CSCI-UA 472 Artificial Intelligence

Muhammad Wajahat Mirza

mwm356@nyu.edu

Homework 07

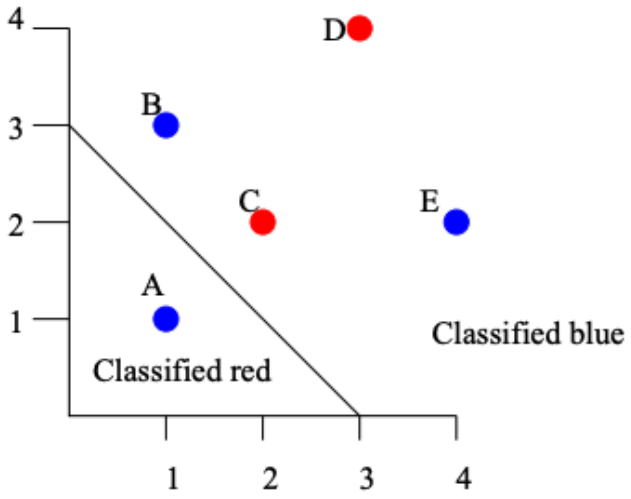
November 19, 2020

Problem 1

Suppose that you have the following collection **T** of data points in two dimensions:

x	1	1	2	3	4
y	1	3	2	4	2
	B	B	R	R	B

The picture below illustrates these points, together with the classifier
IF $x + y - 3 > 0$ **then** BLUE **else** RED,
corresponding to the weight vector $\langle 1, 1, 3 \rangle$



A. Compute the value of the error function for this classifier:

$$E_T(\overrightarrow{w}) = \sum_{p \in T, \mathbf{p} \text{ misclassified}} |w_1 \mathbf{p_x} + w_2 \mathbf{p_y} - w_3|$$

B. Compute the gradient of the error function with respect to the weight vector $\overrightarrow{w}$.
The gradient is a vector $\overrightarrow{\nabla}|_{\overrightarrow{w}} = \langle g_1, g_2, g_3 \rangle$ computed as follows: For a given weight vector $\overrightarrow{w}$ and data point $p \in T$ let

$$s_{\overrightarrow{w}}(\mathbf{p}) = \begin{cases} 1 & \text{if } \mathbf{p} \text{ is labelled RED in T but is classified BLUE by } \overrightarrow{w} \\ -1 & \text{if } \mathbf{p} \text{ is labelled BLUE in T but is classified RED by } \overrightarrow{w} \\ 0 & \text{if } \mathbf{p} \text{ is correctly classified by } \overrightarrow{w} \end{cases}$$

Then

$$\overrightarrow{\nabla} E|_{\overrightarrow{w}} = \sum_{p \in T} s_{\overrightarrow{w}}(\mathbf{p}) \cdot \langle \mathbf{p}_x, \mathbf{p}_y, -1 \rangle$$

C. Compute the new weight vector after one step of gradient descent:
$\overrightarrow{w}' = \overrightarrow{w} - \delta \cdot \overrightarrow{\nabla} E|_{\overrightarrow{w}}$ where $\delta = 0.1$

D. How does the new weight vector $\overrightarrow{w}'$ classify the points?
E. What is the value of the error function at the new weight vector $\overrightarrow{w}'$?

Solution to Problem 1

A. Solution

First, let's identity points that are misclassified with the given linear classifier. Point **A** is misclassified as BLUE while points **C, D** are misclassified as RED. Thus, to calculate $E_T(\overrightarrow{w})$, use the given weight vector and **x, y** coordinates of misclassified points. Thus, we get, ($p \in T, \mathbf{p}$ misclassified) as $p \in \{A, C, D\}$

$$E_T(\overrightarrow{w}) = \sum_{p \in \{A,C,D\}} |w_1 \mathbf{p_x} + w_2 \mathbf{p_y} - w_3|$$

$$E_T(\overrightarrow{w}) = |1 \cdot 1 + 1 \cdot 1 - 3| + |1 \cdot 2 + 1 \cdot 2 - 3| + |1 \cdot 3 + 1 \cdot 4 - 3|$$

$$E_T(\overrightarrow{w} = |-1| + |1| + |4| = 1 + 1 + 4 = 6$$

Therefore, the error value, $E_T(\overrightarrow{w})$, for the given linear classifier is 6.

