**Lab 7: Data Mining.**
**Serii temporale**
**Metode de tip ansamblu**
**Analiza text**

---

## 1. Serii temporale

Analiza seriilor temporale are ca scop sa modeleze si sa explice dependenta unor date de momente de timp successive. Exemple tipice de serii temporale sunt: temperature inregistrate zilnic, curs de schimb valutar, pretul unor actiuni etc.

Principalele prelucrari care pot fi efectuate asupra unei serii de timp sunt:

- *Pre-procesare* (de exemplu, transformarea seriei prin normalizare sau standardizare, completarea valorilor absente prin interpolare, eliminarea zgomotului prin netezire, eliminarea tendintei prin calcularea diferentelor dintre elemente successive etc)
- *Predictie*: estimarea valorilor ulterioare din serie pe baza valorii curente si a celor anterioare (folosind un model care descrie dependenta valorii curente din serie de valorile anterioare).

Un proces de predictie este caracterizat prin:

- *Intrare:* datele de intrare sunt valori anterioare din serie
- *Iesire:* rezultatul reprezinta valoarea/valorile urmatoare din serie
- *Model:* un model de regresie care descrie legatura dintre valoarea curenta a seriei si valorile anterioare (numarul de valori anterioare despre care sa considera ca influenteaza valoarea curenta este denumit intarzierea seriei (*time-lag*))

Consideram seria $X_1$, $X_2$,…., $X_n$ si intarzierea T. Deci valoarea curenta $x_i$ depinde de valorile $X_{i-1}$,$X_{i-2}$,…,$X_{i-T}$ . Prin urmare secventa de valori din serie poate fi transformata un al set de date in care sunt T atribute predictor si un atribut prezis:

| *Atribute predictor* | *Atribut prezis* |
|---|---|
| $X_1$ $X_2$ … $X_i$ … $X_T$ | $X_{T+1}$ |
| $X_2$ $X_3$ … $X_{i+1}$ … $X_{T+1}$ | $X_{T+2}$ |
| … | |
| $X_{n-T}$ $X_{n-T+1}$ … $X_{n-i}$ … $X_{n-1}$ | $X_n$ |

Folosind acest set de date se poate construi un model de regresie (in aceeasi maniera ca pentru date care nu sunt temporale). Una dintre principalele dificultati este alegerea adecvata a valorii T.

**Exercitiul 1.**

a) Deschideti fisierul airlines.arff (continand nr de pasageri ai unei companii aeriene inregistrat lunar in perioada 1949 – 1960)
b) Construiti un nou set de date folosind o intarziere T=12. Indicatie: utilizati Weka pt eliminarea atributului corespunzator date si Excel (sau un limbaj de programare) pt construirea noului set de date
c) Aplicati un model de regresie pentru noul set de date si analizati rezultatele obtinute

**Exercitiul 2.** (optional – doar pt versiune Weka >=3.7.3)

a) Instalati pachetul Time Series Forecasting utilizand Weka GUI Chooser ->Tools->Package manager si selectand pentru instalare timeSeriesForecasting
b) Deschideti fisierul airlines.arff
c) Preziceti urmatoarele 6 utilizand unul dintre urmatoarele modele: (i) linear regression; (ii) multilayer perceptron; (iii) random forests. *Indicatie*: selectia modelului se realizeaza utilizand panelul Advanced Configuration->Based Learner

Obs: detalii privind pachetul TimeSeriesForecasting pot fi gasite la http://wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka.

## 2. Metode de tip ansamblu (ensemble models)

Sunt meta-modele care se obtin din cateva modele de baza antrenate pe acelasi set sau pe seturi diferite de antrenare. Exista mai multe variante de a construe modele de tip ansamblu:

- Utilizand modele bazate pe algoritmi diferiti antrenati pe acelasi set de date (e.g. *bucket of models*)
- Utilizand modele bazate pe acelasi algoritm dar antrenate pe seturi diferite de date (e.g. *bagging* and *boosting*)
- Utilizand diferite modele si impartind setul de date (e.g. *stacking*)

**Exercitiul 3.** Utilizand Weka Experimenter comparati performanta urmatoarelor metamodele: Vote, Bagging, Random Forest, AdaBoost si Stacking pt seturile de date: iris.arff, glass.arff

a) Utilizati valorile implicite ale parametrilor
b) Imbunatatiti comportamentul pt Vote, Bagging si AdaBoost inlocuind clasificatorul de baza cu alt clasificator.

## 3. Analiza textului

Analiza textului 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. based on their content. 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 dictionary in cadrul documentului. D

**Exercitiul 4.**

a) Open the file movieReviews.arff (it contains reviews on movies grouped in two categories: positive and negative)
b) Construct the dataset with the occurrence of terms in the collection of reviews by using Filters->Unsupervised->Attribute->StringToWordVector
c) Apply a classifier (e.g. Naïve Bayes) to the dataset. Remark: it requires to set first the attribute called @@class@@ as class attribute (using Edit, right click on @@class@@ and selecting Attribute as class)
d) Analyze the impact of using a stemming step on the quality of the classification.