COSC 4570/5010 Data Mining
Spring 2020
University of Wyoming

Homework #1

Due: February 07, 2020, 10:00 a.m.

Submission guideline You need to submit only one *.zip* file. Please name the file as "Your Net id_Homework1.zip".

1. Problems from the book

Solve the following from Chapter 2 from Problems 2, 5, 13, 15, and 16.

2. Sampling

- When is sampling with replacement appropriate and when is sampling without replacement more preferable? Provide two examples where each is more appropriate.

- Samples obtained uniformly at random can miss anomalies (underrepresented data points) in datasets. How can we sample more systematically using PCA?

3. Curse of Dimensionality

Due to curse of dimensionality distances become meaningless in high-dimensional spaces. That is, the minimum distance between random pairs of nodes becomes really close to the maximum distance between such pairs. In this problem, you are going to verify this phenomenon.

Write a program in Java, C, C++, Python, or MATLAB that

a. Generates n d-dimensional random points.

b. Computes the maximum and minimum distance using Euclidean distance between all $\binom{n}{2}$ pairs of nodes (i.e. $O(n^2)$ $operations$). Denote these values as $max(d,n)$ and $min(d,n)$ respectively.

c. Computes $\gamma(d,n) = \log \frac{\max(d,n)-\min(d,n)}{\min(d,n)}$

Change n in range $100 \leq n \leq 1{,}000$ and d in range $1 \leq d \leq 100$ and assume that feature values are in range [0,100] (or some other fixed range). Compute $\gamma(d,n)$ using your program and plot the 3-D surface of $\gamma(d,n)$ in MATLAB or your programming language of choice.

How does the surface change with respect to n? Perform the same experiment, but this time using l_1 norm for computing distances (Book, Page 70) and plot the surface. Submit your code and your two plots.

4. Weka

Download and install Weka from http://www.cs.waikato.ac.nz/ml/weka/ . The following tutorial is useful:

https://www.cs.auckland.ac.nz/courses/compsci367s1c/tutorials/IntroductionToWeka.pdf

You can also watch Weka Tutorials on YouTube. Get used to Weka and learn how to modify data in Weka. In particular, learn how to generate ARFF files for Weka from csv files (open an arff file in notepad from Weka data folder + visit http://www.cs.waikato.ac.nz/ml/weka/_arff.html*.) You can use Weka's own CSV-to-ARFF converter or you can write your own header for ARFF files.*

Load a dataset from the data folder of Weka (e.g., weka-3-6/data/weather.numeric.arff). You can always see the current stage of your dataset using the "edit" button. For each of the following, determine how it can be done using Weka and submit as part of your homework, the proper command and parameters. For example, if you want to center your data using Weka, you need to use the Center filter, under **weka.filters.unsupervised.attribute.Center**.

- Center data (having zero mean): Center
- Removing attribute 2 to 4
- Removing all attributes but the last
- ~~Removing~~ Reordering attributes 1,2,3,4,5 as 5,4,1,2,3
- Removing instances with missing values:
- How is a missing value denoted in an ARFF file?
- What does "visualize all" do?
- Removing all instances where the 3nd feature value is equal to 'x'.

5. PCA

The most known dataset in data mining/machine learning is the Iris dataset. Learn about this dataset at the following URL:

https://archive.ics.uci.edu/ml/datasets/iris.

Download the dataset (in Data Folder > iris.data). The dataset contains the information gathered on three types of iris plant: Iris Setosa, Iris Versicolour, and Iris Virginica. To see if different types of Iris plants are distinguishable from one another, we can just

visualize our dataset. The Iris dataset is a four-dimensional dataset; therefore, it cannot be visualized in 2D or 3D. Apply PCA to the dataset and reduce the dimensionality to two. Plot the dimensionality- reduced dataset. Color data points based on their plant type (e.g., Iris-setosa can be red, etc.). The plant type is given as the fifth column in the dataset. The figure should show the type(s) that can be easily distinguished from others. Which Iris type is the easiest to distinguish from the rest? Submit your code and plot.