2) a) Purity : It is defined as the ratio between the dominant class of a cluster ω_i to the size of the ~~cls~~ i cluster.

$$\text{Purity} = \frac{\max(\omega_i)}{\text{Size of } |\omega_i|}$$

Rand Index : It can be defined as the ratio of ~~Sum of~~ # Samples that are correctly ~~classified~~ clustered as the ground truth ~~and~~ to the total number of Samples.

(3)

let us take a matrix of size 2×2

	Same clusters	Different clusters
Same Cluster in ground truth	A	B
Different cluster in ground truth	C	D

$$\text{Rand Index} = \frac{A+D}{A+B+C+D}$$

b) Inter cluster distance can be defined as the distance between two objects in different clusters

Intra cluster distance can be defined as the distance between two objects in same clusters.

Two Interclusters distances are :

Single linkage : It is the distance between any two closest objects placed in different clusters

$$\Rightarrow \min\left(\delta(x_i, x_j)\right)$$
$$x_i \in S$$
$$x_j \in T$$

S & T are two clusters.

Complete linkage : It is the distance between any two farthest objects placed in different clusters.

$$\Rightarrow \max_{\substack{x_i \in S \\ x_j \in T}} \left(\delta(x_i, x_j)\right)$$

Two Intracluster distances are :

Diameter linkage is the distance between any two farthest objects within the same cluster.

$$\Rightarrow \max_{\substack{x_i \in S \\ x_j \in S}} \delta(x_i, x_j)$$

Centroid Average linkage is defined as the two times of average distance of each object within a cluster to its centroid.

(4)

$$\Rightarrow 2 \times \frac{\sum_{i=1}^{v} \delta(x_i, v_s)}{|s|}$$ where s is cluster.

c) Dunn's cluster Validation Index is given as :

$$D.I = \min_{1 \le i \le k} \left\{ \min_{1 \le j \le k} \left\{ \frac{\delta(x_i, x_j)}{\max_{1 \le p \le k} (\Delta x_p)} \right\} \right\}$$

(4)

where, k = total number of clusters,

x_i represents the i^{th} cluster.

$\delta(x_i, x_j)$ = Interclusters distance between the cluster x_i & x_j

Δx_p = Intra cluster distance of the cluster x.

the Dunn's Index is defined in such a way that it tries to maximize the intra cluster distance and minimize the inter cluster distance.

The usefulness of the Dunn's Index is that it is used in the cluster evaluation. It gives us a measure to determine whether the number of clusters that we have made is good or not. The Partition of the clusters that maximizes Dunn's Index is the best.

3) a) An association rule is given as :

$$X \rightarrow Y$$

Support of the rule is given as : $\frac{(X \cup Y).\text{count}}{\text{Total number of transactions.}}$

i.e. the ratio of the number of transaction that contains $(X \cup Y)$ to the total number of transactions.

~~Confidence~~ — X

5) a) Tuple to classify:

$X = <$ Outlook = Overcast, Humidity = High, Wind = Weak $>$

Total number of distinct classes = 2.

5) Total number of tuples = 10.

$P(Yes) = \frac{6}{10} = \frac{3}{5}$

$P(No) = \frac{4}{10} = \frac{2}{5}$

In Naive Bayes theorem the tuple x belongs to class C_i only if

$P(C_i/x) > P(C_j/x)$ for $1 \le j \le m$ & $i \neq j$

where m = total number of clusters.

$$P(C_i|x) = \frac{P(x|C_i) * P(C_i)}{P(x)}$$

$P(x)$ is constant so we maximize $P(x|C_i) * P(C_i)$ only.

Since Naive Bayes assumes attribute as class Independence then,

$$P(x|C_i) = \prod_{i=1}^{n} P(x_i|C_i)$$ where n = total number of attributes.

~~$P(\text{Sunny}|\text{Yes}) = \frac{2}{6} = \frac{1}{3}$, $P(\text{Sunny}|\text{No}) = \frac{2}{4} = \frac{1}{2}$~~

$P($

$P(\text{Outlook} = \text{Overcast} / C_i = \text{Yes}) = \frac{2}{6} = \frac{1}{3}$.

$P(\text{Outlook} = \text{Overcast} / C_i = \text{No}) = \frac{0}{4} = 0.$

$P(\text{Humidity} = \text{High} / C_i = \text{Yes}) = \frac{2}{6} = \frac{1}{3}$.

$P(\text{Humidity} = \text{High} / C_i = \text{No}) = \frac{3}{4}$.

$P(\text{Wind} = \text{Weak} / C_i = \text{Yes}) = \frac{4}{6} = \frac{2}{3}$.

$P(\text{Wind} = \text{Weak} / C_i = \text{No}) = \frac{1}{4}$.

