IAF 604 - Machine Learning and Predictive Analytics

Exam 2

Name: Aman Tej Vidapu

Part 1: Multiple choice questions, each 5 points total 50 points (may contain more than one choice, underline all the correct answer(s))

1. The three phases of supervised learning do NOT include:

A) testing

B) modeling

C) training

D) validation

A,B,D if do include and modeling is done while training phase.

2. Parento principle uses which ratio of training vs testing data set:

A) 70:30

B) 50:50

C) 80:20

D) 90:10

3. Which of the following is NOT distance-based:

A) Entropy

B) MAE

C) Hamming distance

D) MSE

4. Which of the following about support vector machine is/are true:

A) it is one of clustering methods

B) it is an optimization method

C) it is a linear model

D) nonlinear SVM uses kernel function

5. Which of the following about decision tree is/are true:

A) the best tree model will fit the training set perfectly regardless

B) it is a recursive algorithm

C) information gain is the only measure used to evaluate the quality of the splits

D) it is a hierarchical model

6. Which of the following about random forest is incorrect:

A) it is a hierarchical model

B) random sampling is used

C) since it uses multiple trees it may be slower than a decision tree

D) because of bootstrapping, the results would be random and less reliable

7. Which of the following is/are model of deep learning:

A) belief network

B) dropout

C) drop connect

D) drop-and-drip

8. Neural network model includes which of two computing phases:

A) back propagation

B) draw three layers

C) forward propagation

D) input transformation

9. In clustering,

A) labels are known

B) it is also called supervised learning

C) a description is obtained (then able to predict)

D) a prediction is obtained

10. **One advantage** of density based method such as DBSCAN is

A) it can adapt to the shape of data

B) interpretable due to its rectangle ranges

C) it can handle high dimensionality

D) it is incremental

Part 2: short questions, 5 points each

1. **What are the main classification methods covered in the class?**

Ans:

The classification methods covered in class are:

1. Training: from the training dataset, generate a best model with parameters.

2. Validation: adjust the model before testing, stops overfitting problems.

3. Testing: confirm the adjusted model works efficient on another dataset called testing set.

Models:

1. SVM classifier

2. Decision Trees classifier

3. Random Forest classifier

4. Deep learning models to classify

5. KNN for classification

2. **What are the main methods for clustering, each give an example algorithm?**

Ans:

The main methods for clustering:

- Partitioning-based clustering
- Hierarchical clustering
  1.Agglomerative
  2.Divisive
- Density-based clustering
- Grid-based clustering

1. Partitioning-based clustering:
Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors.
Example algorithms: k-means, k-medoids, CLARANS

2. Hierarchical clustering:
Create a hierarchical decomposition of the set of data (or objects) using some criterion.
Example algorithms: Diana, Agnes, BIRCH, CAMELEON.

3. Density-based clustering:
Based on connectivity and density functions.
Example algorithms: DBSACN, OPTICS, DenClue

4. Grid-based clustering:
based on a multiple-level granularity structure.
Example algorithms: STING, WaveCluster, CLIQUE

3. What are the major techniques used in random forest, each with a brief explanation (one sentence or two)?

Ans:

Bootstrapping:

Bootstrapping is applied at the training phase of the random forest algorithm. It aids in the generation of several subsets from a set of data by selecting the same number of observations at random as the original data set, but with the replacement. This allows for some of the original data's observations to be reproduced in a subset of the data set. In random forest models, the goal of bootstrap sampling is to optimize the "class-distance" between each intermediate node and the decision tree leaves.

Bagging:

Bagging is applied at the testing phase of the random forest algorithm. The averaging of the prediction (or classification) answers provided by the bootstrap samples to achieve the final prediction (or classification) result is referred to as bagging. Bootstrap aggregation is where the word Bagging derives from. In the random forest technique, bootstrapping is used as part of the training procedure. It helps develop numerous classification models by generating various domains using simple class overlap thinning. The testing algorithm can efficiently evaluate the performance of the classifiers thanks to the numerous classification models.

4. What are the main cross-validation methods?

Ans:

K-fold cross-validation:
k-fold cross validation technique is a deterministic approach, commonly we use tenfold cross-validation is used in supervised learning. Seen data cross-validation, we use 80-20 split for training and randomly selected validation data. Which is helpful in terms of helping model to overcome over-fitting issue, but both the datasets are seen which might would not give exact better influence towards evaluation of model. Unseen data cross-validation, we use mostly 60-20-20 split for training, validation and testing data. Which overcomes the new data accuracy measure to evaluate

the model performance. In n fold cross-validation, a circular-shift algorithm on a block as folds were used to perform validation. To explain consider 10 fold cross-validation, we will divide the initial block data to 10 disjointed subsets datasets/folds. Now we will train the model using first nine folds and remaining 10th fold for testing the model. Next, the ninth fold will be taken as test data and remaining data as training data and this circular pattern is continued till 10 times in general n-times. Here, we say the initial (9th fold) is seen by training model in first step but, this is still unseen data for further training steps. For better understanding of mentioned process please refer[1] Page No.194 Fig. 8.6. As a result, we will have ten classification accuracies.