B. Solution

To compute the gradient of the error function, let's find the value of $s_{\overrightarrow{w}}(\mathbf{p})$ where $p \in T$.

$$\begin{cases} \mathrm{s}_{\overrightarrow{w}}(\mathbf{A}) = -1 & \text{Because } \mathbf{A} \text{ is labelled RED in T but is classified BLUE by } \overrightarrow{w} \\ \mathrm{s}_{\overrightarrow{w}}(\mathbf{B}) = 0 & \text{Because } \mathbf{B} \text{ is correctly classified by } \overrightarrow{w} \\ \mathrm{s}_{\overrightarrow{w}}(\mathbf{C}) = 1 & \text{Because } \mathbf{C} \text{ is labelled BLUE in T but is classified RED by } \overrightarrow{w} \\ \mathrm{s}_{\overrightarrow{w}}(\mathbf{D}) = 1 & \text{Because } \mathbf{D} \text{ is labelled BLUE in T but is classified RED by } \overrightarrow{w} \\ \mathrm{s}_{\overrightarrow{w}}(\mathbf{E}) = 0 & \text{Because } \mathbf{E} \text{ is correctly classified by } \overrightarrow{w} \end{cases}$$

Using $s_{\overrightarrow{w}}(\mathbf{p})$ values, compute gradient of the error function using

$$\overrightarrow{\nabla} E|_{\overrightarrow{w}} = \sum_{p \in T} s_{\overrightarrow{w}}(\mathbf{p}) \cdot \langle \mathbf{p}_x, \mathbf{p}_y, -1 \rangle$$

$$\begin{aligned} \overrightarrow{\nabla} E|_{\overrightarrow{w}} = s_{\overrightarrow{w}}(\mathbf{A}) \cdot \langle \mathbf{p}_A, \mathbf{p}_A, -1 \rangle + s_{\overrightarrow{w}}(\mathbf{B}) \cdot \langle \mathbf{p}_B, \mathbf{p}_B, -1 \rangle + s_{\overrightarrow{w}}(\mathbf{C}) \cdot \langle \mathbf{p}_C, \mathbf{p}_C, -1 \rangle \\ + s_{\overrightarrow{w}}(\mathbf{D}) \cdot \langle \mathbf{p}_D, \mathbf{p}_D, -1 \rangle + s_{\overrightarrow{w}}(\mathbf{E}) \cdot \langle \mathbf{p}_E, \mathbf{p}_E, -1 \rangle \end{aligned}$$

Use the values from $s_{\overrightarrow{w}}(\mathbf{p})$

$$\overrightarrow{\nabla} E|_{\overrightarrow{w}} = -1 \cdot \langle 1, 1, -1 \rangle + 0 \cdot \langle 1, 3, -1 \rangle + 1 \cdot \langle 2, 2, -1 \rangle + 1 \cdot \langle 3, 4, -1 \rangle + 0 \cdot \langle 4, 2, -1 \rangle$$

$$\overrightarrow{\nabla} E|_{\overrightarrow{w}} = \langle -1, -1, 1 \rangle + \langle 0, 0, 0 \rangle + \langle 2, 2, -1 \rangle + \langle 3, 4, -1 \rangle + \langle 0, 0, 0 \rangle$$

$$\overrightarrow{\nabla} E|_{\overrightarrow{w}} = \langle 4, 5, -1 \rangle$$

Therefore, the gradient of the error function with respect to the weight vector $\overrightarrow{w}$ is $\langle 4, 5, -1 \rangle$.