~~Hence~~

~~$P(x/\text{Yes})$~~

$\therefore P(x/\text{Yes}) = \frac{1}{3} \times \frac{1}{3} \times \frac{2}{3} = \frac{2}{27}$

$P(x/No) = 0 \times \frac{3}{4} \times \frac{1}{4} = 0.$

$P(x/yes) \times P(yes) = \frac{2}{27} \times \frac{3}{\cancel{10}5}$

clearly ~~$P(x/yes) > 0 \times$~~

$P(x/yes) \times P(yes)$ is greater than $P(x/No) \times P(No)$.

Hence, the tuple x is classified as ~~Yes~~ Play Tennis = Yes

b) Yes, there is an error in this prediction, there is not a single value for ~~Overcast~~ Outlook = Overcast & Play Tennis = No which is why $P(\text{Outlook} = \text{Overcast} / C_i = No) = 0$.

4 This error can be corrected using Laplacian correction in which for each distinct values in the attribute Outlook we will add extra one entry in the table. Total number of samples = 10

We add an entry for Outlook = Sunny, Rain & Overcast. with Play Tennis = No

So, ~~$P(\text{Sunny}) = \frac{5}{13}$, $P(\text{Rain}) = \frac{5}{13}$ & $P(\text{Overcast}) =$~~

$P(\text{Outlook} = \text{Sunny} / C_i : \text{No}) = \frac{3}{7}$

$P(\text{outlook} = \text{Rain} / C_i : \text{No}) = \frac{3}{7}$.

$+ P(\text{outlook} = \text{Overcast} / C_i : \text{No}) = \frac{1}{7}$.

This will remove the zero value error and will give correct classification.

c) The naive Bayes algorithm is used for both categorical and continuous values. For continuous value we can use the Gaussian distribution to calculate $P(x_i / C_i)$ where x_i is the ith attribute that have continuous values.

Gaussian distribution is given as :

3) $G(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

where μ = mean of the continuous range of x

σ = Standard deviation.

Q. If the attribute Humidity has a continuous value then we can find the mean ($\mu_{Humidity}$) as well as Standard deviation ($\sigma_{Humidity}$) then for any value x_k in the given tuple we can find $P(x_i/C_i) = g(x_k, \mu_{Humidity}, \sigma_{Humidity})$

$$= \frac{1}{\sqrt{2\pi}\,\sigma_{Humidity}} e^{\frac{-(x_k - \mu_{Humidity})^2}{2\sigma^2_{Humidity}}}$$

4)a) let positive class is yes then

$P = 6$ & $N = 4$.

$$\therefore \quad I(P,N) = -\frac{6}{10}\log_2\left(\frac{6}{10}\right) - \frac{4}{10}\log_2\left(\frac{4}{10}\right)$$

$$= -\frac{3}{5}\left[\log_2(3) - \log_2(5)\right] - \frac{2}{5}\times\left[\log_2(2) - \log_2(5)\right]$$

$$= \frac{-3}{5}\times\left[1.6 - 2.3\right] - \frac{2}{5}\times\left[1 - 2.3\right]$$

$$= \frac{3\times 0.7}{5} + \frac{1.3\times 2}{5}$$

$$= 0.94$$

$$E(A) = \sum_{i=1}^{k} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

For Attribute outlook,

with value Sunny : $p_i = 2$

& $n_i = 2$

$$\therefore I(p_i, n_i) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$$

$$= -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$$

$$= \frac{1}{2} + \frac{1}{2} = 1$$

Value = Overcast,

$p_i = 2$ & $n_i = 0$.

$$\therefore I(p_i, n_i) = -\frac{2}{2}\log_2\left(\frac{2}{2}\right) = -1 \times \log_2(1) = 0.$$

Value = Rain, $p_i = 2$ & $n_i = 2$

$$\therefore I(p_i, n_i) = 1$$

Hence, $E(\text{outlook}) = \frac{2+2}{10} \times 1 + 0 + \frac{2+2}{10} \times 1$

$$= \frac{8}{10} = \frac{4}{5}$$

$$\text{Gain}(\text{outlook}) = 0.94 - 0.8$$

$$= 0.14$$

For Attribute Humidity,

Value = Normal, $P_i = 4$ & $n_i = 1$

$$\therefore I(P_i, n_i) = -\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right)$$

$$= -\frac{4}{5}\log\frac{4}{5} \cdot \frac{4}{5} \times 0.3 + \frac{1}{5} \times 2.3$$

$$= 0.7$$

Value = High, $P_i = 2$ & $n_i = 3$

$$\therefore I(P_i, n_i) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$$

$$= \frac{2}{5} \times 1.3 + \frac{3}{5} \times 0.7$$

$$= 0.94$$

$$E(\text{Humidity}) = \frac{4+1}{10} \times 0.7 + \frac{2+3}{10} \times 0.94$$

$$= 0.82$$

$$\text{Gain}(\text{Humidity}) = 0.94 - 0.82 = 0.12$$

For attribute Wind,

with value = Strong, $P_i = 2$ & $n_i = 3$

$$I(P_i, n_i) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$$

$$= 0.94$$

value = weak. $p_i = 4$ & $n_i = 1$

$$I(p_i, n_i) = -\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right)$$

$$= 0.7$$

(5)

$$\therefore E(\text{wind}) = \frac{5}{10} \times 0.94 + \frac{5}{10} \times 0.7$$

$$= 0.82$$

$$\therefore \text{Gain}(\text{wind}) = 0.94 - 0.82$$

$$= 0.12$$

We can see that the maximum gain is for the attribut outlook with value 0.14

Hence the root of the decision tree is outlook.