Leave-P-Out cross validation:
From the total number of data points in the dataset, we subtract p points (say n). We use these (n – p) data points to train the model, and we use p data points to test it. This technique is repeated for all possible p combinations in the original dataset. Then we average the accuracies from all of these iterations to get the final accuracy. Because we train the model on every possible combination of data points, this is an exhaustive method. As we increase the amount of p, the number of possible combinations increases, and we can say that the approach becomes much more exhaustive.

Leave-one-out cross validation:

The value of p is set to one in this simple or varaition form of Leave-P-Out cross validation. This makes the approach much less exhaustive, as we now have n number of choices for n data points and p = 1.

Random sub-sampling

Pareto principle: 80:20 ratio of training vs testing to validate the model

Part 3: K-means clustering, 30 points

From the following 12 points: (0,1), (0,2), (0,3), (1, 0), (2, 0), (3,0), (3,3), (4,4), (5,5) three points (1,0), (2,0), (3,3) are chosen as the initial centers of 3 clusters. Use k-means algorithm to group these points into three clusters. In each iteration, compute the new means of the centers and which points are in these clusters by drawing diagrams (circle these points if they belong to the same cluster and put an x on the centers).

Ans:

The K-means method searches an unlabeled multidimensional dataset for a predetermined number of clusters, concluding with a simple interpretation of how an optimum cluster can be expressed.

The concept would be divided into two parts:

The cluster center, for starters, is the arithmetic mean (AM) of all the data points in the cluster.

Second, in comparison to other cluster centers, each point is next to its cluster center. The k-means clustering model is built on these two views.

Part-3

Distance formula

Euclidean distance :

$(x_1, x_2) \quad (y_1, y_2)$

$$d((x_1, x_2), (y_1, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Given points

$A_1(0,1)$, $A_2(0,2)$, $A_3(0,3)$, $A_4(1,0)$,

$A_5(2,0)$, $A_6(3,0)$, $A_7(3,3)$, $A_8(4,4)$,

$A_9(5,5)$.

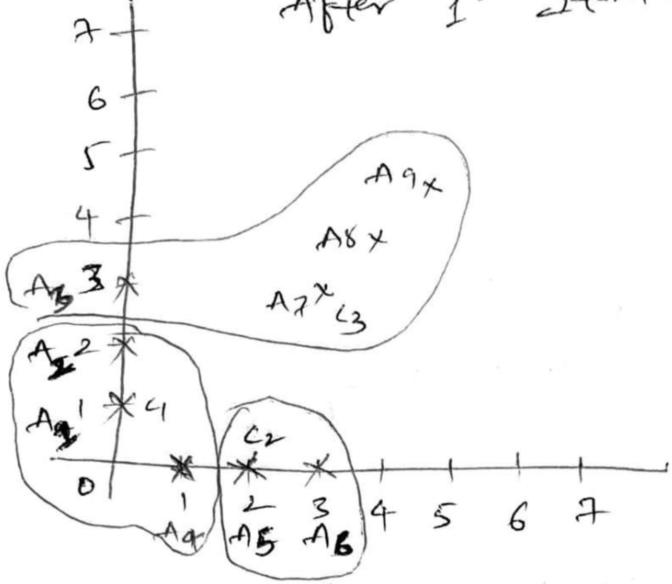
Initial center of each cluster:

$C_1(1,0)$, $C_2(2,0)$; $C_3(3,3)$.

calculating distances from each center & assigning clusters.

Point	$C_1(1,0)$	$C_2(2,0)$	$C_3(3,3)$	cluster
$A_1(0,1)$	1.414	2.236	3.605	$C_1(1,0)$
$A_2(0,2)$	2.236	2.828	3.162	$C_1(1,0)$
$A_3(0,3)$	3.162	3.605	3	$C_3(3,3)$
$A_4(1,0)$	0	1	3.605	$C_1(1,0)$
$A_5(2,0)$	1	0	3.162	$C_2(2,0)$
$A_6(3,0)$	2	1	3	$C_2(2,0)$
$A_7(3,3)$	3.605	3.162	0	$C_3(3,3)$
$A_8(4,4)$	5	4.472	1.412	$C_3(3,3)$
$A_9(5,5)$	6.403	5.830	2.828	$C_3(3,3)$

After 1st Iteration



The three clusters after first iteration

cluster 1 :- {A1, A2, A4}

cluster 2 :- {A5, A6}

cluster 3 :- {A3, A7, A8, A9}

For next iteration update the centriods of the new clusters formed.

