

Domáca úloha č. 3

2-INF-150: Strojové učenie, Zima 2014

Termín: 19.12.2014, 9:00, M163 (pod dvere)

1. Vapnik-Červonenkisova dimenzia.

- a) Uvažujme dva reálne atribúty x_1 a x_2 a množinu hypotéz, ktorá pozostáva zo všetkých obdĺžnikov v rovine, pričom strany obdĺžnika **nemusia byť rovnobežné so súradnicovými osami** (všetky body vo vnútri obdĺžnika sa klasifikujú ako pozitívne a všetky body mimo ako negatívne). Aká je VC dimenzia takejto množiny hypotéz? Dokážte.
- b) (**bonusový príklad**) Uvažujme jeden reálny atribút x a množinu hypotéz s jedným parametrom α , $H = \{\text{sgn}(\cos \alpha x)\}$. Dokážte, že VC dimenzia takejto množiny hypotéz je väčšia ako ľubovoľné konečné k .

2. Segmentácia textu. Na vstupe máte daný dvojjazyčný text, pričom v hociktorom momente sa text môže prepnúť z jedného jazyka do druhého alebo späť. Vašou úlohou je napísať program, ktorý text rozsegmentuje, t.j. označí, ktoré časti textu sú v jednom jazyku a ktoré časti sú v druhom jazyku. Trénovacie a testovacie dáta pre túto úlohu nájdete na stránke

<http://compbio.fmph.uniba.sk/vyuka/ml/a3-data.zip>

Prvý riadok každého súboru obsahuje dvojjazyčný text a druhý riadok obsahuje segmentáciu textu podľa jazyka.

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

- a) Navrhňte stavový diagram pre HMM, ktorý možno použiť pre segmentáciu textu pozostávajúceho z dvoch jazykov a navrhňte spôsob natrénovania jeho parametrov z označovaných dát.
- b) Naimplementujte váš HMM, jeho tréovanie a Viterbiho algoritmus. Natrénujte štyri verzie vášho algoritmu na štyroch trénovacích súboroch (*-train.txt).
- c) Otestujte každú verziu vášho HMM na príslušných testovacích dátach (*-test.txt), pre každú verziu vyhodnoťte úspešnosť. Vysvetlite, prečo sa úspešnosť medzi jednotlivými verziami líši (budete sa pravdepodobne musieť pozrieť na dáta a na to, ako sa natrénovali jednotlivé parametre modelu).
- d) (**bonusový príklad**) V predchádzajúcej časti ste pravdepodobne postrehli nízku úspešnosť segmentácie česko-slovenského textu. Pokúste sa navrhnúť lepší HMM pre túto úlohu. Hint: Niektoré dvojice písmen sa veľmi často vyskytujú v jednom z jazykov, ale sú pomerne vzácne v druhom jazyku.