Exercise Sheet 5

Data Mining II – Classification

*Submit your solutions until **Monday, 11.04.2016, 12h00** by uploading them to ILIAS. Later submissions won't be considered. Every solution should contain the **name(s), email adress(es)** and **registration number(s)** of its (co-)editor(s).*

1 Classification (6 Points)

A machine learning technique that is not mentioned in the lecture is a *neural network* which follows the idea of modeling the functionality of a brain using connected neurons that react depending on the neurons they are connected to. Because of a neural network's structure it is not easy to see what was learned from the given training data. Read the following story about an experiment from the 80s where the US military tried to detect hidden tanks in images using a neural network classifier: http://neil.fraser.name/writing/tank/

a) (4 Points)

Why did the neural network fail? Would another classifier have had better results?

b) (2 Points)

If the scientists had trained a decision tree to classify the pictures they would have known the reason for the misclassification a lot faster. Why?

2 *k* Nearest Neighbors (12 Points)

The training data in Figure 1 consists of four classes, namely $C = \{$*red triangle, blue square, green star, black heart*$\}$. No label is assigned to the data points *A, B, C, D, E,* and *F*, yet.

2.1 kNN Classification (8 Points)

Use the k nearest neighbor algorithm with $k = 4$ on the training data to assign each of the unlabeled instances to one class. Euclidean distance is employed to compute distances.

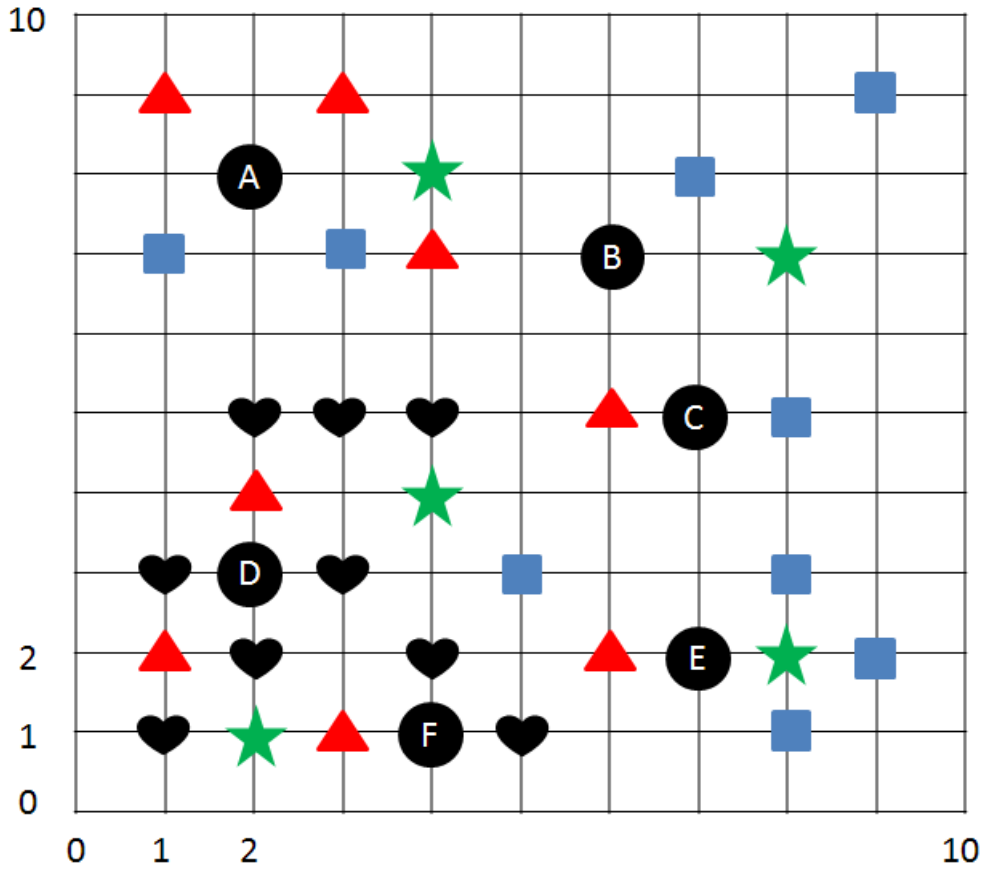

Figure 1: Data for k nearest neighbor classification

Do **not** use the newly classified data points to classify any of the other data points. For instance, if you classify point B as a *blue square*, do not use this additional information for the classification of point C.

