Domáca úloha č. 3

2-INF-150: Strojové učenie, Zima 2009

Termín: 14.12.2009, M163 (pod dvere)

1. Vapnik-Červonenkisova dimenzia.

a) Uvažujme dva reálne atribúty x_1 a x_2 a množinu hypotéz, ktorá pozostáva zo všetkých kruhov v rovine (všetky body vo vnútri kruhov sa klasifikujú ako pozitívne a všetky body mimo ako negatívne). Koľko má každá takáto hypotéza parametrov? Aká je VC dimenzia takejto množiny hypotéz? Dokážte.

b) (**bonusový príklad**) Uvažujme jeden reálny atribút x a množinu hypotéz s jedným parametrom α, $H = \{\text{sgn}(\sin \alpha x)\}$. Dokážte, že VC dimenzia takejto množiny hypotéz je ∞.

2. Priemerné ceny domov. V tomto príklade odporúčame použiť octave. V prípade použitia jazyka octave môžete použiť funkcie eig, sort, mean, std a bežné funkcie na prácu s maticami (nepoužívajte zložitejšie funkcie). V prípade iného programovacieho jazyka môžete použiť knižnice na hľadanie vlastných hodnôt, ale nie knižnice ktoré priamo vedia počítať PCA.

a) Napíšte funkciu `normalize`, ktorá znormalizuje vstupné dáta tak, aby priemer každého atribútu bol 0 a štandardný odchýlka každého atribútu bola 1.

b) Napíšte funkciu `pca(x,k)`, ktorá pre danú vstupnú normalizovanú maticu x nájde a vypíše prvých k hlavných komponentov. Postupujte podľa algoritmu z prednášky, váš program by mal mať nasledovné identifikovateľné časti: vytvorenie kovariačnej matice, vypočítanie vlastných čísel a vektorov, utriedenie podľa vlastných čísel, vrátenie prvých k hlavných vektorov.

c) Aplikujte vašu implementáciu na dáta o priemerných cenách domov v Bostone, ktoré sú k dispozícii v UCI repository `http://archive.ics.uci.edu/ml/datasets/Housing`

Z týchto dát použite pre analýzu hlavných komponentov prvých 13 atribútov, cieľový atribút 14 (priemerná cena domov) z analýzy vynechajte.

Zobrazte formou dvojrozmerného grafu prvý vs. druhý hlavný komponent, pričom v grafe vyznačíte (napr. značkou "x") všetky príklady, kde atribút 14 je menší ako 15 (tzv. lacné domy) a iným spôsobom (napr. značkou "o") vyznačíte príklady, kde atribút 14 je väčší ako 30 (tzv. drahé domy).

Aké závery by ste z grafu vyvodili?