CSCI-B565: Data Mining

Homework Assignment # 1

Assigned: 01/18/2018 *Due: 02/03/2018, 11:59pm, through Canvas*

Six questions, 125 points in total. Good luck!
Prof. Predrag Radivojac, Indiana University, Bloomington

Read Chapter 2 of the Tan et al. textbook. Problems 1-5 are related to this material. Problem 6 covers Chapter 4.

Problem 1. (10 points) Term frequency-inverse document frequency transformation. Suppose you are given a document-term matrix corresponding to a set of n documents and a dictionary of m terms (words). Suppose further m_{ij} is the number of times that the j-th term appears in the i-th document, m_i is the number of terms in the i-th document and n_j is the number of documents containing the j-th term in the data set. Consider the following feature representation for the ij-th element of the document-term matrix

$$x_{ij} = \frac{m_{ij}}{m_i} \cdot \log \frac{n}{n_j},$$

and answer the following questions:

a) (4 points) What might be the benefits of this encoding in mining text documents?

b) (4 points) What might be disadvantages of this encoding compared to $x_{ij} = \frac{m_{ij}}{m_i}$ or $x_{ij} = m_{ij}$?

c) (2 points) What is the effect of this transformation if a term occurs in one document or in every document?

Problem 2. (10 points) Let $\mathbf{x}$ and $\mathbf{y}$ be k-dimensional column vectors from $\mathbb{R}^k$. Prove that

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|}$$

where $\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots x_k^2}$ is the length of vector $\mathbf{x}$. A good way to start might be to consider a right triangle and use the definition of a cosine; i.e., cosine of an angle is defined as the ratio of the lengths of the adjacent side and the hypotenuse.

Problem 3. (25 points) Metrics on sets. Let each $d : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ below be a distance function on pairs of sets from space $\mathcal{S}$. For each of the functions below, either prove it is a metric or provide a counterexample. Consider four distance functions:

a) (5 points) $d_1(A, B) = |A - B| + |B - A|$

b) (5 points) $d_2(A, B) = \frac{|A-B|+|B-A|}{|A \cup B|}$

c) (5 points) $d_3(A, B) = 1 - \left(\frac{1}{2} \cdot \frac{|A \cap B|}{|A|} + \frac{1}{2} \cdot \frac{|A \cap B|}{|B|} \right)$

d) (10 points) $d_4(A,B) = 1 - \left(\frac{1}{2} \cdot \frac{|A|}{|A \cap B|} + \frac{1}{2} \cdot \frac{|B|}{|A \cap B|}\right)^{-1}$

where $A - B$ is the set difference, $A \cup B$ is the set union, $A \cap B$ is the set intersection, and $|A|$ is the number of elements in the set A, also called the cardinality of A.

Problem 4. (20 points) Metrics on k-dimensional real vectors. Let $d : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ below be a distance function on pairs of vectors from $\mathbb{R}^k$; i.e.,

$$d(x,y) = \left(\left(\sum_{i:x_i>y_i}(x_i - y_i)\right)^p + \left(\sum_{i:x_i<y_i}(y_i - x_i)\right)^p\right)^{\frac{1}{p}}$$

where $x = (x_1, x_2, \ldots x_k)$, $y = (y_1, y_2, \ldots y_k)$, and p is a non-negative constant. Prove that d satisfies metric properties or provide a counterexample

a) (10 points) when $p \geq 1$.

b) (10 points) when $0 < p < 1$.

Problem 5. (25 points) Understanding the curse of dimensionality. Consider the following experiment: generate n data points with dimensionality k. Let each data point be generated using a uniform random number generator with values between 0 and 1. Now, for a given k, calculate

$$r(k) = \log_{10} \frac{d_{\max}(k) - d_{\min}(k)}{d_{\min}(k)}$$

where $d_{\max}(k)$ is the maximum distance between any pair of points and $d_{\min}(k)$ is minimum distance between any pair of points (you cannot use identical points to obtain the minimum distance of 0). Consider five distance functions: (i) Euclidean distance, (ii) citiblock distance, (iii) Minkowski distance when $p = 3$, (iv) distance from Problem 4 when $p = 2$, and cosine distance; i.e., $d(x,y) = 1 - \cos(x,y)$. Let k take each value from $\{1, 2, \ldots, 99, 100\}$. Repeat each experiment multiple times to get stable values by averaging the quantities over multiple runs for each k.

a) (15 points) Plot $r(k)$ as a function of k for three different values of n; $n \in \{100, 1000, 10000\}$, and for each distance function. Label and scale each axis properly to be able to make comparisons over different n's. Embed your final picture(s) in the file you are submitting for this assignment.

b) (5 points) Repeat everything but replace the uniform data generator with a zero-mean unit-covariance-matrix Gaussian source.

c) (5 points) Discuss your observations and also compare the results to your expectations prior to carrying out the experiment.

Problem 6. (35 points) Implementing classification trees and evaluating their accuracy.

a) (15 points) Implement the greedy algorithm that learns a classification tree given a data set. Assume that all features are numerical and properly find the best threshold for each split. Use Gini and information gain, as specified by the user, to decide on the best attribute to split in every step. Stop growing the tree when all examples in a node belong to the same class or the remaining examples contain identical features.

b) (15 points) Evaluate your classifier. Visit the UCI Machine Learning Repository and select three datasets for evaluating your tree. Select only those data sets where all features are numerical. In certain cases you can convert categorical features into numerical by encoding them using sparse binary representation. That is, if feature values belong to a set $\{\mathsf{blue}, \mathsf{yellow}, \mathsf{red}, \mathsf{green}\}$, encode this feature using 4-dimensional binary vectors such that if the feature value is blue, the encoding is $(1, 0, 0, 0)$, if the feature value is yellow, the encoding is $(0, 1, 0, 0)$, etc. Split each data set into two halves. Construct a tree on one half (training set) and then calculate the fraction of correctly classified data points on the other half (test set). Repeat this process 5 times by making a different random split and average the results for the final report. Discuss your findings and compare training and test accuracies.

c) (5 points) Compare Gini and information gain as splitting criteria and discuss your observations on the quality of splitting.

You can implement a classification tree using standard recursive partitioning of the data or you can choose to implement it without recursion.

To run your program log in to the following server

`hulk.soic.indiana.edu`

by using your university ID and password. Before submitting a report, you must give appropriate access to your folder to the professor and the teaching assistants (TA/AI) so that they can run and evaluate your code. You should also submit your code as part of the homework package. Make sure that the last modification in the code is before the submission deadline.

Note that you are not allowed to use any statistical or machine learning packages in this assignment. Allowable libraries may include Standard Template Library in C/C++, sorting libraries etc. If in doubt, consult with the instructors.

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.