Assignment 3

This is the first part of Assignment 3. The deadline for the assignment is **January 14** at the beginning of the class. The assignment should be submitted either via turn-in, or via email. The details for the turn-in are on the web page of the course.

Question 1

In this question you will experiment with classification in WEKA. You will experiment with three different datasets, the Iris dataset, the Mushroom dataset, and the Spambase dataset, which you can get from the Material page of the course. You can find more details for each dataset on the UCI repository page. You will experiment with three different classification algorithms of WEKA: Naïve Bayes, Logistic Regression, and SimpleCart decision tree. For each algorithm and each dataset, you will do 10-fold cross validation, and you will submit the summary produced by WEKA at the end. For the decision tree, you will also submit the tree produced by WEKA.

Together with the results you will also submit a short report discussing the results: how hard the classification problem for each dataset is, and which algorithm performs best. Examine the produced decision trees, and discuss which attributes seem to be the most important in the classification.

Question 2

For this question you will experiment with classification of tips. So far you have been working with a set of tips about restaurants. We will now use tips from different categories. You are given a new file "category_tips.txt" which you can download from the Material page of the class. The file contains the tips you have been working with so far, plus tips about nightlife venues and about shops. Each line in the files consists of two tab-separated fields. The first is the category of the venue: it is either "food", "nightlife" or "shops"'; the second is the text of the tip.

Your task is to create a classifier which given a tip it predicts the category of the tip. As features you will use the words in the tip text, which will be weighted by the tf-idf value (similar to what you did for Assignment 2). You should remove the stop words, and you can also do other kinds of feature selection if you want. You will use a Logistic Regression classifier, using the implementation in some existing software package. You can use either WEKA (using sparse representation of the data), or the LibLinear package (there is a link to the LibLinear package on the Material page of the class).

You will use a randomly selected subset of 80% of the data for training, and the remaining 20% of the data for testing (you can use the code for sampling that you created for Assignment 1). The Logistic Regression classifier will produce a probability for each tip to belong to a class, and you will assign the tip to the class with the highest probability. Compute the accuracy of the classifier. Consider also the case that you assign a tip to the class with the highest probability only if the probability is above a

threshold θ. For each class, compute the precision-recall curve for threshold values $\theta \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

Write a short report about your implementation, your results, and the quality of your classifier. Is there some threshold that seems to work best? Hand in your code and your report.

Question 3

For this question you will implement a Naïve Bayes text classifier as described in class. You should do your own implementation using any language you want. You will use the same training and test data as in the previous question. You should do the same preprocessing as before, but in this case you will not use the tf-idf score.

Compute the accuracy of your classifier and compare it with the accuracy of the Logistic Regression classifier in Question 2. Append to the report of Question 2 your observations about the quality of the Naïve Bayes classifier and the comparison with the Logistic Regression classifier. Hand in your code and the report.

Note: When processing tab-separated data, the Unix commands on the Material page of the course may be useful.