b) The possible terminating criterias for the decision tree algorithm (3) are :-

i) When all the samples belong to the same class.

ii) When no samples are left.

iii) When there are no other attributes present. In this case we determine the leaf

class by majority.

c) Model overfitting occurs in the presence of the outliers and inconsistent data present in the training sample. Model tries to overfit the training example hence resulting in lower accuracy in test data. The overfitting in Decision Tree is avoided by Pruning :

i) Pre Pruning
ii) Post Pruning.

In Pre Pruning, we stop the growth of the ~~the~~ tree before it perfectly classify the training data. In ~~each~~ decision tree growth for every split we determine the gain of the split, if it is ~~is~~ less than some threshold value then we stop the growth of the tree.

3

In postpruning, we ~~let~~ the tree grow to the its maximum size then we remove the branches because of which overfitting may

occur. The branch is replaced by its leaf. In case of multiple classes we determine the new class with majority rule.

6)b) i) Holdout Method : In this method we simply divide the sample data into training and test data generally the partition is done as $\frac{2}{3}$ data is taken as the training data and $\frac{1}{3}$ data is taken as the test data. We train the model using training data and calculate the accuracy using test data.

ii) Cross-Validation : In this method the sample data is divided into K datasets It may use stratification as well in which the distribution of classes in each dataset is approximately equal. In the ith iteration the model is trained with all the dataset except the i^{th} data set which is used as the test data to determine accuracy.

- the overall accuracy can be determined by taking the average of all accuracies.

iii) Bootstrap: In this method the training data is sampled with replacement from the original dataset, i.e. same training data can appear more than once and some tuples might not even be in the training data. The tuples that doesn't end in training data will be used as test data.

4

c) The main purpose of ensembles of classifiers is to increase the overall accuracy and decrease the error. It trains multiple base models and the final output is determined by taking the output from each base model. Since we use multiple base models then the probability of error made by majority of the models is less.

The Adaboost algorithm works in the following way.

i) Initially all the Samples are given equal weight $\frac{1}{d}$, where d = total number of Samples.

ii) At the i^{th} iteration we train the i^{th} Model with weighted Samples.

iii) We determine the error of the model. If it is greater than 0.5 then we train the model again other wise we continue.

iv) We multiply $\frac{error(M_i)}{1-error(M_i)}$ to ~~each~~ the weight of each Sample which was correctly classified.

v) Then we normalize the weights by dividing the new weights with old weights. for all Samples. ~~of the~~ In this way the weight of the misclassified Samples are increased for the $i+1^{th}$ iteration.

vi) We Repeat the same Process again.

How to classify a given tuple:

i) Calculate the weight (w_i) of each model with $\log\left(\frac{1-error(M_i)}{error(M_i)}\right)$.

4 ii) Find the class predicted by the model M_i

iii) Add the weight w_i to class predicted by Model M_i

Now, the class with highest weight will be taken as the final prediction

6)a) For tuple, $x = \langle Sunny, Normal, Strong \rangle$, Actual class = Yes

Model prediction = Yes.

Similarly, $x = \langle Overcast, Normal, Strong \rangle$

Actual class = No, Model prediction = Yes.

$x = \langle Rain, High, Strong \rangle$

Actual class = Yes, Model prediction = No.

$x = \langle Sunny, High, Weak \rangle$

Actual class = No, Model Prediction = No

$X = \langle Rain, High, Strong \rangle$

Actual class = No, Model Prediction = No.

Hence, $TP = 1$, $FP = 1$

$TN = 2$, $FN = 1$

Confusion matrix is

	Predicted as positive	Predicted as negative
Actual positive class	1	1
Actual negative class	1	2

4

i) $\text{Precision} = \frac{TP}{TP+FP} = \frac{1}{1+1} = 0.5$

ii) $\text{Recall} = \frac{TP}{TP+FN} = \frac{1}{1+1} = 0.5$

iii) $\text{F-Score} = \frac{\text{Precision} \times \text{Recall}}{2 \times \text{Precision} + \text{Recall}}$

$= \frac{0.5 \times 0.5}{1+0.5} = \frac{0.5 \times 0.5}{1.5} = \frac{0.5}{3}$