Assignment 4

Question 1: See hw4q1.pdf.
Question 2: Consider the following dataset

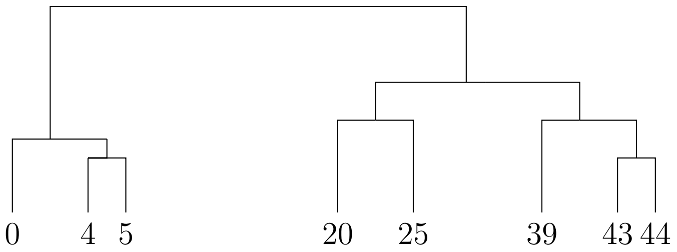
$$\{ \texttt{0, 4, 5, 20, 25, 39, 43, 44} \}$$

1. Build a dendrogram for this dataset using the **single-link, bottom-up** approach.

   Distance Matrix:

	0	4	5	20	25	39	43	44
0	0	4	5	20	25	39	43	44
4		0	1	16	21	35	39	40
5			0	15	20	34	38	39
20				0	5	19	23	24
25					0	14	18	19
39						0	4	5
43							0	1
44								0

   Dendrogram:



2. List the data points in each of the two top level clusters.
   In one of the top level clusters, there is 0, 4 and 5. In the other top level cluster, there is 20, 25, 39, 43 and 44.

Question 3: Given two clusters

$$C_1 = \{(1,1),(2,2),(3.3)\}$$
$$C_2 = \{(5,2),(6,2),(7,2),(8,2),(9,2)\}$$

compute the following values. Use the definition for scattering criteria. Note that `tr` in the scattering criterion is referring to the trace of the matrix.

1. the mean vectors m_1 and m_2

$$m_1 = \frac{1}{3}\left[\begin{bmatrix}1\\1\end{bmatrix} + \begin{bmatrix}2\\2\end{bmatrix} + \begin{bmatrix}3\\3\end{bmatrix}\right] = \frac{1}{3}\begin{bmatrix}6\\6\end{bmatrix} = \begin{bmatrix}2\\2\end{bmatrix}$$
$$m_2 = \frac{1}{5}\left[\begin{bmatrix}5\\2\end{bmatrix} + \begin{bmatrix}6\\2\end{bmatrix} + \begin{bmatrix}7\\2\end{bmatrix} + \begin{bmatrix}8\\2\end{bmatrix} + \begin{bmatrix}9\\2\end{bmatrix}\right] = \frac{1}{5}\begin{bmatrix}35\\10\end{bmatrix} = \begin{bmatrix}7\\2\end{bmatrix}$$

2. the total mean vector m

$$m = \frac{1}{8}\left[3\begin{bmatrix}2\\2\end{bmatrix} + 5\begin{bmatrix}7\\2\end{bmatrix}\right] = \frac{1}{8}\left[\begin{bmatrix}6\\6\end{bmatrix} + \begin{bmatrix}35\\10\end{bmatrix}\right] = \frac{1}{8}\begin{bmatrix}41\\16\end{bmatrix} = \begin{bmatrix}5.125\\2\end{bmatrix}$$

