Qian Zhang

G1939418

Question 1:

(a). First Run

	①	②	③	④
Cluster:	{(0,0)},	{(1,1)}.	{(4,2)},	{(6,0)}
Centroid:	(0,0)	(1,1)	(4,2)	(6,0)

Distance:

①	0	–	–	–
②	$\sqrt{(0-1)^2+(0-1)^2}$ $= 1.41$	0	–	–
③	$\sqrt{(0-4)^2+(0-2)^2}$ $= 4.47$	3.16	0	–
④	6	5.10	2.83	0

{(0,0)}and {(1,1)} are the closet.

Second Run:

	①	②	③
Cluster:	{(0,0), (1,1)}	{(4,2)}	{(6,0)}
Centroid:	(0.5, 0.5)	(4,2)	(6.0)

Distance:

①	0	–	–
②	3.81	0	–
③	5.52	2.83	0

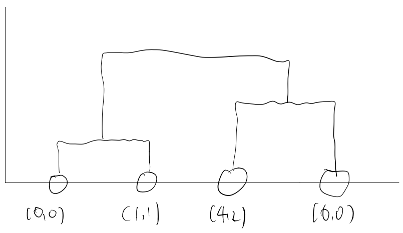
now (4,2) and (6,0) are the closet

Third Run: ① ②

Cluster: {(0,0). (1,1)} {(4,2). (6,0)}

Only two left

Dendrogram:



(b).

Yes, the result will be the same. There are two steps of Hierarchical Cluster Algorithms: the first step is to start with a single cluster for each data point, and the second step is to merge the two clusters with the minimum distance. The first step is always the same no matter each time of applying the algorithm to the same dataset. The second step might generate different results when there are two pairs of clusters that have the same distance and the algorithm need to merge one of them randomly, but in this question there are no pairs of clusters that have the same distance when using Euclidean distance and Centroid Distance.

Question 2

(a).

No. The Naive Bayes model relies on the calculation of the possibility. Therefore, if we discord any training sample, the number of samples will change and affect the calculation of possibility, and consequently affect the parameter learning of the Naive Bayes model.

(b). $P(Y=0) = \frac{1}{3}$, since in the 12 rows of data. 4 of them have value of 0. Therefore. $P(Y=0) = \frac{1}{3}$

$P(X_1=0 | Y=1) = \frac{2}{7}$, in the 12 rows of data. 8 rows have $Y=1$, among the 8 rows, 1 row has missing value of X_1, 2 rows have $X_1=0$. Therefore, $P(X_1=0 | Y=1) = \frac{2}{7}$

$$P(Y=0 | X_3=1) = \frac{P(X_3=1 | Y=0) P(Y=0)}{P(X_3=1)}$$

$$= \frac{\frac{1}{3} \times \frac{4}{12}}{\frac{7}{13}}$$

$$= 0.206$$

(c). $P(Y=1 \mid X_1=1, X_2=1, X_3=1)$

$$= \frac{P(X_1=1, X_2=1, X_3=1 \mid Y=1)\, P(Y=1)}{P(X_1=1, X_2=1, X_3=1)}$$

$$P(X_1=1, X_2=1, X_3=1) = P(X_1=1, X_2=1, X_3=1 \mid Y=1) \times P(Y=1)$$
$$+ P(X_1=1, X_2=1, X_3=1 \mid Y=0) \times P(Y=0)$$

since we apply a Naive Bayes Model here,

$P(X_1=1, X_2=1, X_3=1 \mid Y=1) \propto P(X_1=1 \mid Y=1) \cdot P(X_2=1 \mid Y=1) \cdot P(X_3=1 \mid T=1)$ and

$P(X_1=1, X_2=1, X_3=1 \mid Y=0) \propto P(X_1=1 \mid Y=0) \cdot P(X_2=1 \mid Y=) \cdot P(X_3=1 \mid T=0)$

$P(X_1=1 \mid Y=1) = \frac{5}{7}$, $P(X_2=1 \mid Y=1) = \frac{1}{7}$, $P(X_3=1 \mid Y=1) = \frac{6}{8}$

$P(X_1=1 \mid Y=0) = \frac{2}{4}$, $P(X_2=1 \mid Y=0) = \frac{3}{4}$, $P(X_3=1 \mid Y=0) = \frac{1}{3}$

$P(Y=1) = \frac{8}{12}$, $P(Y=0) = \frac{4}{12}$

so $P(X_1=1, X_2=1, X_3=1)$

$$= \frac{5}{7} \times \frac{1}{7} \times \frac{6}{8} \times \frac{8}{12} + \frac{2}{4} \times \frac{3}{4} \times \frac{1}{3} \times \frac{4}{12} = 0.0927$$

$$P(X_1=1, X_2=1, X_3=1 \mid Y=1) = \frac{5}{7} \times \frac{1}{7} \times \frac{6}{8} \times \frac{8}{12} = 0.051$$

$$P(Y=1 \mid X_1=1, X_2=1, X_3=1) = \frac{0.051 \times \frac{8}{12}}{0.0927} = 0.367$$

since it is smaller than 0.367,
then we predict that the label is 0

Question 3:

(a).

