Učení z textů

Základy umělé inteligence – úloha 3

LS 2009

1 Odevzdání a hodnocení

1. Úlohu vypracovává každý student samostatně.
2. Odevzdání má dvě části:
   (a) demonstrace experimentů, předvedení výsledků
      - na cvičení 10.5., resp. 13.5. dle rozvrhu,

   (b) odevzdání zprávy o řešení
      - systém https://cmp.felk.cvut.cz/ulohy,
      - termín odevzdání do 10.5., resp. 13.5.
3. Za úlohu lze získat 15 bodů
   (a) 3 body za demonstraci experimentů
   (viz bod 2a, lze získat pouze na uvedeném cvičení),

   (b) 12 bodů za zprávu o řešení
   (viz bod 2b),

   (c) za každý započatý týden zpoždění odevzdání do systému na https://cmp.felk.cvut.cz/ulohy ztrácíte 3 body.

2 Zadání

Učení z textových dat, obecné zadání:
Vytvořte a testujte klasifikátory, které na základě textu diskusního příspěvku přidělí tento příspěvek do jedné z dvaceti různých diskusních skupin. Cílem je vytvořit srozumitelný klasifikátor s maximální klasifikační přesností. Důležitou součástí řešení je vhodné předzpracování vstupních textů.

Postup řešení:

1. Seznamte se s 20_Newsgroups daty na stránce:
   http://people.csail.mit.edu/jrennie/20Newsgroups/.

2. Seznamte se s nástrojem strojového učení WEKA:
   viz http://www.cs.waikato.ac.nz/ml/weka/ a WEKA_guide.pdf.

3. Předzpracujte vstupní data do ARFF formátu. Zvolte si úroveň obtížnosti řešení:

   (a) **Snadná.** Použijte předpřipravený soubor news_1000.arff. V nástroji WEKA experimentujte s různými klasifikátory, popřípadě výběrem relevantních slov. Počet sledovaných slov je omezen na 1000, výběr je dán pouze pořadím slov. Řešení bude mít rezervy v přesnosti. Nelze získat plný počet bodů za úlohu, maximem je 11 bodů (2+9).

   (b) **Střední.** Vyjděte ze zahuštěné reprezentace dat pro Matlab ve formátu (docIdx, wordIdx, count) dostupné na: http://people.csail.mit.edu/jrennie/20Newsgroups/20news-bydate-matlab.tgz. Tuto reprezentaci převeďte do arff WEKA formátu, zvolte vhodný rozsah a výběr množiny slov. Můžete využít python konverter matlab_arff_sparse_w.py.

   (c) **Obtížná.** Vyjděte z původních textů dostupných na: http://people.csail.mit.edu/jrennie/20Newsgroups/20news-bydate.tar.gz. Navrhněte vlastní způsob předzpracování těchto dat. Použijte stemmery implementované přímo v prostředí WEKA (weka.filters.unsupervised.attribute.StringToWordVector), popřípadě uvažujte zachycení sekvenční povahy textu (nezáleží pouze na četnosti výskytu slov, ale také jejich interakci). Případně použijte specializované nástroje typu Bow (http://www.cs.cmu.edu/~mccallum/bow/).

4. Pokud je počet atributů nebo počet příkladů příliš velký, použijte filtry.

   (a) Pro předvýběr relevantních slov lze použít například míru weka.attributeSelection.GainRatioAttributeEval.

(b) Jaká slova se jeví jako nejdůležitější? Odpovídá to očekáváním?

(c) Instance lze filtrovat například filtrem weka.filters.unsupervised.instance.RemovePercentage (náhodné filtrování) nebo weka.filters.unsupervised.instance.RemoveWithValues (omezení na podmnožinu kategorií).

5. Na základě dat sestavte symbolický klasifikátor (rozhodovací strom – weka.classifiers.trees.J48, množina rozhodovacích pravidel – weka.classifiers.rules.DecisionTable, weka.classifiers.rules.JRip).

   (a) vytvořený model interpretujte (popište slovně), je srozumitelný?, dává smysl?, které diskusní skupiny se daří nebo naopak nedaří oddělit (vysvětlete matici záměn, popřípadě použijte F-míru)?

   (b) ověřte přesnost modelu křížovou validací,

   (c) experimentujte s parametry klasifikátoru nebo jeho variantami,

   (d) sestavte křivku učení (zkušenost může být parametrizována rostoucícm počtem trénovacích příkladů nebo relevantních slov), lze manuálně opakovaným použitím filtru weka.filters.unsupervised.instance.RemovePercentage, návod na automatické vytvoření křivky učení je na http://weka.wikispaces.com/Learning+curves.

6. Na základě dat sestavte neuronovou síť (vícevrstvý perceptron – weka.classifiers.functions.MultilayerPerceptron).

   (a) ověřte přesnost modelu křížovou validací,

   (b) experimentujte s parametry klasifikátoru nebo jeho variantami,

   (c) srovnejte vlastnosti modelu s modelem symbolickým vytvořeným v předchozím kroku,

   (d) sestavte křivku učení (zkušenost může být parametrizována rostoucícm počtem trénovacích příkladů nebo relevantních slov).

7. O řešení vypracujte písemnou zprávu

   (a) zpráva bude obsahovat řešení pro body 3-6,

   (b) do zprávy zařaďte i jakákoli zajímavá pozorování nad rámec bodů 3-6, napište krátké shrnutí,

   (c) odevzdání na https://cmp.felk.cvut.cz/ulohy,

   (d) název odevzdaného ZIP archivu = vaše uživatelské jméno.