3. the scatter matrices S_1 and S_2

$$\begin{aligned}
S_1 &= \left(\begin{bmatrix}1\\1\end{bmatrix} - \begin{bmatrix}2\\2\end{bmatrix}\right)\left(\begin{bmatrix}1\\1\end{bmatrix} - \begin{bmatrix}2\\2\end{bmatrix}\right)^T + \left(\begin{bmatrix}2\\2\end{bmatrix} - \begin{bmatrix}2\\2\end{bmatrix}\right)\left(\begin{bmatrix}2\\2\end{bmatrix} - \begin{bmatrix}2\\2\end{bmatrix}\right)^T \\
&+ \left(\begin{bmatrix}3\\3\end{bmatrix} - \begin{bmatrix}2\\2\end{bmatrix}\right)\left(\begin{bmatrix}3\\3\end{bmatrix} - \begin{bmatrix}2\\2\end{bmatrix}\right)^T \\
&= \begin{bmatrix}-1\\-1\end{bmatrix}\begin{bmatrix}-1 & -1\end{bmatrix} + \begin{bmatrix}0\\0\end{bmatrix}\begin{bmatrix}0 & 0\end{bmatrix} + \begin{bmatrix}1\\1\end{bmatrix}\begin{bmatrix}1 & 1\end{bmatrix} \\
&= \begin{bmatrix}1 & 1\\1 & 1\end{bmatrix} + \begin{bmatrix}0 & 0\\0 & 0\end{bmatrix} + \begin{bmatrix}1 & 1\\1 & 1\end{bmatrix} = \begin{bmatrix}2 & 2\\2 & 2\end{bmatrix} \\
S_2 &= \left(\begin{bmatrix}5\\2\end{bmatrix} - \begin{bmatrix}7\\2\end{bmatrix}\right)\left(\begin{bmatrix}5\\2\end{bmatrix} - \begin{bmatrix}7\\2\end{bmatrix}\right)^T + \left(\begin{bmatrix}6\\2\end{bmatrix} - \begin{bmatrix}7\\2\end{bmatrix}\right)\left(\begin{bmatrix}6\\2\end{bmatrix} - \begin{bmatrix}7\\2\end{bmatrix}\right)^T \\
&+ \left(\begin{bmatrix}7\\2\end{bmatrix} - \begin{bmatrix}7\\2\end{bmatrix}\right)\left(\begin{bmatrix}7\\2\end{bmatrix} - \begin{bmatrix}7\\2\end{bmatrix}\right)^T + \left(\begin{bmatrix}8\\2\end{bmatrix} - \begin{bmatrix}7\\2\end{bmatrix}\right)\left(\begin{bmatrix}8\\2\end{bmatrix} - \begin{bmatrix}7\\2\end{bmatrix}\right)^T \\
&+ \left(\begin{bmatrix}9\\2\end{bmatrix} - \begin{bmatrix}7\\2\end{bmatrix}\right)\left(\begin{bmatrix}9\\2\end{bmatrix} - \begin{bmatrix}7\\2\end{bmatrix}\right)^T \\
&= \begin{bmatrix}-2\\0\end{bmatrix}\begin{bmatrix}-2 & 0\end{bmatrix} + \begin{bmatrix}-1\\0\end{bmatrix}\begin{bmatrix}-1 & 0\end{bmatrix} + \begin{bmatrix}0\\0\end{bmatrix}\begin{bmatrix}0 & 0\end{bmatrix} \\
&+ \begin{bmatrix}1\\0\end{bmatrix}\begin{bmatrix}1 & 0\end{bmatrix} + \begin{bmatrix}2\\0\end{bmatrix}\begin{bmatrix}2 & 0\end{bmatrix} \\
&= \begin{bmatrix}4 & 0\\0 & 0\end{bmatrix} + \begin{bmatrix}1 & 0\\0 & 0\end{bmatrix} + \begin{bmatrix}0 & 0\\0 & 0\end{bmatrix} + \begin{bmatrix}1 & 0\\0 & 0\end{bmatrix} + \begin{bmatrix}4 & 0\\0 & 0\end{bmatrix} \\
&= \begin{bmatrix}10 & 0\\0 & 0\end{bmatrix}
\end{aligned}$$

4. the within-cluster scatter matrix S_W

$$S_W = \begin{bmatrix}2 & 2\\2 & 2\end{bmatrix} + \begin{bmatrix}10 & 0\\0 & 0\end{bmatrix} = \begin{bmatrix}12 & 2\\2 & 2\end{bmatrix}$$

