COSC 4570/5010 Data Mining
Spring 2020
University of Wyoming

Homework #3

Due: **April 03, 2020, 10:00 a.m.**

Submission guideline You need to submit only one *.zip* file. Please name the file as "Your Net id_Homework3.zip".

1. Problems from the book (Introduction to Data Mining 2nd Edition by Tan, Steinbach et al.)

Solve the following:

Chapter 4: Problems 16 and 21.

Chapter 7: Problems 7 and 11.

OR

Problems from the book (Introduction to Data Mining 1st Edition by Tan, Steinbach et al.)

Solve the following:

Chapter 5: Problems 17 and 23.

Chapter 8: Problems 7 and 11.

2. Clustering

a) Given k clusters and their respective cluster sizes $s_1, s_2, \dots, s_k$, what is the probability that two random (with replacement) data vectors (from the clustered dataset) belong to the same cluster?

b) Now assume you are given this probability (you don't have s_i's and k), and the fact that clusters are equally sized, can you find k? This gives you an idea for predicting the number of clusters in a dataset.

c) Give an example of a dataset consisting of 4 data vectors where there exist two different optimal (minimum SSE) 2-means (k-means, k=2) clustering of the dataset.

- Calculate the optimal SSE value for your example.
- In general, how should datasets geometrically look like so that we have more than one optimal solution?
- What defines the number of optimal solutions?
- This problem provides an example of situations where k-means does not necessarily con- verge to the same optimal all the time.

3. KDD Cup 2009

A very popular intrusion detection dataset is the KDD Cup 2009 dataset. The dataset was collected at MIT Lincoln labs for 1998 DARPA Intrusion Detection Evaluation Program. Read about the dataset and its features (that describe network traffic) here:

http://kdd.ics.uci.edu/databases/kddcup99/task.html

The class attribute value is the network attack type associated with the instance. In this homework, your task is to perform intrusion detection using classification. You are going to use the dataset that is uploaded in ARFF format with this homework and Weka to perform the following:

a) Download the dataset kddcup99.zip here.
b) The dataset has around 500 thousand records. (1) Randomize your dataset and (2) take a 10% sample of your dataset. Save your sample ("Save" from the "Preprocess" tab). For this problem to be graded, we need this sample ARFF file, so please submit it with assignment.
c) Classify your sample using Naive Bayes, Decision Tree Learning (J48 in Weka), and K-NN (IBk) in Weka. Classify using 10-fold cross validation. Use default parameters for all, except for IBk (use k=10). Save result buffers (right click on the classifier name in "Result list") and submit your three result buffers.

4. Text Clustering (Takes Time!)

Download the fine foods dataset from:

http://snap.stanford.edu/data/web-FineFoods.html

Perform the following:

a) Identify all the unique words that appear in the "review/text" field of the reviews. Denote the set of such words as L.
b) Remove from L all stop words in "Long Stop word List" from https://www.ranks.nl/stopwords. Denote the cleaned set as W.

c) Count the number of times each word in *W* appears among all reviews ("review/text" field) and identify the top 500 words.
d) Vectorize all reviews ("review/text" field) using these 500 words.
e) Cluster the vectorized reviews into 10 clusters using k-means. You are allowed to use any program or code for k-means (Weka has k-means too). This will give you 10 centroid vectors.
f) From each centroid, select the top 5 words that represent the centroid (i.e., the words with the highest feature values)

Submit the following:

1. Top 500 words + counts for these words.
2. The top 5 words representing each cluster and their feature values (50 words + 50 values).
3. **IMPORTANT:** your code and a step-by-step readme to help reproduce your results. I should be able to get the same results by running your code and by following your readme for this problem to get graded.