Data Mining: Practical Assignment #3

Due on Thu & Fri, May 08-09 2014,

Task 1

The test of a 3-class classification system on 100 test data points yields the following confusion matrix:

Predicted \Actual	Class 1	Class 2	Class 3
Class 1	15	1	9
Class 2	3	5	6
Class 3	2	4	55

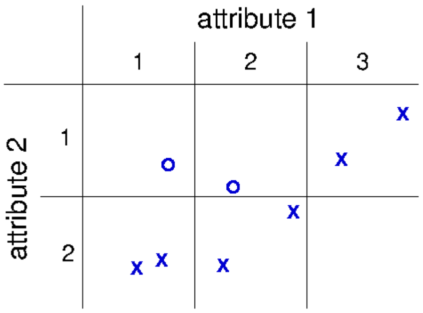
Let us assume that we are interested in results for 'class 1' and want to apply a number of metrics from the lecture. Write down the confusion matrix that results from combining 'class 2' and 'class 3' to 'other' (this should result in four entries). Then compute the following evaluation metric values and give an intuitive description of their meaning:

1. accuracy
2. error
3. true positive rate
4. sensitivity
5. recall
6. false positive rate
7. specificity
8. precision
9. F1-score (sometimes simply called F-score, if $\alpha = 1$ is implied)

Task 2

In the lecture you learnt about the apriori algorithm. In brief, the algorithm gets data and a minimum support value serving as a threshold. Now, it generates iteratively chunks of data items occuring together. Open the DAMI apriori.m file and go through the code to understand the implementation. Take the data provided in the data.zip and run the algorithm. What does the output depict and why? You may need to comment the code (replace the ??? comments with meaningful information) to provide a better understanding of the steps.

Task 3



attribute 1	attribute 2	Class
1	1	1
1	2	2
1	2	2
2	1	1
2	2	2
2	2	2
3	1	2
3	1	2

From large data we can learn a tree of decisions that help us to very quick classify new data. Here we want to practice the underlying computations, namely the information (entropy) of certain features in the data. To get started have a look on the toy data set in the figure on the left. In the given data set circles denote class 1 members and crosses denote class 2 members. Discretised into bins, these data can also be written in form of a table (figure, right).
The ID3 and C4.5 decision tree algorithms would first split the data either into the 3 partitions of attribute 1 or into the 2 partitions of attribute 2, depending on information gain. Which split would be made? Verify this choice by computing which split leads to the larger information gain.
Hint: you need to compute the following values

1. the entropy $\mathit{Info}(D)$ of the entire data (irrespective the attribute values)
2. the entropy $\mathit{Info}(D_{a_i})$ for each value i of each attribute a (irrespective the other attribute)
3. the average entropy $\mathit{Info}(D_a)$ for each attribute
4. the information gain $\mathit{Info}(D) - \mathit{Info}(D_a)$ for each attribute

Note: the logarithm to base 2 function in Matlab is called `log2`.

Task 4

Now lets compute a decision tree on the Auto Miles-per-gallon Data Set with Matlab. A thorough description can be found in the ReadMe.txt file. Before you can call the according tree, you have to prepare the data. From task 2 you should be familiar with the cell array concept.

1. Import the data into Matlab.
2. Split the data into the classes to decide upon and the features.
3. Inform yourself about the Matlab function for decision trees and perform the computation with the preprocessed data.
4. Visualize the tree.
5. With the gained knowledge do the same process again for the data in task 3, to verify your earlier result.

Clarify how and why the algorithm split the tree.