BINUS UNIVERSITY
GLOBAL CLASS

Final Exam Answer

Student name: *Christopher Angelo - 2440041503*

Course: *Data Mining (COMP6746001)* – Professor: *Ms. Dewi Suryani*
Due date: *February 7th, 2023*

Question 1

Table 1 is a dataset of golf data. It comprises the attributes of the weather as the parameter to determine whether it is okay to play golf. Suppose we want to use a decision tree classifier. Please calculate the entropy and gain from each entity and display your result in a tabular form

No.	Play	Outlook	Temp.	Humidity	Wind
1.	No	Sunny	85	85	False
2.	No	Sunny	80	90	True
3.	Yes	Overcast	83	78	False
4.	Yes	Rain	70	96	False
5.	Yes	Rain	68	80	False
6.	No	Rain	65	70	True
7.	Yes	Overcast	64	65	True
8.	No	Sunny	72	95	False
9.	Yes	Sunny	69	70	False
10.	Yes	Rain	75	80	False
11.	Yes	Sunny	75	70	True
12.	Yes	Overcast	72	90	True
13.	Yes	Overcast	81	75	False
14.	No	Rain	71	80	True

Note: The dataset has only 14 records. The attributes are 4: Outlook, Temperature, Humidity, and Wind. The class label is 2: Yes and No.

Answer 1. Entropy is the degree of randomness in a dataset. The entropy of a dataset is calculated by using the following formula:

$$H(S) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{1}$$

Play. Calculating the entropy of the Play attribute, we have 9 values positive and 5 values negative. The entropy of the Play attribute is calculated as follows:

$$\begin{aligned} H(Play) &= -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} \\ &= 0.94 \end{aligned} \tag{2}$$

Outlook. The Outlook attribute has 3 values: Sunny, Overcast, and Rain. The entropy of the Outlook 'Sunny' is calculated as follows:

$$\begin{aligned} H(Outlook, Sunny) &= -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} \\ &= 0.97 \end{aligned} \tag{3}$$

The entropy of the Outlook 'Overcast' is calculated as follows:

$$\begin{aligned} H(Outlook, Overcast) &= -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} \\ &= 0 \end{aligned} \tag{4}$$

The entropy of the Outlook 'Rain' is calculated as follows:

$$\begin{aligned} H(Outlook, Rain) &= -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} \\ &= 0.97 \end{aligned} \tag{5}$$

After those calculations, we can calculate the gain of the Outlook attribute as follows:

$$\begin{aligned} Gain(Outlook) &= H(Play) - \frac{5}{14}H(Outlook, Sunny) - \frac{4}{14}H(Outlook, Overcast) - \frac{5}{14}H(Outlook, Rain) \\ &= 0.94 - \frac{5}{14}0.97 - \frac{4}{14}0 - \frac{5}{14}0.97 \\ &= 0.246 \end{aligned} \tag{6}$$

Temperature. Before we calculate the entropy of the Temperature attribute, we need to sort the values of the Temperature attribute. The sorted values including their play values are as follows:

Play	Temp
Yes	64
No	65
Yes	68
Yes	69
Yes	70
No	71
No	72
Yes	72
Yes	75
Yes	75
No	80
Yes	81
No	85

After calculating entropy and gain splits for each midpoints the value, the best split point that will produce the highest gain after split is 84. Such so, every temperature will be grouped into 2 groups: Cold and Warm where 84 is the splitting point. So, the table data will be as follow:

Play	Temp	
Yes	Cold	64
No	Cold	65
Yes	Cold	68
Yes	Cold	69
Yes	Cold	70
No	Cold	71
No	Cold	72
Yes	Cold	72
Yes	Cold	75
Yes	Cold	75
No	Cold	80
Yes	Cold	81
No	Warm	85

The entropy of the Temperature where the attribute is Cold is calculated as follows:

$$\begin{aligned} H(Temperature, Cold) &= -\frac{8}{13}\log_2\frac{8}{13} - \frac{4}{13}\log_2\frac{4}{13} \\ &= 0.890 \end{aligned} \tag{7}$$

The entropy of the Temperature where the attribute is Warm is calculated as follows:

$$\begin{aligned} H(Temperature, Warm) &= -\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1} \\ &= 0 \end{aligned} \tag{8}$$

So, the gain of the Temperature attribute is calculated as follows:

$$\begin{aligned} Gain(Temperature) &= H(Play) - \frac{13}{14}H(Temperature, Cold) - \frac{1}{14}H(Temperature, Warm) \\ &= 0.94 - \frac{13}{14}0.890 - \frac{1}{14}0 \\ &= 0.151 \end{aligned} \tag{9}$$

Humidity. Before we calculate the entropy of the Humidity attribute, we need to sort the values of the Humidity attribute. The sorted values including their play values are as follows:

