# Domáca úloha č. 3

2-INF-150: Strojové učenie, Zima 2017

Termín: 03.01.2018, M163 (pod dvere)

**1. Vapnik-Červonenkisova dimenzia.**

a) Uvažujme množinu hypotéz, ktorá pozostáva zo všetkých kruhov v rovine (všetky body vo vnútri kruhov sa klasifikujú ako pozitívne a všetky body mimo ako negatívne). Aká je VC dimenzia takejto množiny hypotéz? Dokážte.

b) Na prednáške sme ukázali, že obdĺžniky s osovorovnobežnými stranami (všetky body vo vnútri obdĺžnika sa klasifikujú ako pozitívne a všetky body mimo ako negatívne) v priestore $\mathbb{R}^2$ majú VC dimenziu 4. Aká je VC dimenzia osovorovnobežných obdĺžnikov v priestore $\mathbb{R}^d$? Dokážte.

**2. Priemerné ceny domov.** V tejto úlohe naimplementujete a použijete jednoduchú funkciu pre PCA v jazyku python. Môžete vychádzať z cvičení a z prednášok. Vo vašej implementácii môžete použiť základné knižničné funkcie napríklad pre výpočet priemerov, štandardných odchýliek, vlastných hodnôt a pod. Ak ste niektoré časti kódu naimplementovali v rámci cvičení, môžete ich použiť. **Nesmiete však použiť knižničné funkcie, ktoré vedia priamo počítať PCA.**

a) Napíšte funkciu `mojaNormalizacia(x)`, ktorá znormalizuje vstupné dáta v matici $x$ tak, aby priemer každého atribútu bol 0 a štandardná odchýlka každého atribútu bola 1. (Riadky matice reprezentujú jednotlivé príklady a stĺpce reprezentujú jednotlivé atribúty.)

b) Napíšte funkciu `mojePCA(x,k)`, ktorá pre danú vstupnú normalizovanú maticu $x$ nájde a vráti prvých $k$ hlavných komponentov. Postupujte podľa algoritmu z prednášky, váš program by mal mať nasledovné jasne 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 archíve StatLib `http://lib.stat.cmu.edu/datasets/boston`

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?

Všetky časti odovzdávate písomne. Ku častiam a) a b) odovzdajte okomentovaný kód implementujúci dané funkcie. Ku časti c) odovzdajte výsledný graf a krátke závery. Ďalšie obslužné časti vášho kódu nie je potrebné odovzdávať.