Data Mining
Homework Assignment #9

Dmytro Fishman, Anna Leontjeva and Jaak Vilo

April 11, 2014

You are free to use any programming language you are comfortable with (unless otherwise stated). All the terms and notions used in this homework can be found on the lecture sliides `https://courses.cs.ut.ee/MTAT.03.183/2014_spring/uploads/Main/DM_L06_ML.pdf`, please take a look.

Task 1

Consider the following dataset:

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

It is quite *windy* and *sunny* today, temperature is *cold* and humidity is *high*. 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? [1]

[1] The task is based on `http://courses.cs.ut.ee/2009/dm/uploads/Main/DM_HW7_ML_II.pdf`

Task 2

Use the same dataset as from Task 1 and manually construct an 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 prize 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 (algorithm that was used by Netflix before competition started) by more than 10%. I.e. everyone could evaluate their best algorithm against this test data and get their current position 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.[1]

Task 5

Read the article by Domingos: A few useful things to know about machine learning (communications of the ACM, Vol. 55 No. 10, Pages 78-87 doi: 10.1145/2347736.2347755 via ACM Digital library, `https://courses.cs.ut.ee/MTAT.03.183/2012_fall/uploads/Main/domingos.pdf`). Make a list of key messages with a supporting 1-2 sentence example or clarification of that message (something like short summary of the article).

Task 6

Try building a classifier for discriminating spam messages. You have two options: use my_mails.txt for building your own classifier with any programming language of your preference (worth 2 points) or my_mails.arff and Weka (worth 1 point). This data (both files) contains e-mails subjects that were made public

as a part of the data mining competition associated with ICONIP 2010. How good is your classifier? What words contribute mostly to its decisions? You are free to approach this task in any way you deem appropriate.[1]

Note, after loading the file into Weka you will need to convert the string to a set of binary variables, one for each word. This can be done using the *StringToWordVector* filter. Also you need to transform your data from numerical to nominal by applying corresponding filter.