C. Solution

With the calculated gradient, $\overrightarrow{\nabla} E|_{\overrightarrow{w}}$, use the following given equation to find the new weight vector.

$$\overrightarrow{w}' = \overrightarrow{w} - \delta \cdot \overrightarrow{\nabla} E|_{\overrightarrow{w}}$$

where $\delta = 0.1$.

$$\begin{aligned} \overrightarrow{w}' &= \langle 1, 1, 3 \rangle - 0.1 \cdot \langle 4, 5, -1 \rangle \\ \overrightarrow{w}' &= \langle 1, 1, 3 \rangle - \langle 0.4, 0.5, -0.1 \rangle \\ \overrightarrow{w}' &= \langle 0.6, 0.5, 3.1 \rangle \end{aligned}$$

Hence, new vector weight, $\overrightarrow{w}'$, with respect to calculated gradient is $\langle 0.6, 0.5, 3.1 \rangle$

D. Solution

To classify points using the new weight vector, use the following equation:
$w_1 \mathbf{p_x} + w_2 \mathbf{p_y} - w_3$. If < 0, then RED. Else, BLUE.

$$\begin{cases} A = 0.6 \cdot 1 + 0.5 \cdot 1 - 3.1 = -2 & \text{A is classified as RED because} < 0 \\ B = 0.6 \cdot 1 + 0.5 \cdot 3 - 3.1 = -1 & \text{B is classified as RED because} < 0 \\ C = 0.6 \cdot 2 + 0.5 \cdot 2 - 3.1 = -0.9 & \text{C is classified as RED because} < 0 \\ D = 0.6 \cdot 3 + 0.5 \cdot 4 - 3.1 = 0.7 & \text{D is classified as BLUE because} > 0 \\ E = 0.6 \cdot 4 + 0.5 \cdot 2 - 3.1 = 0.3 & \text{E is classified as BLUE because} > 0 \end{cases}$$

Thus, points **A, B, C** are classified as RED whereas points **D, E** are classified as BLUE.

E. Solution

Again, let's identity points that are misclassified with the given linear classifier and new weight vector. Points **A, B** are misclassified as RED while point **D** is misclassified as BLUE. Thus, to calculate $E_T(\overrightarrow{w})$, use the given weight vector and **x, y** coordinates of misclassified points. Thus, we get, ($p \in T, \mathbf{p}$ misclassified) as $p \in \{A, B, D\}$

$$E_T(\overrightarrow{w}) = \sum_{p \in \{A,B,D\}} |w_1 \mathbf{p_x} + w_2 \mathbf{p_y} - w_3|$$

$$E_T(\overrightarrow{w}) = |0.6 \cdot 1 + 0.5 \cdot 1 - 3.1| + |0.6 \cdot 1 + 0.5 \cdot 3 - 3.1| + |0.6 \cdot 3 + 0.5 \cdot 4 - 3.1|$$

$$E_T(\overrightarrow{w} = |-2| + |-1| + |0.7| = 2 + 1 + 0.7 = 3.7$$

Therefore, the new error value, $E_T(\overrightarrow{w})$, for the given linear classifier is 3.7 which is much smaller than 6. Improvement shown!

Problem 2: Precision and Recall

Suppose that a classifier computes a numeric score to an item based on the classifier's "confidence" that the item is a member of the target category. In using the classifier, you set a threshhold, accept the items whose score is higher than the threshhold, and reject items whose score is lower.

For instance, suppose you have the following training set and outputs:

Name	a	b	c	d	e	f	g	h	i	j
Label	T	T	F	T	F	T	T	F	F	F
Score	.95	.92	.85	.84	.81	.75	.71	.69	.62	.56

Name	k	l	m	n	o	p	q	r	s	t
Label	F	F	T	F	T	F	F	F	T	F
Score	.51	.48	.43	.42	.32	.25	.21	.15	.08	.01

If you then set the threshhold at 0.50, the classifier will accept items a-k and reject items l-t.
Compute precision, recall, and F-score for the following threshholds: 0.9, 0.6, 0.4, 0.1.

Solution to Problem 2

Using the training set and outputs given above, member count for target can be done. Each True assigned with given confidence value is a member of the target category, thus, $C_T = 8$ in the data set. The following Table summarizes the calculations done.

