciu $h_{d,t}$ natrénovanú v časti c) použite aj na predpoveď hodnôt pre príklady z testovacej množiny z časti a) a spočítajte tzv. *testovaciu chybu na jeden príklad*, t.j. sumu štvorcov chýb predelenú počtom príkladov v testovacej množine. Prezentujte výsledky formou tabuľky a prehľadného grafu.

e) Vysvetlite priebehy a rozdiely grafov z častí c) a d)

3. (Bonus) L_1 norma. V jazyku Octave napíšte funkciu $L1(X, y)$, ktorá pre dizajnovú maticu X a cieľový vektor y použitím lineárneho programovania spočíta koeficienty lineárnej regresie θ, ktoré minimalizujú chybovú normu L_1 (namiesto obvyklej L_2). Za pomoci simulovaných dát demonštrujte, že vaša funkcia funguje a porovnajte výsledky takéhoto trénovania s klasickým trénovaním (norma L_2) pomocou normálnych rovníc.

Poznámka: Bonusový príklad bude hodnotený oveľa prísnejšie ako zvyšok domácej úlohy, čiastočné riešenia nezískajú žiadne body. Ak ste na pochybách, radšej investujte čas do príkladov 1 a 2.