

Data Mining

Lab 5:

Association rules
Regression models

1. Association rules

Example (market basket problem). Let us consider a set of transactions (T_1, T_2, \dots, T_n) , each one containing a set of items. For instance:

T1: {bread, milk, water}
T2: {bread, meat, water}
T3: {bread, butter, meat, water}
T4: {fruits, water}

We are looking for items which are frequently purchased together and for IF... THEN rules expressing associations between items (e.g. IF bread AND water THEN meat).

In an association rule IF A THEN B (also denoted as $A \rightarrow B$) the left hand side term (A) is an antecedent, and the right hand side term (B) is the consequent.

From a given set of transactions one can extract many rules – it is necessary to evaluate their relevance in order to provide ranked list of rules (with the most relevant rules in top).

To evaluate the relevance of a rule we can use at least two measures:

- **Support:** $\text{supp}(A \rightarrow B)$ = the number of transactions which contain both A and B divided by the total number of transactions
- **Confidence:** $\text{conf}(A \rightarrow B)$ = the number of transactions which contain both A and B divided by the number of transactions which contain A

Example: IF bread AND water THEN meat

$A = \{\text{bread, water}\}, B = \{\text{meat}\}$

$\text{Supp}(A \rightarrow B) = 2/4 = 0.5$

$\text{Conf}(A \rightarrow B) = 2/3 = 0.6$

Remark: besides these measures there are other indicators which quantify the degree of novelty (or interestingness) of the rule. Such an indicator is the lift, computed as in the following equation:

$\text{Lift}(A \rightarrow B) = \text{prob}(A, B) / (\text{prob}(A)\text{prob}(B))$

The probability involved in the computation can be estimated as the relative frequency. The rule is interesting if lift value is large. If the lift value is close to 1 this suggests that A and B are not correlated thus one cannot extract useful association rules of type $A \rightarrow B$

Example: R=IF bread AND meat THEN water
Conf(R)=2/2=1
Lift(R)=0.5/(0.5*1)=1

APRIORI algorithm

Input data: set of transaction (each transaction contains a list of items)

Control parameters:

- Minimum support threshold (e.g.: 0.2)
- Minimum confidence threshold (e.g.: 0.9)

The general structure of the algorithm:

Step 1: identify the frequent itemsets (itemsets with a support higher than the threshold):

- Identify the frequent 1-itemsets (sets containing only one frequent item) - list L_1
- FOR $k=2, K$ DO construct the list L_k containing frequent k-itemsets by joining elements from L_{k-1} (two elements from L_{k-1} having k-2 common elements are joined)

Step 2: construct rules by partitioning the itemsets identified at Step 1 in two parts (one part for the antecedent and the other part for the consequent of the rule); only the rules with a confidence level higher than the threshold are kept.

Exercise 1.

- a) Open in Weka the file [supermarket.arff](#)
- b) Find association rules using [Associate->Apriori](#) (with the default values of the parameters)
- c) Apply the same algorithm for other values of the thresholds for the support ([lowerBoundMinSupport=0.2](#)) and for the confidence ([minMetric=0.75](#)).

2. Regression models

2.1. Linear regression

In the linear models, the dependence between the predicted variables and the predictors is described by a linear function $Y=WX$.

The parameters of the model (elements of matrix W) are estimated based on the data by using a least squares minimization procedure.

Exercise 2.

- a) Open in Weka the file [autoPrice.arff](#)
- b) Use [Classify->Functions->SimpleLinearRegression](#) to find a linear relationship between the output attribute (price) and the most relevant input attribute. Analyze the values corresponding to the [Correlation Coefficient](#) and [Mean Absolute Error](#).
- c) Use [Classify->Functions->LinearRegression](#) to do the same thing

2.2. Nonlinear regression

Exercise 3. (also for the file [autoPrice.arff](#))

- d) Use [Classify->Functions->MultilayerPerceptron](#) (with the default values of the parameters). Analyze the values corresponding to [Correlation Coefficient](#) and [Mean Absolute Error](#).
- e) Identify in the category [Classify->Trees](#) the variant which allows the construction of a regression tree

Remark. Versions of Weka less than 3.8 contained also a simple [RBF Network](#) implementation

Exercise 4. Perform same operations as in Exercises 2 and 3 in the case of the dataset [autoMPG.arff](#) and analyze the differences (particularly with respect to the regression tree).