Threshold	0.9	0.6	0.4	0.1
Precision	1	0.5556	0.4286	0.3889
Recall	0.25	0.625	0.75	0.875
F-Score	0.4	0.5883	0.5456	0.5385

- **Threshold 0.9**:
  - The classifier only accepts items with score > 0.9. Therefore, items **a-b** are accepted and items **c-t** are rejected.
  - True Quantity items, $Q_T = 2$, for this threshold are accepted.
  - $Q_T \cap C_T$ gives True target category members. $Q_T \cap C_T = 2$

$$Precision \rightarrow P = \frac{C_T \cap Q_T}{Q_T} = \frac{2}{2} = 1$$

$$Recall \rightarrow R = \frac{C_T \cap Q_T}{C_T} = \frac{2}{8} = 0.25$$

$$F - Score \rightarrow F = 2 \times \frac{P \cdot R}{P + R} = 2 \times \frac{1 \cdot 0.25}{1 + 0.25} = 2 \times \frac{0.25}{1.25} = 0.4$$

- **Threshold 0.6**:
  - The classifier only accepts items with score > 0.6. Therefore, items **a-i** are accepted and items **j-t** are rejected.
  - True Quantity items, $Q_T = 9$, for this threshold are accepted.
  - $Q_T \cap C_T$ gives True target category members. $Q_T \cap C_T = 5$

$$Precision \rightarrow P = \frac{C_T \cap Q_T}{Q_T} = \frac{5}{9} = 0.5556$$

$$Recall \rightarrow R = \frac{C_T \cap Q_T}{C_T} = \frac{5}{8} = 0.625$$

$$F - Score \rightarrow F = 2 \times \frac{P \cdot R}{P + R} = 2 \times \frac{0.5556 \cdot 0.625}{0.5556 + 0.625} = 2 \times \frac{0.3473}{1.1806} = 0.5883$$

- **Threshold 0.4**:

- The classifier only accepts items with score > 0.4. Therefore, items **a-n** are accepted and items **o-t** are rejected.
- True Quantity items, $Q_T = 14$, for this threshold are accepted.
- $Q_T \cap C_T$ gives True target category members. $Q_T \cap C_T = 6$

$$Precision \rightarrow P = \frac{C_T \cap Q_T}{Q_T} = \frac{6}{14} = 0.4286$$

$$Recall \rightarrow R = \frac{C_T \cap Q_T}{C_T} = \frac{6}{8} = 0.75$$

$$F - Score \rightarrow F = 2 \times \frac{P \cdot R}{P + R} = 2 \times \frac{0.4286 \cdot 0.75}{0.4286 + 0.75} = 2 \times \frac{0.3215}{1.1786} = 0.5456$$

- **Threshold 0.1**:
  - The classifier only accepts items with score > 0.1. Therefore, items **a-r** are accepted and items **s-t** are rejected.
  - True Quantity items, $Q_T = 18$, for this threshold are accepted.
  - $Q_T \cap C_T$ gives True target category members. $Q_T \cap C_T = 7$

$$Precision \rightarrow P = \frac{C_T \cap Q_T}{Q_T} = \frac{7}{18} = 0.3889$$

$$Recall \rightarrow R = \frac{C_T \cap Q_T}{C_T} = \frac{7}{8} = 0.875$$

$$F - Score \rightarrow F = 2 \times \frac{P \cdot R}{P + R} = 2 \times \frac{0.3889 \cdot 0.875}{0.3889 + 0.875} = 2 \times \frac{0.3403}{1.2639} = 0.5385$$

Problem 3: K-means

Consider the following collection of data points in two dimensions:

	A	B	C	D	E	F	G	H	I	J
x	1	1002	498	6	510	503	4	1010	1006	502
y	6	20	651	10	622	632	9	25	30	680

Trace the behavior of the k-means algorithms, with $k = 3$, starting from the centers $\langle 500, 10 \rangle$, $\langle 200, 700 \rangle$, $\langle 800, 200 \rangle$. (Use floating point values for the center points.) Your trace should show the alternating computing center points, and assignments of points to cluster.

