Data Mining
Homework Assignment #9

Dmytro Fishman, Anna Leontjeva and Jaak Vilo

April 10, 2014

You are free to use any programming language you are comfortable with (unless otherwise stated).

Task 1

Consider the following dataset:

Sunny	Windy	PlayTennis
Yes	No	Yes
Yes	No	Yes
Yes	No	Yes
No	Yes	Yes
No	Yes	Yes
Yes	Yes	No
Yes	Yes	No
No	No	No
No	No	No
No	No	No

It is windy and not sunny today. Manually calculate a prediction for playing tennis using a Bayes classifier. Do the same using Naïve Bayes classifier. Which prediction is correct? Why? What properties of the Naïve Bayes classifier are involved?

Task 2

Use the same dataset as in Task 1 and manually calculate ID3 tree. Provide some intermediate steps and draw the final decision tree. Compare with the results in Task 1.

Task 3

Netflix was running a $1M challenge for the best possible machine learning algorithm (`http://courses.washington.edu/css490/2012.Winter/lecture_slides/08a_Netflix_Prize.pptx`). The test set was used to measure the goodness of the current best method (and call the competition to an end when the first team would beat the state of the art method by more than 10%. I.e. everyone could evaluate their best algorithm against this test data and get their current standing in the rankings. But the final evaluation happened on the third data set that was completely hidden from any contestants until the competition had ended. Why was that? Explain the reasons for this third data set for evaluations.

Task 4

Suppose you are given a task of classifying texts (e.g. sorting e-mail as spam or not). Is it a good idea to apply the K-nearest neighbors algorithm? How could you apply it? What if your training set is very large, how would you solve the algorithm performance problems? Be reasonably brief: no more than two-three short paragraphs total.

Task 5

Find yourself a project group (up to 3 person). Agree on a project topic area. Write group members, project title and an abstract of your project. During practice session we will discuss the proposed projects. Later you can decide whether to stick with your topic or to pick the one that will be proposed.

Task 6

(2 points)