Data Mining and Exploration
Interim Report

Thorvaldur Helgason (s1237131)
Daniel Stanoescu (s0838600)
Maria Alexandra Alecu (STUDENT NUMBER)

February 27, 2013

What we have done so far:

Data

The data we have chosen for this project is the Orange telecom customer behaviour dataset. We have been provided with a labelled training set and unlabeled test set for the data focusing on predicting the customer churn (switch providers), appetancy (tendancy to buy new products) or purchase upgrades. Given the proposed dataset, we are looking at solving a classification problem. To respect customer privacy, both the order of the customers as well as the variables in the dataset have been shuffled.

There are two datasets for this task, one small with 230 features and another big with 15000 features. We have decided to work solely on the small dataset to begin with because of lack of proper computational resources, but we might try working on the big dataset later given time. Also, we are only going to be using the training data and labels because we do not have labels for the test data and can therefore not use it for evaluation. Instead we are splitting the training data and labels up so 80% of it will be our training data and 20% our test data. We also make sure that the split is balanced in terms of classes.

After inspecting the data, we have identified that it contains both numerical and categorical values. We used R to explore as well as preprocess the data. We began with preprocessing the numerical data. We replaced all the entries that were empty with zero. Secondly we replaced the zero entries with the mean of the values of each variable in the dataset. In the case of the categorical data, we changed the empty strings to Not Available. We then converted all the strings to numeric values so we can handle data operations easier. Additionally, we created a binary bag-of-features dataset that was specifically created to see if it would work better with the Naive Bayes classifier.

Classifiers

We began experimenting with a series of algorithms in order to figure out which ones would perform better. We looked at employing SVMs but we quickly abandoned the idea as they would require a higher computational requirement than other methods. We have also familiarized ourselves with Naive Bayes and Decision Trees.

We are looking at Naive Bayes because that was used as a basis in the competition and we want to create our own baseline for our modified dataset. So far, it looks like Naive Bayes is performing better on the raw and binary datasets than the ones were missing values are replaced, although the overall performance is pretty poor (AUC ~ 0.5).

Decision trees - what we have done so far

We have implemented decision trees using the package "rpart" from R and tested it on the **pre-processed** data, using the first **50 features** of the data set.
As an **evaluation** of this classifier, we have used the following methods:

- 2-fold cross validation

- the "printcp" function available in the "rpart" package. This function gives the error from 10-fold cross validation performed on the data set, using different values for the number of nodes in the decision tree.

The **results** we obtained are as follows:

- With 2-fold cross validation, we obtained a very good accuracy, namely of 91
- The results from the "printcp" function seem to contradict this result, in the sense that this function shows that no matter how many nodes we include in our decision tree, there is no decision tree that fits the data.

What we plan on doing:

- See which pre-processing techniques and features work best for different classifiers.
- Familiarize ourselves with more classifiers.
- Compare the performance of the classifiers with the criteria described below.

We plan experiment with a linear classification method, more exactly, employing logistic regression methods. If there is time, a look into a mixture of gaussians to model the date will be attempted.

Decision trees - what we plan on doing

From the **data set** point of view, our goals are the following:

- use the same functions on the **unprocessed** data (since decision trees can handle missing data and categorical data)
- use **all the features** of the data set
- perform cross-validation using **more folds**.

In what concerns the **methods** we want to use, our main goal is to use the classifiers we obtained at the previous step in **ensemble** methods, such as **random forests**, **bagging** and **boosting**. We plan to use the built-in function in R, such as "cforest" and also experiment with the same functions in Weka.

What comparisons we want to run:

- Perform 5-fold cross-validation on the training set and do final evaluation on the test set.
- For each classifier we store their accuracy, confusion matrix, ROC curve and AUC.