At each stage of the algorithm — that is, after the center points have been recomputed and after the new cluster assignments have been recomputed — compute the value of the cost function:

$$Cost(C) = \sum_{p \in S} D^2(p, C(p)$$

In the above formula:

- S is the set of all the points.
- C is the mapping from data points to the associated center point.
- D is the Euclidean distance, and D^2 is the square of the Euclidean distance.
- Thus, $\mathrm{D}^2(\langle p_x, p_y \rangle, \langle q_x, q_y \rangle) = (p_x - q_x)^2 + (p_y - q_y)^2$

Solution to Problem 3

To compute $Cost(C)$, first assign each data point to one of the three given cluster centers, $\langle 500, 10 \rangle$, $\langle 200, 700 \rangle$, $\langle 800, 200 \rangle$. To assign each point to a cluster center, calculate Euclidean distance, D^2, of each point with each center and pick the closest center to assign the point to it.

In the following table, each row represents calculated squared Euclidean distance of a data point to each of the given cluster centers. The last column of each row shows which cluster data point has been assigned based on the shortest distance.

Data Point	D^2 from Center 1	D^2 from Center 2	D^2 from Center 3	Assign point to
A	249017	521237	676037	**cluster 1**
B	252104	1105604	73204	**cluster 3**
C	410885	91205	294605	**cluster 2**
D	244036	513736	666536	**cluster 1**
E	374644	102184	262184	**cluster 2**
F	386893	96433	274833	**cluster 2**
G	246017	515897	670097	**cluster 1**
H	260325	1111725	74725	**cluster 3**
I	256436	1098536	71336	**cluster 3**
J	448904	91604	319204	**cluster 2**

Hence, using the last column from the table given above, it can seen that clusters with data points are formed as follows:

- Cluster 1: **{A, D, G}**
- Cluster 2: **{C, E, F, J}**
- Cluster 3: **{B, H, I}**

Using these clusters, compute the value of the cost function:

$$Cost(C) = \sum_{p \in S} D^2(p, C(p)$$

We have already calculated $D^2(p, C(p)$ as shown in the above table. Thus, use the compu-

tations from the table to find $Cost(C)$.

$Cost(C) = 249017+73204+91205+244036+102184+96433+246017+74725+71336+91604$

$$Cost(C) = 1339761$$

With the given cluster assignments, value of the cost function is 1339761.

Use these data point cluster assignments to find new cluster centers, x_c, y_c and repeat the process.

$$x_c = \frac{\sum_{p \in \text{cluster}} x_p}{\text{total points in cluster}}$$

$$y_c = \frac{\sum_{p \in \text{cluster}} y_p}{\text{total points in cluster}}$$

Cluster 1: {A, D, G}

$$x_1 = \frac{x_A + x_D + x_G}{3} = \frac{1+6+4}{3} = \frac{11}{3} = 3.667$$

$$y_1 = \frac{y_A + y_D + y_G}{3} = \frac{6+10+9}{3} = \frac{25}{3} = 8.333$$

Cluster 2: {C, E, F, J}

$$x_2 = \frac{x_C + x_E + x_F + x_J}{4} = \frac{498+510+503+502}{4} = \frac{2013}{4} = 503.25$$

$$y_2 = \frac{y_C + y_E + y_F + y_J}{4} = \frac{651+622+632+680}{4} = \frac{2585}{4} = 646.25$$

Cluster 3: {B, H, I}

$$x_3 = \frac{x_B + x_H + x_I}{3} = \frac{1002+1010+1006}{3} = \frac{3018}{3} = 1006$$

$$y_3 = \frac{y_B + y_H + y_I}{3} = \frac{20+25+30}{3} = \frac{75}{4} = 25$$

Thus, the new centers are $\langle 3.667, 8.333 \rangle$, $\langle 500.25, 646.25 \rangle$, $\langle 1006, 25 \rangle$

In the following table, new cluster centers are used to do computations.

