Lab 7: Data Mining.
Time series
Ensemble methods
Text mining

---

1. Time series

Time series analysis aims to model and explain a time-dependent series of data points. The analysis is usually based on statistical techniques. Time series appear in various domains and can be of various types. Typical examples are: daily temperature recorded during one month/year, daily exchange rate, prices of stocks etc.

The main tasks concerning time series are:

- *Pre-processing* (e.g. transform the time series by normalization or standardization, fill the missing values by interpolation, remove the noise by smoothing, remove the trend by differencing)
- *Forecasting* (prediction): estimate future values in the series based on known past values, by using a model.

The forecasting task is characterized by:

- *Input:* The input data are previous values in a time series
- *Output:* The result is the next value(s) in the series
- *Model:* a regression model describing the relationship between the current value in the timeseries and several previous values (the number of previous values which are considered to have an influence on the current value is the so-called *time-lag*)

Let suppose that the we have a timeseries X_1, X_2,…., X_n and let us consider that T is the time lag. Thus it is considered that a current value x_i depends on X_{i-1},X_{i-2},…,X_{i-T} . Therefore the sequence of values in the timeseries can be transformed in another dataset having T attributes playing the role of predictors and an attribute playing the role of predicted value:

T predictor attributes	*attribute to be predicted*
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

Based on the dataset containing the lagged attributes one can construct a regression model (in the same way as in the case of non-temporal data). One of the main issues is the choice of the time lag

Exercise 1.

a) Open the file airlines.arff (containing the monthly passenger numbers for an airline for the years 1949 – 1960)
b) Construct a new dataset by using a time lag of 12. Hint: you can use the Weka editor, Excel or any text editor
c) Apply a linear regression model to the new dataset and predict the next value in the timeseries

Exercise 2. (optional – only for Weka versions >=3.7.3)

a) Install the Time Series Forecasting package by using Weka GUI Chooser ->Tools->Package manager and by selecting for installation timeSeriesForecasting
b) Open the file airlines.arff
c) Forecast the next 6 values by using one of the following models for regression: (i) linear regression; (ii) multilayer perceptron; (iii) random forests. Hint: the selection of the model is done by using the Advanced Configuration->Based Learner panel

Remarks: details on the package TimeSeriesForecasting can be found at http://wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka.

2. Ensemble methods

Ensemble methods construct *meta-models* consisting of several base models (e.g. classifiers). They are constructed such that to enhance the performance of component models by reducing the *bias* and *variance* components of the error (see Lecture 12). There are different ways of constructing ensemble methods:

- By generating different models based on one training dataset (e.g. *bucket of models*)
- By generating different models through a sampling process of the training dataset (e.g. *bagging* and *boosting*)
- By using different models and in the same time splitting the dataset (e.g. *stacking*)

Exercise 3. By using Weka Experimenter compare the performance of the following ensemble models (metamodels): Vote, Bagging, Random Forest, AdaBoost and Stacking for the following datasets: iris.arff, glass.arff

a) Use the default values of the parameters
b) Try to improve the behavior of Vote, Bagging and AdaBoost by replacing the default individual Classifier with other models.

3. Text mining

Text mining refers to extracting information from documents (interpreted as sequence of words). The main text mining tasks are classification and clustering of documents based on their content. The simplest approach for classification/clustering documents is based on the following steps:

- Pre-process the text by:
  - Removing the *stop words* (words which do not provide specific information being rather syntactic components used to link various parts of speech). Lists of stopwords corresponding to different languages can be found at http://www.ranks.nl/stopwords
  - Transform the words by *stemming* (i.e. reduces the inflected variants of words to their root form). The most popular stemming algorithm is that proposed by Porter (see http://tartarus.org/martin/PorterStemmer/). A web service for stemming in various languages is available at http://text-processing.com/demo/stem/

- Construct for each document a *frequency vector* containing quantitative measures of the presence of words belonging to a dictionary in each of the documents. If the dictionary contains N words then to each document in the collection of documents to be processed one have to associate a vector of N elements specifying the number of occurrences of the corresponding word in the document. Since words which are specific to only some document have a higher discriminative power, instead of using frequencies of terms it is used the so-called *TF-IDF* (term frequency – inverse document frequency) encoding characterized by the fact that the frequency of a term in a given document is divided by the number of documents in the collection which contain that term. Once these numerical vectors are constructed then one can apply any classification/clustering technique.

Exercise 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.