CS 6350 – Big Data Analytics and Management

Spring 2019

Assignment# 4

Due: April 29 by 11.59 p.m.

Part I: Clustering

Question A (K-means algorithm)

Consider the following eight points in 2-dimensional space: (2,10); (2,5); (8,4); (5,9); (7,5); (6,4); (1,2); (4,9); (10,10). Suppose we plan to use the Euclidean distance metric and that we are interested in clustering these points into 3 clusters.

(i).Plot the data to see what might be appropriate clusters.

(ii) Beginning with the points (2,5), (5,8) and (4,9) as initial cluster centers, form the three initial clusters.

(iii) Use the k-means clustering algorithm to get the final three clusters. What are the resulting centers and resulting clusters? (Here K = 3)

Question B

(i). Use the similarity matrix in Table below to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

	P1	P2	P3	P4	P5
P1	1.00	0.10	0.41	0.55	0.35
P2	0.10	1.00	0.64	0.47	0.98
P3	0.41	0.64	1.00	0.44	0.85
P4	0.55	0.47	0.44	1.00	0.76
P5	0.35	0.98	0.85	0.76	1.00

Question C.

(ii) Hierarchical clustering is sometimes used to generate K clusters, K > 1 by taking the clusters at the Kth level of the dendrogram. (Root is at level 1.) By looking at the clusters produced in this way, we can evaluate the behavior of hierarchical clustering on different types of data and clusters, and also compare hierarchical approaches to K-means.

The following is a set of one-dimensional points: {6, 12, 18, 24, 25, 28, 30, 42, 48}.

(a) For each of the following sets of initial centroids, create two clusters by assigning each point to the nearest centroid, and then calculate the total squared error for each set of two clusters.

Show both the clusters and the total squared error for each set of centroids.

1) { 5, 7.5}

2) {15, 25}

b) Do both sets of centroids represent stable solutions, i.e., if the K-means algorithm was run on this set of points using the given centroids as the starting centroids, would there be any change in the clusters generated?

c) What are the two clusters produced by MIN? (MIN is single-link clustering, also called minimum method)

d) Which technique, K-means or MIN, seems to produce the "most natural" clustering in this situation?

e) What well-known characteristic of the K-means algorithm explains the previous behavior?

Part II: Classification

Question D (Decision Tree).

Given the following training data and attributes with their respective possible values: X1 ∈ {a,b}, X2 ∈ {c,g,u,w}, X3 ∈ {k,s,v}. The class is Y ∈ {+1,-1}.

x1	x2	x3	y
a	c	k	-1
a	w	k	-1
b	w	v	+1
a	c	v	+1
b	w	k	-1
a	c	s	+1
b	w	s	+1
a	u	v	-1
b	c	v	-1
b	c	s	+1
b	g	v	-1

Learn a decision tree using the ID3 algorithm (information gain heuristic with entropy) and draw the resulting decision tree (with all parts labeled accordingly). Please show all calculations justifying your answer.

b. Given the following training data and attributes with their respective possible values: X 1 ∈ {a,b}, X 2 ∈ N, X 3 ∈ {e,f}, X 4 ∈ {c,d}. The class is Y ∈ {+1, -1}.

X_1	X_2	X_3	X_4	Y
b	185	f	d	+1
b	180	f	c	+1
b	170	f	c	-1
b	140	e	d	-1
a	176	f	d	+1
b	179	f	d	-1

i. Draw a decision tree with only 2 attribute node 2 attribute nodes and 3 leaf class nodes that will get 100% accuracy on the training data. No need to show calculations, but please make sure all parts of the tree are labels accordingly.

ii. What is the accuracy of the following test data using your learned decision tree?

X_1	X_2	X_3	X_4	Y
b	170	f	d	-1
a	150	f	d	+1
b	60	f	d	+1

Part III: Programming

Question E (Deep Learning - CNN).

You will be working on MNIST data, a dataset of thousands of images of handwritten digits. You can download the dataset here - https://www.kaggle.com/c/digit-recognizer/data. The data files train.csv and test.csv contain gray-scale images of hand-drawn digits, from zero through nine. Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255, inclusive. The training data set, (train.csv), has 785 columns. The first column, called "label", is the digit that was drawn by the user. The rest of the columns contain the pixel-values of the associated image." You only need to download the train.csv file. test.csv is not required. The train.csv file contains 42k samples of images. To reduce time of running the program, you will only work with 10,000 randomly selected samples out of these, although make sure you have equal number of samples belonging to each label (i.e.,1,000 samples of label '0', 1,000 samples of label '1' and so on).

Apply PyTorch package to implement the CNN algorithm, and train the model using the training set generated in the last task and use your model to predict labels in the test set. The structure of CNN isdefined as follows. You have the flexibility to define the parameters of different layers (e.g., No. of nodes, filter dimension, pooling dimension etc.). Show your results on both training and testing sets.

Structure:

Convolutional layer -> Max pooling layer -> Convolutional layer - > Max pooling layer -> Fully connected layer x2 -> Softmax layer

Deliverables:

1. Python source codes in a zipped file;
2. Brief report including all your results and observations;

Question F (Recommendation Systems).

Use Collaborative filtering to find the accuracy of ALS model. Use ratings.dat file. It contains

User id :: movie id :: ratings :: timestamp.

Your program should report the accuracy of the model.

For details follow the link: https://spark.apache.org/docs/latest/mllib-collaborative-filtering.html

Please use 60% of the data for training and 40% for testing and report the MSE of the model.