Data Point	D^2 from Center 1	D^2 from Center 2	D^2 from Center 3	Assign point to
A	12.556	662175.125	1010386	**cluster 1**
B	996805.556	640940.625	41	**cluster 3**
C	657385.889	50.125	649940	**cluster 2**
D	8.222	652071.625	1000225	**cluster 1**
E	632960.222	633.625	602425	**cluster 2**
F	638293.889	203.125	621458	**cluster 2**
G	0.556	655338.125	1004260	**cluster 1**
H	1012984.556	642747.125	16	**cluster 3**
I	1005141.556	632521.625	25	**cluster 3**
J	699472.222	1040.625	683041	**cluster 2**

Hence, using the last column from the table given above, it can seen that clusters with data points are formed as follows:

- Cluster 1: {**A, D, G**}
- Cluster 2: {**C, E, F, J**}
- Cluster 3: {**B, H, I**}

The data point assignment has not changed for any of the cluster, thus, algorithm stops with this cluster formation. Using these clusters, compute the value of the cost function:

$$Cost(C) = \sum_{p \in S} D^2(p, C(p)$$

We have already calculated $D^2(p, C(p)$ as shown in the above table. Thus, use the computations from the table to find $Cost(C)$.

$$Cost(C) = 12.556 + 41 + 50.125 + 8.222 + 633.625 + 203.125 + 0.556 + 16 + 25 + 1140.625$$

$$Cost(C) = 2130.834$$

With the given cluster assignments, value of the cost function is 2130.834 which is significantly smaller than the previous cost calculated using old cluster centers.

Problem 4: Agglomerative clustering

Show the tree of clusters generated by the agglomerative clustering algorithm applied to the data in problem 3, assuming that the distance between two clusters C1 and C2 is the maximal distance from a point in C1 to a point in C2.

Solution to Problem 4

For the tree of clusters, at the beginning, it has each data point as separate cluster. Thus, there are ten clusters in the start:

$$\{\{A\},\{B\},\{C\},\{D\},\{E\},\{F\},\{G\},\{H\},\{I\},\{J\}\}$$

Now, by doing calculations, we find the minimal distance between two clusters C1 and C2 that is the maximal distance from a point in C1 to a point in C2. Repeat this until we get one cluster with all data points.

- **Max. Distance = 5**. The minimal distance between clusters {D} and {G} that is the maximal distance between points **D** and **G**.

$$\{\{A\},\{B\},\{C\},\{D,G\},\{E\},\{F\},\{H\},\{I\},\{J\}\}$$

- **Max. Distance = 41**. The minimal distance between clusters {A} and {D, G} that is the maximal distance between points **A** and **D**.

$$\{\{A,D,G\},\{B\},\{C\},\{E\},\{F\},\{H\},\{I\},\{J\}\}$$

- **Max. Distance = 41**. The minimal distance between clusters {H} and {I} that is the maximal distance between points **H** and **I**.

$$\{\{A,D,G\},\{B\},\{C\},\{E\},\{F\},\{H,I\},\{J\}\}$$

- **Max. Distance = 116**. The minimal distance between clusters {B} and {H, I} that is the maximal distance between points **B** and **I**.

$$\{\{A,D,G\},\{B,H,I\},\{C\},\{E\},\{F\},\{J\}\}$$

- **Max. Distance = 149**. The minimal distance between clusters {E} and {F} that is the maximal distance between points **E** and **F**.

$$\{\{A,D,G\},\{B,H,I\},\{C\},\{E,F\},\{J\}\}$$

- **Max. Distance = 857**. The minimal distance between clusters {C} and {J} that is the maximal distance between points **C** and **J**.

$$\{\{A,D,G\},\{B,H,I\},\{C,J\},\{E,F\}\}$$

- **Max. Distance = 3428**. The minimal distance between clusters {C,J} and {E,F} that is the maximal distance between points **J** and **E**.