Play	Humidity
Yes	65
No	70
Yes	70
Yes	70
Yes	75
Yes	78
No	80
Yes	80
Yes	80
No	85
No	90
Yes	90
No	95
Yes	96

After calculating entropy and gain splits for each midpoints the value, the best split point that will produce the highest gain after split is 82.5. Such so, every humidity will be grouped into 2 groups: Humid and dry where 82.5 is the splitting point. So, the table data will be as follow:

Play	Humidity Group	Humidity
Yes	Dry	65
No	Dry	70
Yes	Dry	70
Yes	Dry	70
Yes	Dry	75
Yes	Dry	78
No	Dry	80
Yes	Dry	80
Yes	Dry	80
No	Humid	85
No	Humid	90
Yes	Humid	90
No	Humid	95
Yes	Humid	96

The entropy of the Humidity where the attribute is Dry is calculated as follows:

$$\begin{aligned} H(Humidity, Dry) &= -\frac{7}{9}\log_2\frac{7}{9} - \frac{2}{9}\log_2\frac{2}{9} \\ &= 0.991 \end{aligned} \tag{10}$$

The entropy of the Humidity where the attribute is Humid is calculated as follows:

$$\begin{aligned} H(Humidity, Humid) &= -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} \\ &= 0.971 \end{aligned} \tag{11}$$

So, the gain of the Humidity attribute is calculated as follows:

$$\begin{aligned} Gain(Humidity) &= H(Play) - \frac{9}{14}H(Humidity, Dry) - \frac{5}{14}H(Humidity, Humid) \\ &= 0.94 - \frac{9}{14}0.991 - \frac{5}{14}0.971 \\ &= 0.151 \end{aligned} \tag{12}$$

Wind. The entrop of the Wind where the attribute is True is calculated as follows:

$$\begin{aligned} H(Wind, True) &= -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} \\ &= 1 \end{aligned} \tag{13}$$

The entropy of the Wind where the attribute is False is calculated as follows:

$$\begin{aligned} H(Wind, False) &= -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} \\ &= 0.811 \end{aligned} \tag{14}$$

So, the gain of the Wind attribute is calculated as follows:

$$\begin{aligned} Gain(Wind) &= H(Play) - \frac{6}{14}H(Wind, True) - \frac{8}{14}H(Wind, False) \\ &= 0.94 - \frac{6}{14}0.918 - \frac{8}{14}0.954 \\ &= 0.048 \end{aligned} \tag{15}$$

Question 2

Good clustering method is considered by high intra-class similarity and low inter-class similarity. Elaborate on the statement above with your own words and provide the example.

Answer 2. The process of clustering refers to the grouping / clumping of data points into a group (clusters). When we have a dataset, we can use clustering to group the data points into a group. The more data points that are similar to each other, the more likely they will be grouped into the same cluster. The more data points that are different from each other, the more likely they will be grouped into different clusters.

It is naturally great to have a little amount of cluster which have large encompassing group in the dataset while still having a big distance between the clusters. Though, these type of analysis won't always result in a clear-cut outcome of the clustering.

An **Intra**-class similarity is the similarity between the data points within the same cluster. As it is stated in the question: *A good clustering method is considered by high intra-class similarity*, which means that the more similar the data points within the same cluster, the better the clustering method is.

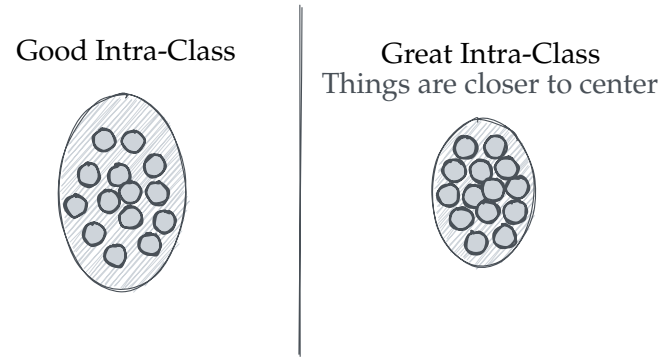

Figure 1: Intra-class similarity

An **Inter**-class similarity is the similarity between the data points between different clusters. Continuing on the statement from the question, the more different the data points between different clusters, the better the clustering result is.