5. the between-cluster scatter matrix S_B

$$\begin{aligned}
S_B &= 3\left(\begin{bmatrix}2\\2\end{bmatrix} - \begin{bmatrix}5.125\\2\end{bmatrix}\right)\left(\begin{bmatrix}2\\2\end{bmatrix} - \begin{bmatrix}5.125\\2\end{bmatrix}\right)^T + 5\left(\begin{bmatrix}7\\2\end{bmatrix} - \begin{bmatrix}5.125\\2\end{bmatrix}\right)^T\left(\begin{bmatrix}7\\2\end{bmatrix} - \begin{bmatrix}5.125\\2\end{bmatrix}\right) \\
&= 3\begin{bmatrix}-3.125\\0\end{bmatrix}\begin{bmatrix}-3.125 & 0\end{bmatrix} + 5\begin{bmatrix}1.875\\0\end{bmatrix}\begin{bmatrix}1.875 & 0\end{bmatrix} \\
&= 3\begin{bmatrix}9.76 & 0\\0 & 0\end{bmatrix} + 5\begin{bmatrix}3.51 & 0\\0 & 0\end{bmatrix} \\
&= \begin{bmatrix}29.29 & 0\\0 & 0\end{bmatrix} + \begin{bmatrix}17.57 & 0\\0 & 0\end{bmatrix} = \begin{bmatrix}46.875 & 0\\0 & 0\end{bmatrix}
\end{aligned}$$

6. the scatter criterion $\frac{tr(S_B)}{tr(S_W)}$

$$\text{Scatter Criterion} = \frac{tr(S_B)}{tr(S_W)} = \frac{46.875 + 0}{12 + 2} = \frac{46.875}{14} = 3.348$$

Question 4: Consider density-based clustering algorithm DBSCAN with parameters $\epsilon = \sqrt{2}$, MinPts = 3 and Euclidean distance measures. Given the following points:

$$(0,0), (1,2), (1,6), (2,3), (3,4), (5,1), (4,2), (5,3), (6,2), (7,4)$$

1. List the clusters in term of their points.

$$C_1 : \{(1,2), (2,3), (3,4)\}$$
$$C_2 : \{(4,2), (5,1), (5,3), (6,2)\}$$

2. What are the density-connected points?
   Cluster 1 and 2 both form its own set of density-connected points.
   Namely, $\{(1,2), (2,3), (3,4)\}$ is one set of density-connected points and $\{(4,2), (5,1), (5,3), (6,2)\}$ is another set of density-connected points.

3. What points (if any) does DBSCAN consider as noise?

$$(0,0), (1,6), (7,4)$$

Question 5: A Naive Bayes Classifier gives predicted probability of each data point belonging to the positive class, sorted in a descending order:

Instance #	True Class Label	Predicted Probability of Positive Class	Predicted Class Label	Type
1	P	0.95	P	TP
2	N	0.85	P	FP
3	P	0.78	P	TP
4	P	0,66	P	TP
5	N	0.60	P	FP
6	P	0.55	P	TP
7	N	0.43	N	TN
8	N	0.42	N	TN
9	N	0.41	N	TN
10	P	0.40	N	FN

Suppose 0.5 is the threshold to assign the predicted class label to each data point, i.e., if the predicted probability ≥ 0.5, the data points is assigned to the positive class; otherwise, it is assigned to the negative class. Calculate the confusion matrix, accuracy, precision, recall, F1 score and specificity of the classifier.

Confusion Matrix:

		Truth		
		Positive	Negative	Total
Prediction	Positive	40%	20%	60%
	Negative	10%	30%	40%
	Total	50%	50%	100%

Calculations:

$$\begin{aligned}
\text{Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN} = \frac{7}{10} = 70\% \\
\text{Precision} &= \frac{TP}{TP+FP} = \frac{4}{6} = 66\% \\
\text{Recall} &= \frac{TP}{TP+FN} = \frac{4}{4+1} = 80\% \\
\text{F1 score} &= \frac{2TP}{2TP+FP+FN} = \frac{2\cdot 4}{2\cdot 4+2+1} = \frac{8}{11} = 72\% \\
\text{Specificity} &= \frac{TN}{FP+TN} = \frac{3}{2+3} = 60\%
\end{aligned}$$