$$\{\{A,D,G\},\{B,H,I\},\{C,J,E,F\}\}$$

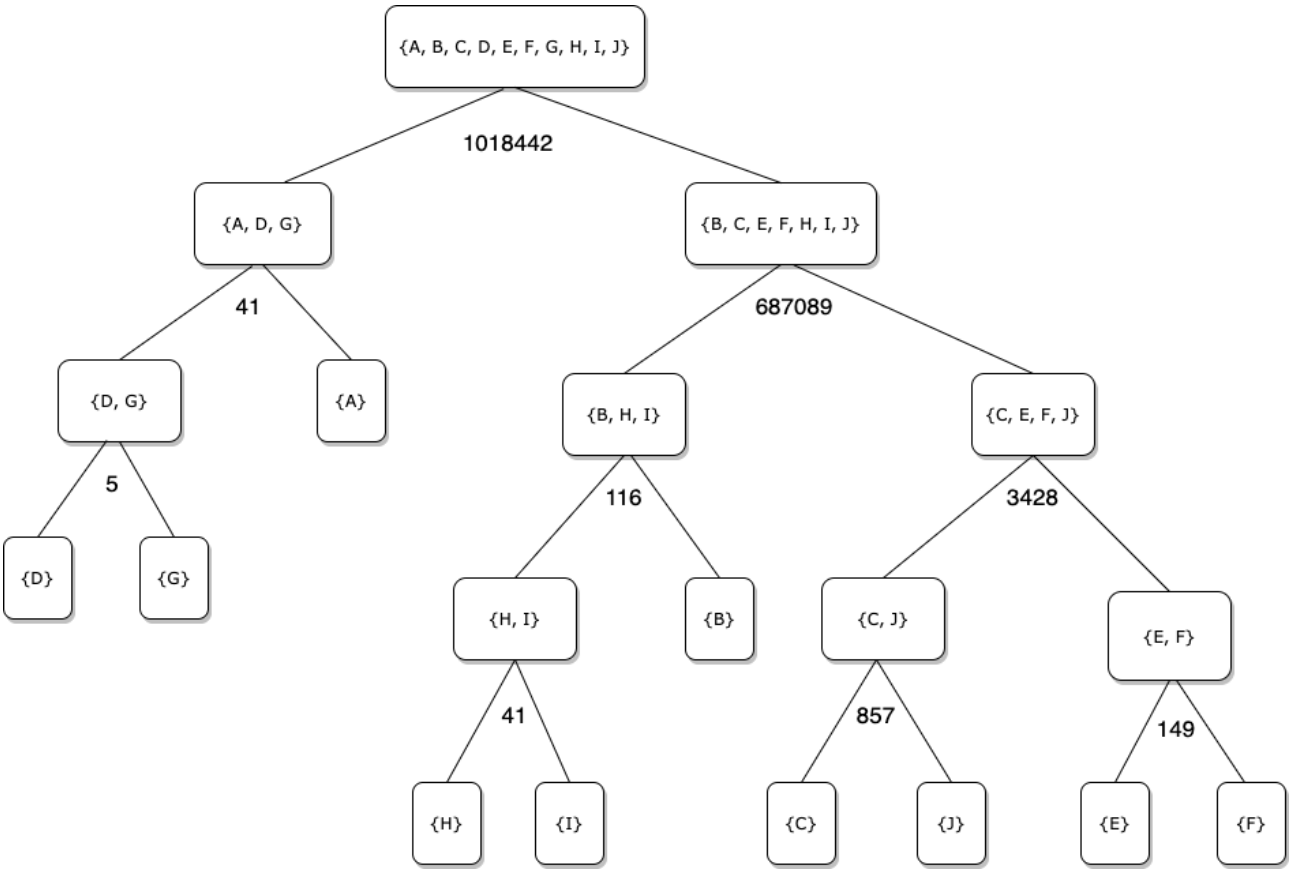
- **Max. Distance = 687089**. The minimal distance between clusters {C,E,F,J} and {B,H,I} that is the maximal distance between points **J** and **H**.

$$\{\{A, D, G\}, \{B, H, I, C, J, E, F\}\}$$

- **Max. Distance = 1018442**. The minimal distance between clusters {A,D,G} and {B, H, I, C, J, E, F} that is the maximal distance between points **A** and **H**.

$$\{A, B, C, D, E, F, G, H, I, J\}$$

From 10 clusters, we now have one giant cluster. By using the "Max Distance" as well as gradual clustering, we can show the tree generated by the agglomerative clustering algorithm.



End of Assignment. Thank you!