CSCI-B565: Data Mining

Homework Assignment # 5

Assigned: 04/16/2018 *Due: 04/23/2018, 11:59pm, through Canvas*

Three questions, 80 points in total. Good luck!
Prof. Predrag Radivojac, Indiana University, Bloomington

Problem 1. (30 points) Naive Bayes classifier. Consider a binary classification problem where there are only four data points in the training set. That is $\mathcal{D} = \{(0, 0, -), (0, 1, +), (1, 0, +), (1, 1, -)\}$, where each tuple (x_1, x_2, y) represents a training example with input vector (x_1, x_2) and class label y.

a) (15 points) Construct a naive Bayes classifier for this problem and evaluate its accuracy on the training set..

b) (15 points) Transform the input space into a six-dimensional space $(1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$ and repeat the previous step.

Carry out all steps manually and show all your calculations.

Problem 2. (20 points) Let $\{i_1, i_2, \ldots, i_n\}$ be a set of n items that are used in sequential pattern discovery (Textbook, Section 7.4). Derive a formula or give an algorithm for calculating the number of potential k-sequences, where $k \geq 1$. Explain what each part in your expression means and then provide a table of the total number of possible sequential patterns for $k \in \{1, 2, \ldots, 5\}$.

Problem 3. (30 points) Mining time series data. In this question you will read two scientific publications and in your own words summarize each of them in 1000 words or less. Both papers are available on-line. Your summary should demonstrate the ability to succinctly present the main points of these papers and provide a short critical assessment of each paper. A critical assessment should demonstrate strengths and weaknesses and well as your opinion on the quality of these papers.

a) (15 points) Patel P, Keogh E, Lin J, Lonardi S. Mining motifs in massive time series databases. *IEEE International Conference on Data Mining*, ICDM 2002.

b) (15 points) Lin J, Keogh E, Lonardi S, Chiu B. A symbolic representation of time series, with implications for streaming algorithms. *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, DMKD 2003.

It is absolutely not allowed to copy any single sentence from the papers and use it in your summary. All sentences must be your own.

Homework Directions and Policies

Submit a single package containing all answers, results and code. You submission package should be compressed and named firstnamelastname.zip (e.g., predragradivojac.zip). In your package there should be a single pdf file named main.pdf that will contain answers to all questions, all figures, and all relevant results. Your solutions and answers must be typed[1] and make sure that you type your name and IU username (email) at the beginning of the file. The rest of the package should contain all code that you used. The code should be properly organized in folders and subfolders, one for each question or problem. All code, if applicable, should be turned in when you submit your assignment as it may be necessary to demo your programs to the associate instructors. Use Matlab, Python, R, or C/C++.

Unless there are legitimate circumstances, late assignments will be accepted up to 5 days after the due date and graded using the following rules:

on time: your score $\times$ 1

1 day late: your score $\times$ 0.9

2 days late: your score $\times$ 0.7

3 days late: your score $\times$ 0.5

4 days late: your score $\times$ 0.3

5 days late: your score $\times$ 0.1

For example, this means that if you submit 3 days late and get 80 points for your answers, your total number of points will be $80 \times 0.5 = 40$ points.

All assignments are individual, except when collaboration is explicitly allowed. All the sources used for problem solution must be acknowledged; e.g., web sites, books, research papers, personal communication with people, etc. Academic honesty is taken seriously! For detailed information see Indiana University Code of Student Rights, Responsibilities, and Conduct.

[1] We recommend Latex; in particular, TexShop-MacTeX combination for a Mac and TeXnicCenter-MiKTex combination on Windows. An easy way to start with Latex is to use the freely available Lyx. You can also use Microsoft Word or other programs that can display formulas professionally.