Explain how you chose the labels for A and C.

#	Heart Rate	Blood Pressure	Class
1	irregular	normal	Ill
2	regular	normal	Healthy
3	irregular	abnormal	Ill
4	irregular	normal	Ill
5	regular	normal	Healthy
6	regular	abnormal	Ill
7	regular	normal	Healthy
8	regular	normal	Healthy

Table 1: Medical Training data

2.2 The Green Star (4 Points)

a) (2 Points)

Can you find a point in the data that would be classified as belonging to the green star class? Why?

b) (2 Points)

How would your answer in a) change for different values of k?

3 Decision Trees (12 Points)

3.1 A Health Status Classifier (6 Points)

A hospital wants to help doctors with their diagnose of patients. Health data of several patients were collected for that task and can be seen in Table 1. To automatically classify a patient's health status, build a *decision tree* using the *Information Gain*. List all computations involved in detail and draw the resulting decision tree!

3.2 Deep vs. Broad Trees (4 Points)

Should a decision tree be rather deep or broad? Explain why!

3.3 Feature Value Range (2 Points)

Imagine that the blood pressure would be stored in absolute values. How would the decision tree look like if we use the same strategy to build it as in 3.1? An informal description is sufficient.

4 Classification with RapidMiner (12 Points + 13 Bonus Points)

RapidMiner is a framework that provides a user interface to create processing pipelines using all sorts of machine learning algorithms and data processing libraries. Download and install it from http://www.rapid-i.com.

Besides the videos on the RapidMiner website there is a useful playlist on youtube with a tutorial on text analytics with RapidMiner. The first part starts here: https://www.youtube.com/watch?v=hpvda_Rfg3s&list=PL7669FFBBA1825900 Watch these videos to learn how to work with RapidMiner, the following tasks assume that you have seen them.

We will work with an example corpus *sunburn.arff* containing information on *factors affecting sunburn* that you can download from ILIAS. The *.arff* file format is optimized to store values for machine learning task and was developed at the university of Waikato for their machine learning framework Weka (http://www.cs.waikato.ac.nz/ml/weka/).

4.1 Learning a decision tree using ID3 (12 Points)

The ID3 algorithm presented in the lecture can learn decision trees from a labeled data set.

a) 5 Points

Load the *sunburn.arff* file into RapidMiner.[1] The .arff file does not deliver the information which feature provides the class label to be learned. You can set the *target role* of the attribute "result" to the value "label" with the *set role* operation to let RapidMiner know that this is the class to be learned.

Find the *ID3* operator and apply it to the data in order to learn a decision tree. Set the *criterion* to "information_gain" Report the result in text format (click on "Text View" in the resulting tree view).

b) 3 Points

- Edit the .arff file that you used in the previous task and uncomment the last line (remove the percent symbol). Rerun your pipeline to create the decision tree and again report the result in text format.
- Inspect the features that are used in the resulting decision tree. Would you have chosen the same features for classifying the examples? Why, or why not?

[1] *HINT*: You can search for functions with the *Filter* textfield of the *Operators* tab.

c) 4 Points

Delete the additional line from Task b). Now add a new entry so that the first split in the learned decision tree is made at the feature *lotion* instead of *hair*. Report the line that you added and the resulting tree (text format).

4.2 Cross Validation (13 Bonus Points)

a) 5 Bonus Points

Explain how n-fold cross validation works and why it is needed to evaluate the performance of a model.

b) 3 Bonus Points

Per default RapidMiner's cross validation uses *stratified sampling* rather than linear sampling. Explain what stratified sampling is and why it is generally safer to use it than linear sampling.

c) 5 Bonus Points

RapidMiner comes with example data sets, one of them is the Iris data set containing descriptions of flowers with length and width of the sepals and petals in centimetres.

Apply k nearest neighbor classification (k-NN in Lazy Modeling) on the Iris data set and perform 10-fold cross validation (*X-validation*)[2]. Report the accuracy of the results and show the confusion matrix.

[2] see https://jeszysblog.wordpress.com/2012/04/13/cross-validation-in-rapidminer/