This new distance definition discards only the missing value but not the whole record. However, the missing value of the input variable might be very important of determining the property of this variable. By omitting the missing feature but not the whole record, the KNN algorithm might fail to find the closest neighbor of the test sample.

(b).

Since we don't want to ignore the missing value, we could utilize the KNN model to estimate the value of the missing input variables.

For example, if we have a new inputs variables X1, X2, X3 and new output variable Y, but the first row has a missing value of X1. We could treat the X1 as the new output variable, and treat X2 and X3 as the new input variables. Then, we could run a KNN model to predict the value of X1, and fill it in the table.

We could apply this strategy to fill all missing values in the table. After all the missing value are refilled, the calculation of distance between the test sample and the i-th training sample is normal, since all missing values have been replaced.

Since we estimate the most possible value of the missing features, we minimize the possibility of finding the wrong closest neighbor to the test sample.

Question 4:

(a).

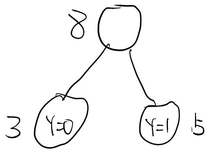
Taking the average of entropy and Gini index will not be an appropriate one, since Gini and Entropy has different interval. The largest possible number for Gini index is 0.5, while the largest possible value for Entropy index is 1. Therefore, when measuring a node which has identical numbers of class, Gini index returns 0.5 while Entropy returns 1. The two measures have different intervals, and taking the average of them will distort the judgment about the purity of the node.

(b).

The choose of using Entropy index or Gini index depends on the number of records of data. Since the Entropy index uses logarithms to calculate the purity of a node, the time of calculation is longer than Gini index. . On the other hand, if the number of records is not large, then we could use training set-validation set approach to determine which measure to use.

(c). $G(N)=\sum_{i=1}^{m} P_i(1-P_i)$

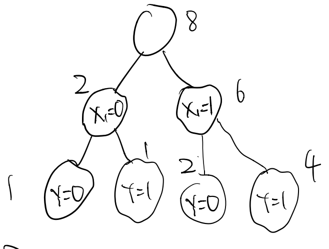
① Initial Run:



Purity of the root node: $\frac{3}{8}(1-\frac{3}{8})+\frac{5}{8}(1-\frac{5}{8}) = 0.469$

② First split

split by X_1



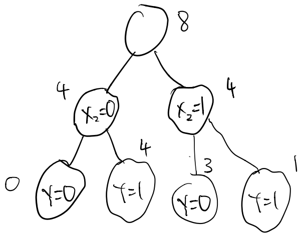
Purity: G(root Node) = 0.469

$G(X_1=0) = \frac{1}{2}(1-\frac{1}{2})+\frac{1}{2}(1-\frac{1}{2})$

$= 0.5$

$G(X_1=1) = 0.44$

split by X_2



Purity: G(root Node) = 0.469

$G(X_2=0) = \frac{0}{4}(1-\frac{0}{4})+\frac{4}{4}(1-\frac{4}{4})$

$= 0$

$G(X_2=1) = 0.375$

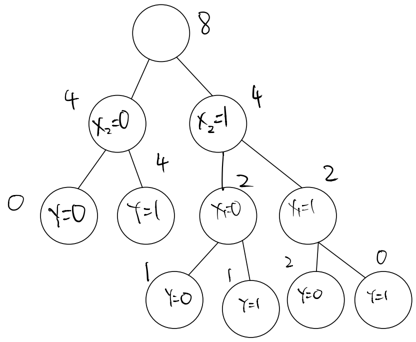
Information Gain: $0.469 - (\frac{2}{8} \times 0.5 + \frac{6}{8} \times 0.44)$ Information Gain: $0.469 - (\frac{4}{8} \times 0 + \frac{4}{8} \times 0.375)$

$= 0.014$ $= 0.285$

Therefore we choose to split by X_2

②. second split:

Since node ($X_2=0$) could not be further split. we only split node ($X_2=1$)



Question (d).

Since the X1 and X2 of test sample are both 1, we could simply filter the table by finding those record with X1 and X2 both equal to 1. In this table, the record 3 and 7 have the two values equal to 1, and the Y value of both the records are 0. Therefore, without training the tree, we could know that the predicted value of Y of the test sample is 0.