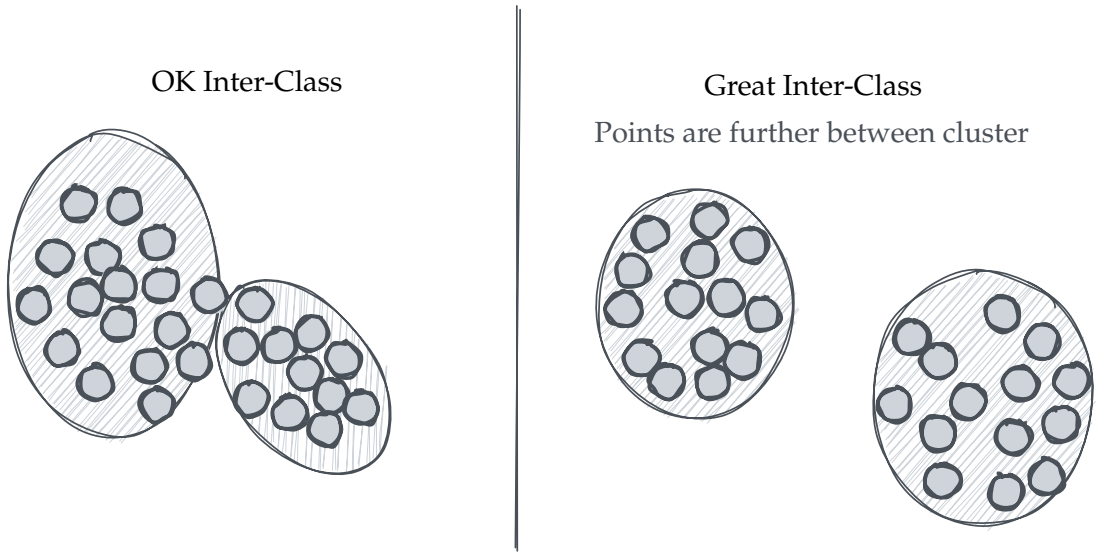

Figure 2: Inter-class similarity

Question 3

The following is the actual and the predicted result of the prediction

Both tables omitted for brevity . . .

The experiment result is to predict each square with the 2 class label: BLUE and RED. Other squares which do not meet the threshold remain empty. Your task is:

a. Create your own predicted-B table based on the predicted-A table by changing the color of the grid to the opposite color (if the grid in the predicted-A is RED, then change it to BLUE and vice versa). This changing grid is based on the first letter of your name and mod by 10 for the X axis, and the second letter of your name is mod by 5 for your Y axis. The coordinates X and Y will show you the location of the grid. If you have 2 words in your name, you must change 2 grids, and so on

b. Generate your confusion matrix based on the actual data and (your own) predicted-B data.

c. Based on the confusion matrix you generated, calculate the accuracy, recall, precision, and F1-Score.

Answer 3a. My first two letters in each word of my name is CH and AN. CH is the 3rd and the 8th letter in the alphabet, while AN is the 1st and 14th letter in the alphabet. Thus, the following is the grid that I have changed:

Letter	X	Y
CH	3	3
AN	1	4

So the following is the predicted-B table (Bolded cells are the cells that I have changed):

Answer 3b.

		Actual	
		BLUE	RED
Predicted	BLUE	9	1
	RED	4	7

	0	1	2	3	4	5	6	7	8	9
0	BLUE		BLUE						RED	
1		BLUE	RED				RED		RED	
2	BLUE	BLUE	RED					RED		BLUE
3		BLUE	RED	**BLUE**		RED	RED		BLUE	BLUE
4					RED	RED				

Table 1: Predicted-B table

Answer 3c. The accuracy of the model is calculated by the following formula:

$$\text{Accuracy} = \frac{9+7}{9+1+4+7} = \frac{16}{21} = 0.7619 \tag{16}$$

The recall of the model is calculated by the following formula:

$$\text{Recall} = \frac{9}{9+4} = \frac{9}{13} = 0.6923 \tag{17}$$

The precision of the model is calculated by the following formula:

$$\text{Precision} = \frac{9}{9+1} = \frac{9}{10} = 0.9 \tag{18}$$

The F1-Score of the model is calculated by the following formula:

$$\text{F1-Score} = \frac{2 \times 0.6923 \times 0.9}{0.6923 + 0.9} = \frac{1.2571}{1.5923} = 0.7895 \tag{19}$$

Question 4

From the AoL assignments you have worked on with the group, please elaborate on the parts of your contribution to the project you are working on

Answer 4. Early on in the AoL assignment just before the mid-exam, I have helped the team on structuring the assignment where we've discussed on the question / problem that were to be answered upon. We have looked on Kaggle for datasets that we can use for the assignment. We have also discussed on the tools that we can use for the assignment. We have decided to use Python and Jupyter Notebook for the assignment.

We have thought on using a multiplayer tool for the assignment (such as deploying our own tool via using Coder OSS or using a PaaS such as Deepnote). But since it was more known, we have decided to use Google Colab for the assignment. Since Google Colab multiplayer support is not that great, we have also decided to have a main person to do the code while other members brainstorm and write the needed writings.