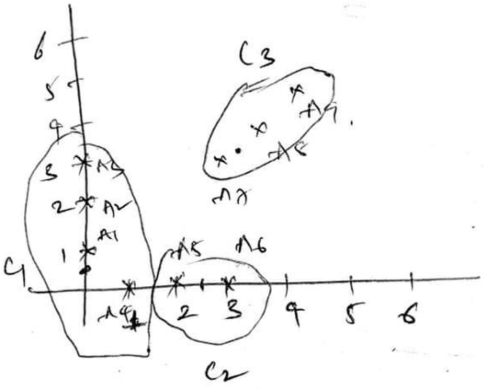
center 1 = $((0+1+0)/3, (1+0+2)/3) = (0.33, 1)$

center 2 = $((2+3)/2, (0+0)/2) = (2.5, 0)$

center 3 = $((0+3+4+5))/4, (3+3+4+5)/4) = (3, 3.75$

- Assign clusters to points by calculating distance with new clusters.

	c1	c2	c3	cluster
A1	0.33	2.69	4.06	c1
A2	1.05	3.20	3.47	c1
A3	2.02	3.90	3.09	c1
A4	1.20	1.5	4.25	c1
A5	1.94	0.5	3.88	c2
A6	2.85	0.5	3.75	c2
A7	3.33	3.04	0.75	c3
A8	4.74	4.27	1.03	c3
A9	6.14	5.59	2.35	c3



After 2nd iteration, clusters are:-

cluster 1:- $\{A_1, A_2, A_3, A_4\}$

Cluster 2:- $\{A_5, A_6\}$

Cluster 3:- $\{A_7, A_8, A_9\}$

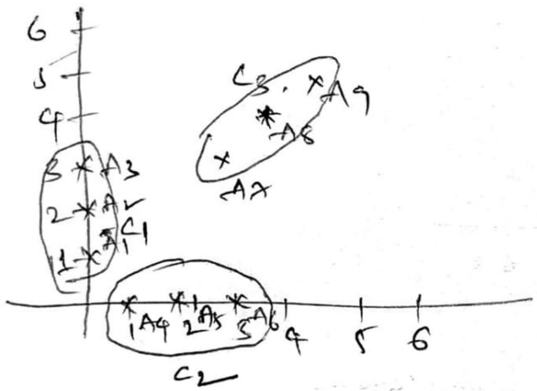
For Next iteration update centroids,

Center1 :- $((0+0+0+1)/4,\ (1+2+3+0)/4) = (0.25, 1.5)$

center2:- $((2+3)/2,\ (0+0)/2) = (2.5, 0)$

Center3:- $((3+4+5)/3,\ (3+4+5)/3) = (4, 4)$

	C1	C2	C3	cluster.
A1	0.55	2.69	5.0	C1
A2	0.55	3.90	4.47	C1
A3	1.52	3.90	4.12	C1
A4	1.67	1.5	5.0	C2
A5	2.30	0.5	4.47	C2
A6	3.18	0.5	4.12	C2
A7	3.13	3.04	1.41	C3
A8	4.50	4.27	0	C3
A9	5.90	5.59	1.41	C3



After iteration 3, clusters are:

C1 = {A1, A2, A3}

C2 = {A4, A5, A6}

C3 = {A7, A8, A9}.

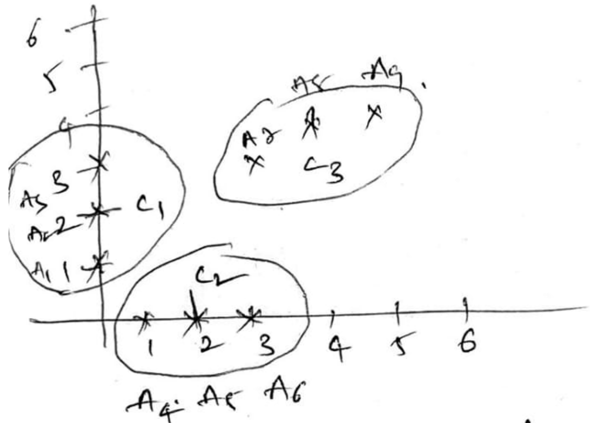
calculating mediods for new clusters.

center 1 :- {(0+0+0)/3, (1+2+3)/3} = (0, 2)

center 2 :- {(1+2+3)/3, (0+0+0)/3} = (2, 0)

center 3 :- (4, 4)

	C_1	C_2	C_3	cluster
A_1	1.0	2.23	5.0	C_1
A_2	0.0	2.82	4.47	C_1
A_3	1.0	3.60	4.12	C_1
A_4	2.23	1.0	5.0	C_2
A_5	2.82	0.0	4.47	C_2
A_6	3.60	1.0	4.12	C_2
A_7	3.16	3.16	1.41	C_3
A_8	4.47	4.47	0.0	C_3
A_9	5.83	5.83	1.41	C_3 .



