CSCI-B565: Data Mining

Homework Assignment # 2

Assigned: 02/10/2018 *Due: 02/25/2018, 11:59pm, through Canvas*

Five questions, 200 points in total. Good luck!
Prof. Predrag Radivojac, Indiana University, Bloomington

Problem 1. (20 points) Consider a binary classification scenario. Suppose you are given a data set $\{(x_i, y_i)\}_{i=1}^n$, where $y_i \in \{-1, +1\}$ and a set of predictions $\{p_i\}_{i=1}^n$, where $p_i \in \mathbb{R}$, obtained through cross-validation. What are the minimum and maximum number of thresholds for which the true positive and false positive rates have to be computed in order to exactly calculate the area under the ROC curve? Prove it. You may assume that the data set contains n_+ positive and $n_- = n - n_+$ negative examples.

Problem 2. (30 points) Understanding outliers and hubs in low- and high-dimensional spaces. Use random number generators to make a data set $\mathcal{D} = \{x_1, x_2, \dots x_n\}$ consisting of n real-valued examples with dimensionality k. Consider now a data point $x \in \mathcal{D}$. Count the number of times this data point is among $K = 5$ nearest neighbors over all other $n - 1$ data points, and refer to that number as $n_5(x)$. Calculate now n_5 for every data point in $\mathcal{D}$ to create a vector N_5 where the i-th value is $n_5(x_i)$ and x_i is the i-th data point in $\mathcal{D}$. Finally, make a plot in which the x-axis is the value of n_5 and the y-axis is the fraction of times in the data set this value of $n_5(x)$ occurred (roughly speaking, you are plotting the histogram of n_5 but not using the hist function).

Vary k from $\{3, 30, 300, 3000\}$ and keep n fixed at 10000. To generate data sets, use a uniform number generator as well as multi-variate zero-mean Gaussian generator with unit covariance matrix. Use the Euclidean distance, cosine distance, and the two functions provided below ($p = 2$) to measure the proximity between k-dimensional points. Make one plot for each combination of (k, distance measure, random number generator). However, it is recommended that you group the plots for the same k and use different colors for different distance functions. Clearly label axes in every figure.

The new distance functions are given as

$$d^p(x, y) = \left(\left(\textstyle\sum_{i=1}^k (x_i - y_i)^+ \right)^p + \left(\textstyle\sum_{i=1}^k (x_i - y_i)^- \right)^p \right)^{\frac{1}{p}} \tag{1}$$

and

$$d_N^p(x, y) = \frac{\left(\left(\sum_{i=1}^k (x_i - y_i)^+ \right)^p + \left(\sum_{i=1}^k (x_i - y_i)^- \right)^p \right)^{\frac{1}{p}}}{\sum_{i=1}^k \max\{|x_i|, |y_i|, |x_i - y_i|\}} \tag{2}$$

where $x, y \in \mathbb{R}^k$, $t^+ = \max(t, 0)$ and $t^- = \max(-t, 0)$.

Operationally, consider outliers to be data points that are among K-nearest neighbors of few points and hubs to be data points to be in the K-nearest neighborhoods of many points (consider only $K = 5$ but choose thresholds for “few” and “many” reasonably). Comment on what you observed with respect to outliers and hubs as k grows. Experiment with plotting in the log-log and semi-log scale if this better supports your argument. Could you propose and justify a numerical measure, say a formula, that supports your argument and could be used to quantify the problems with particular combinations (k, distance measure).

Problem 3. (80 points) K-means clustering. Implement k-means clustering and apply it on three data sets you selected from UCI Machine Learning Repository.

a) (10 points) The basic k-means algorithm that is intended to minimize the sum of squared errors objective function. The algorithm should take the data set $\mathcal{D}$ and the number of clusters K as inputs and return cluster assignments for each data point; i.e., an integer between 1 and K. The algorithm should also return the number of iterations and distance calculations as well as the total sum of squared errors.

b) (30 points) K-means clustering with Elkan's acceleration based on triangle inequality. The inputs and outputs should be exactly the same as in the previous question. Repeat Elkan's experiment for at least one data set in Table 2 of his 2003 ICML paper and confirm his observations. Use $K = 3$, $K = 20$, and $K = 100$ as in the original paper. It might be necessary to repeat experiment multiple times to stabilize the results.

c) (15 points) K-means clustering that enables the use of various other distance functions. Your algorithm should be able to select among: (i) Euclidean distance (default, as above), (ii) cosine distance, (iii) Cityblock distance, (iv) distance from Eq. 1 when $p = 2$, (v) distance from Eq. 2 when $p = 2$. In all cases the algorithm should compute the centroid as the mean of data points that belong to the cluster.

d) (25 points) Download three data sets of your choice from UCI ML Repository and evaluate the quality of clustering for all methods. Your data sets should be classification data sets such that the cluster indices can be compared with class labels after clustering. You should avoid binary classification problems; furthermore, it is recommended that one of the data sets be Iris. Set K to the number of classes in the original data set. In your comparisons, it may be necessary to perform multiple restarts for each algorithm and you can select the best run based on the objective function criterion (note: you have some flexibility in setting the objective function for various distances mentioned above). Therefore, your evaluations should also discuss the influence of multiple restarts.

Problem 4. (30 points) Formalizing clustering is difficult. 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 informed opinion on the quality of these papers.

a) (15 points) Kleinberg J. The impossibility theorem for clustering. *Advances in Neural Information Processing Systems*, NIPS 2002.

b) (15 points) Ackerman M, Ben-David S. Measures of clustering quality: a working set of axioms for clustering. *Advances in Neural Information Processing Systems*, NIPS 2008.

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.

Problem 5. (40 points) Chapter 8, Section 8.2.6 of the Tan et al. textbook gives a centroid derivation when the objective function is the sum of squared errors as well as the sum of absolute errors, in both cases for one-dimensional data.

a) (20 points) Derive the centroid when the objective function is the square of Eq. 1 with $p = 2$.

b) (20 points) Derive the centroid when the objective function is the square of Eq. 2 with $p = 2$.

Consider both one-dimensional and multi-dimensional inputs.

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 × 1

1 day late: your score × 0.9

2 days late: your score × 0.7

3 days late: your score × 0.5

4 days late: your score × 0.3

5 days late: your score × 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.