CMPE558 Data Mining
Fall 2018

Programming Assignment#2
Classification

Task

In this task you will experiment with classification using Support Vector Machines (SVM). SVMs have recently become very popular as machine learning based classifiers in many classification tasks. You should read Section 5.5 in the Text Book to understand the main principles of SVM.

Dataset

You should use the BiocreativeII Protein Interaction Article Selection Task dataset which contains 5495 articles each one represented with a set of features (each article may be represented with a different number of features). The first column denotes the class for each article. Download this dataset from the course web page (558PA2data)

Tool

You will use the SVM^{light} tool available at http://svmlight.joachims.org/ for classification. Download this tool and read the page which describes its functionality and usage from the web site above. However you may use any other SVM package you are comfortable with.

What to do

The data is organized in such a way that the first 3536 articles belong to class 1 and the remaining articles belong to class -1. You must first randomly mix the articles in the dataset such that this order is no longer preserved.

a) Train 5 different SVM classifiers using the first 10%, 20%, 30%, 40%, and 50% of the samples in the dataset for training. For each case use the remaining samples as validation set and calculate the precision, recall, F-score, and accuracy of each classifier. Investigate the effect of train data size and write your conclusions (backed up with your results) in your report.

b) Now perform a 10-fold cross validation task using the dataset and find the 10-fold cross validation of the SVM you are using. Give precision, recall, and F-scores for the 10-fold cross validation. How does the performance differ for each fold?

Report

You should write a short report (6-8 pages) describing the work you have done. Your report should briefly describe problem, the data used, the methods you have used to process the data and your results. The answer to all questions above must be in your report. You need to discuss the results highlighting main findings and possible reasons and conclusions. Just stating the results is never enough. Use figures and tables where necessary. You should conclude very briefly. Your report should make reference to all resources you have accessed and/or used. In case you have used readily available software you should both cite it and briefly explain its usage to me in your report. If you have worked with a partner, your report cover must include the name and number of your partner as well.

Programming Language

I suggest that you use the SVM*light* mentioned above or any other SVM package available. In addition, you are free to use any programming language you feel comfortable with. Please don't try to discuss with me the difficulties you may encounter during programming such as reading, writing data files, fixing bugs, codes not working etc. ☺. If you use publicly available software, again make sure it is properly cited and its usage is explained in the report.

Grading

This assignment is worth 15% of your CMPE558 course grade.

Working with Teams

- You may want to do the assignment on your own or in groups of 2.
- Members of the same team will may not receive the same grade.

Schedule:

- Assigned date: December 3, 2018
- Due date: December 17, 2018 at 9.30 a.m in class. There will be a 10% penalty applied to your grade for every day that it is not turned in.

What to turn in

- Your report.
- A CD which contains your program code(s) and an electronic version of your report. All material must be given under a folder whose name is your student id.
- All material must be handed it to me in person (in class); if you turn in your report later than the due date/time, drop it down at my pigeon box at the secretary's room if I am not around but make sure you send me an e-mail as soon as possible telling me that you have left it there. I will assume no responsibility for work slid under my door or left in the secretary's office without an e-mail notification that follows it.

Academic Honesty

As stated in the CMPE558 Course outline:

"All work submitted must be of your own. You are allowed to discuss solutions to programming assignments with your friends but you must not share codes or ideas in detail. In any case you must acknowledge the person you have shared ideas with in name and in writing. For programming assignments you are normally permitted to use code that is publicly available but it must be properly acknowledged and referenced. For all written work, you must quote (using quotation marks " ") any sentences taken from other published material For ideas used from other written sources, the source must be properly referenced. Plagiarism is a type of cheating (if you are not sure what is considered as plagiarism, please ask) and it will be dealt accordingly. It will result in a final course grade of "F" and may be referred to the EMU disciplinary committee."