After iterat 4, cluster are unchanged.

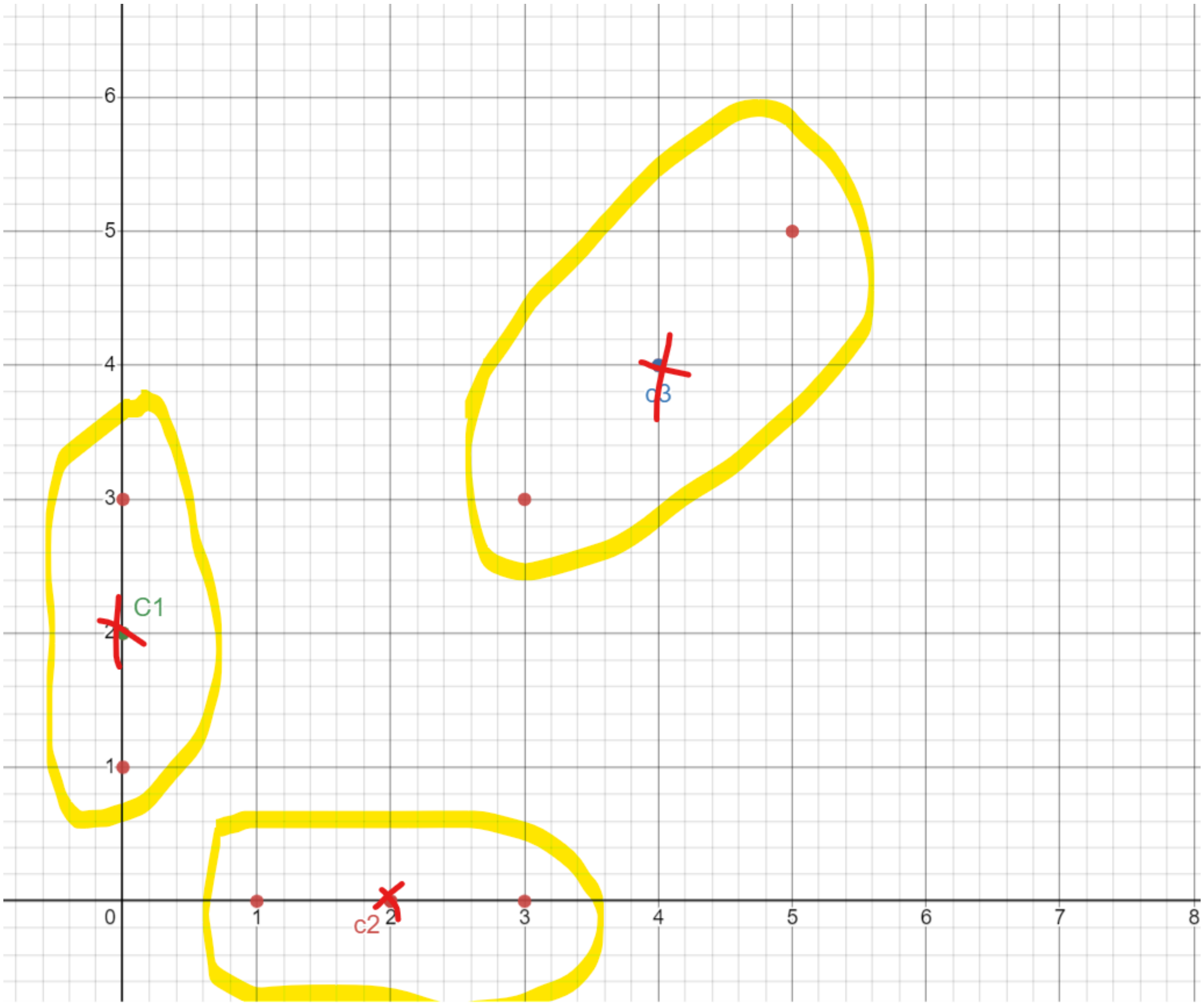
So, we stop K-means Algorithm here itself.

And, Final clusters are:-

cluster 1 :- $\{A_1, A_2, A_3\}$
Cluster 2 :- $\{A_4, A_5, A_6\}$
Cluster 3 :- $\{A_7, A_8, A_9\}$

With centers as :- $C_1 (0, 2)$
$C_2 (2, 0)$
$C_3 (4, 4)$

Final clusters:



References:

[1] Book: Machine Learning Models and Algorithms for Big Data Classification, Author: Shan Suthahara