TDT4300 Datavarehus og datagruvedrift - Spring 2012
Assignment 3: Classification

Spring 2013

1 Decision Trees

The first task is dealing with decision trees. Read the following example and answer the questions.

In a small computer business, Compushop, which only sells large computer equipment for youth and students, they want to predict who can pay for themselves and thus are allowed to buy the PC or not. The table below is an example of the decisions they made earlier that they think they can use as a basis for new decisions. Assume that each user record has five attributes as follows:

- Age: which may have values {Young, Medium young, Old}
- Income: that may have values {Low, Medium, High}
- Student: who may have values {yes, no}
- Creditworthiness: that can have values {Pass, High}
- Buy a PC: can have values {yes, no}

1. Compute the Gini index for the entire training set (Table 1).
2. Compute the Gini index for the UserID attribute.
3. Compute the Gini index for the Age attribute.
4. Compute the Gini index for Student attribute.
5. Compute the Gini index for Creditworthiness attribute.
6. Which attribute is the best?

UserId	**Age**	**Income**	**Student**	**creditworthiness**	**Buy PC**
1	young	high	no	pass	No
2	young	high	no	high	No
3	medium young	high	no	pass	Yes
4	old	medium	no	pass	Yes
5	old	low	no	pass	Yes
6	old	low	yes	high	No
7	medium young	low	yes	high	Yes
8	young	medium	no	pass	No
9	young	low	yes	pass	Yes
10	old	medium	yes	pass	Yes
11	young	medium	yes	high	Yes
12	medium young	medium	no	high	Yes
13	medium young	high	yes	pass	Yes
14	old	medium	no	high	No
15	medium young	medium	yes	pass	No
16	medium young	medium	yes	high	Yes
17	young	low	yes	high	Yes
18	old	high	no	pass	No
19	old	low	no	high	No
20	young	medium	yes	high	Yes

Table 1: Sample Dataset

7. Explain why UserID should not be used as attributtestbetingelse?

Suppose we have the following two users we want to predict whether they should be able to buy a PC or not. Explain how you want to proceed.

1. User # 21: A young student with medium income and ”high” creditworthiness.
2. User # 22: A young non-student with low income and ”pass” creditworthiness.

2 Classification

The second task is dealing with classification. You will use an open source toolkit to test the algorithms we had in the lectures (and others).

2.1 Software

Weka can be downloaded at:
`http://www.cs.waikato.ac.nz/ml/weka/`

Once unzipped, the main directory contains the file *weka.jar*, which can be started by typing “java -jar weka” (for larger datasets it might be necessary to pass on extra memory, e.g. 2 giga with “-Xmx2g”). From there choose the “Explorer” application to work with.

The data directory contains a list of sample data sets in Weka’s arff format. These files can be loaded in the Weka explorer (Preprocess tab). Once a dataset is loaded, you can switch to the Classify tab to perform classification.

More specifically, we will look at the following datasets:

- iris
- diabetes
- spambase

2.2 Datasets

Summarise the unique properties of the datasets. How many variables do they contain? How many instances? What kind of data do they describe? Which one do you assume will be the most difficult to classify?

2.3 Classification

Classify each of the datasets with at least the following algorithms (you can use “percentage split” for now, the 66% standard setting is an ok value:

- J48 (decision trees, “trees/J48” in weka
- k-NN (instance based learning, “lazy/IBk” in weka

- Support vector machines, “functions/SMO” in weka

For each of them list the most important parameters and describe how changing them influences the overall result (e.g. k in k-NN). As overall result, you can report “Accuracy” unless you think it’s not fitting the particular use-case.

2.4 Evaluation

Until now you used percentage split. Describe cross-validation and in what cases cross-validation might be a better approach. Address the issue of ordering of instances. What happens if the dataset is ordered according to classes (i.e. first all instances of one class, then all instances of another). Imagine you have three classes and a percentage split of 66%. Is it reasonable to perform percentage split here?

2.5 Best Classifiers

Which of the classifiers report the best results? Is any one of them consistently better than the others? Can you find other classifiers in weka which produce better results?

Notes

Your submission in its learning is a **pdf** file with your report. Include tables to show the classification results. Your submission should be between two and three pages.