Lab 7: Data Mining.

Metode de tip ansamblu
Analiza documentelor text (Text mining)

---

1. Metode de tip ansamblu

Sunt meta-modele care se obţin din câteva modele de bază antrenate pe acelaşi set sau pe seturi diferite de antrenare. Există mai multe variante de a construi modele de tip ansamblu:

- Utilizând modele bazate pe algoritmi diferiţi antrenati pe acelaşi set de date (e.g. *bucket of models*)
- Utilizând modele bazate pe acelaşi tip de algoritm dar antrenate pe seturi diferite de date (e.g. *bagging* and *boosting*)
- Utilizând diferite modele şi împărţind setul de date (e.g. *stacking*)

Motivația principală este de a reduce eroarea la predicție acționând asupra celor două componente: deplasarea (bias) și varianța (variance).

Exercițiul 1/ Rattle. Aplicați clasificatorul de tip Random Trees din Rattle pentru a clasifica datele din seturile folosite la Lab 3: (a) iris; (b) breast cancer (Wisconsin); (c) Titanic.

Exercițiul 2/R. Testați pachetul SuperLearner. Exemplu: ExampleSuperLearner.r

Exercitiul 3/Weka. Utilizand Weka Experimenter comparaţi performanţa următoarelor metamodele: Vote, Bagging, Random Forest, AdaBoost si Stacking pt seturile de date: iris.arff, glass.arff

a) Utilizaţi valorile implicite ale parametrilor
b) Imbunătăţiţi comportamentul pt Vote, Bagging si AdaBoost înlocuind clasificatorul de bază cu alt clasificator.

2. Text mining

Analiza textului (text mining) are ca scop extragerea de informatii din documente (documentele sunt interpretate ca secvente de cuvinte). Principalele tipuri de prelucrari sunt: clasificarea si gruparea documentelor pe baza continutului lor. Cea mai simpla abordare se bazeaza pe aplicarea urmatoarelor etape:

- Pre-procesarea textului prin:
  - Eliminarea cuvintelor de legatura (*stop words).* Liste cu cuvinte de legatura pt diferite limbi pot fi gasite la http://www.ranks.nl/stopwords
  - Transformarea cuvintelor prin *stemming* (i.e. reducerea la radacina cuvintului). Cel mai popular algoritm de stemming este cel propus de Porter (vezi http://tartarus.org/martin/PorterStemmer/). Un serviciu web pt stemming este disponibil la http://text-processing.com/demo/stem/
- Construirea pt fiecare document a vectorului de frecvente(*frequency vector):* nr de aparitii ale fiecarui cuvant din dictionar in cadrul documentului.

Exercitiul 4/Weka.

a) Deschideţi fişierul movieReviews.arff (conţine recenzii ale unor filme clasificate în două categorii: pozitive şi negative)
b) Construiţi setul de date cu ocurenţele termenilor în colecţia de recenzii folosind Filters->Unsupervised->Attribute->StringToWordVector
c) Aplicaţi un clasificator (e.g. Naïve Bayes) asupra setului de date. Obs: e necesar ca primul atribut (@@class@@) să fie definit ca atribut de clasă (folosind Edit, click dreapta pe @@class@@ şi selectând Attribute as class)
d) Analizaţi impactul utilizării etapei de stemming asupra calităţii clasificării.