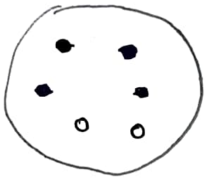
2. Cluster Validation indices.

Purity: ~~of a~~ Purity of cluster is defined as follows.

$$\pi_j = \frac{1}{n_i} \max(n_{ij})$$



Purity of the given cluster

$$\pi_j = \frac{1}{6} \max(3, 2, 1)$$

$$= \frac{3}{6}$$

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

$$= 0.5$$

3

Rand Index: Rand Index can be comprehende from the given analogy.

	Ground Truth Value with class X	Ground Truth Value not having X
~~Cluste~~ Sampling having class label X	A	C
Sample not having class label X	B	D

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

b.

Intercluster Distance. Intercluster cluster distance gives the similarity between two different clusters.

Single Linkage intercluster distance. The minimum distance between two datapoints ~~poi~~ present in the cluster gives single linkage distance.

a. Let us say S & T are two different clusters. So single linkage intercluster distance

$$\Delta_1(S,T) = \min_{\substack{\forall x \in S \\ \forall y \in T}} (\delta(x,y))$$

4

Complete Linkage Intercluster Distance: The maximum distance between two data points present in two different clusters.

Let us say S & T are two different clusters. So the distance is defined as.

$$\Delta_2(S,T) = \max_{\substack{\forall x \in S \\ \forall y \in T}} (\delta(x,y))$$

Intra Cluster Distance. Similarity between two ~~clusters~~ datapoints present in same cluster.

(diameter cluster distance)

Type 1. Intra cluster distance; for a cluster S

$$\Delta_1(S) = \max_{\substack{\forall x \in S \\ \forall y \in S \\ x \neq y}} (\delta(x,y))$$

Average Centroid Distance = ~~①~~ Average centroid distance is defined. ~~fo~~ as follows for a cluster S.

$$\Delta_1(S) = \frac{1}{|S|\ (|S|-1)} \sum_{x \in S} \delta(x, \bar{v})$$

where $\bar{v} = \frac{1}{|S|} \sum_{y \in S} y$

Dunn's Cluster Validation Index. The dunn's cluster validation index is defined as follows

3

$$D_{index} = \min_{1 \leq i \leq c} \left\{ \min_{\substack{1 \leq j \leq c' \\ i \neq j}} \left\{ \frac{\delta(x_i, x_j)}{\max(\Delta(x_{ic}))} \right\} \right\}$$

~~In~~ To obtain a good cluster the dunn's clustering index is ~~alway~~ ~~maximized~~. minimized. It is always preferred that the inter cluster distance of any two cluster is high but the intra cluster distance is low. ~~The~~ ~~dunn's~~ ~~index~~ ~~the numerator~~ ~~value~~ ~~is related~~ the inter ~~cluster~~ ~~between two~~ ~~cluster~~. ~~Dun~~ ~~Minimum~~ ~~so~~ Minimization of the ~~two~~ term $\frac{\delta(x_i, x_j)}{\max(\Delta(x_{ic}))}$ gives the same.

Q4.

$P = 6, \; n = 4$, $I(P,n) = -\frac{6}{10}\log\frac{6}{10} - \frac{4}{10}\log\frac{4}{10}$

$= 0.29$

Outlook	P_i	n_i	$I(P_i, n_i)$
Sunny	2	2	1
Overcast	2	0	0
Rain	2	2	1

$I(2,2) = \cancel{\frac{4}{10} \times 0.29} = \cancel{0.029}$

$= \cancel{0.049} \; \cancel{0.116} = 1$

$I(2,0) = \cancel{\frac{2}{10} \times 0.29} \quad 0$

$= \cancel{0.058}$

$E(A) = \sum \frac{P_i + n_i}{P + n} I(P_i, n_i)$

$= \frac{4}{10} \times 1 + \frac{2}{10} \times 0 + \frac{4}{10} \times 1$

$= 0.8$

GAIN (OUTLOOK) = 0.29 - 0.8

= -0.51

3½

Humidity	P_i	n_i	$I(P_i, n_i)$
Normal	4	1	~~0.145~~ 0.72
High	2	3	~~0.145~~ 0.97

$I(4,1) = \cancel{\frac{5}{10} \times 0.29} \quad$

$= \cancel{0.145} \quad 0.72$

$I(2,3) = \cancel{\frac{5}{10} \times 0.29}$

$= \cancel{0.145} \quad 0.97$

$E(A) = \sum \frac{P_i + n_i}{P + n} I(P_i, n_i)$

$= \frac{5}{10} \times 0.72 + \frac{5}{10}\, 0.97$

$= 0.845$

Gain (Humidity) = 0.555

Wind	P_i	n_i	$I(P_i, n_i)$
Strong	2	3	0.97
Weak	4	1	0.72

$I(2,3) = 0.97$

$I(4,1) = 0.72$

$$E(A) = \sum \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$= \frac{5}{10} \times 0.97 + \frac{5}{10} \times 0.72$$

$$= 0.845$$

$\text{Gain(wind)} = -0.555$

So ~~both wind and humidity~~ Outlook can be the root.

b. Possible terminating Criteria for decision tree:

i. The generation of decision during the training stops if the obtained gini index impurity at some node is higher than its parent.

15

ii. In ID3 approach, depending on the gain value ~~the~~ generated during the training process.

iii. During testing when the process reaches the leaf node it returns the possible output and terminates.

c.

A model generally overfits when there is high ~~low~~ variance in the datapoints but low bias.

The possible causes

i. There is a high variance present in the dataset.

ii. There is low bias present in the data-set.

iii. The model is ~~not~~ trained in such a way, i.e the weights are initialized in such a way, that there model cannot predict unseen data.

Decision tree ~~is~~ classifier is prone to overfit. ~~as~~ solution to overfitting

1. Use of Random forest: ~~A~~ In random forest classifier multiple decision trees are used during training. And after the classification is done the majority output is returned as predicted class.

2. Meta Modelling: Over fitting is also dealt with a techqonique known as meta modelling Here multiple decision tree classifier trained over different instances are taken. And there outputs are in turn fed into a different classifier to predict the values.

5.a ~~We need to calculate P(outlook/play Tennis = "yes") /outlook~~

~~& P(outlook/play Tennis = "no")~~

~~P(outlook = "overcast"/play Tennis = "yes")~~

~~P(play Tennis = "yes")/~~

Let us say X = < outlook = "overcast", humidity = "high", wind = "weak" >

4

~~P(~~ we are to calculate

P(play Tennis = "yes"/X) and P(play Tennis = "no"/X)

$$P(\text{play Tennis} = \text{"yes"}/X) = \frac{P(X/\text{play Tennis} = \text{"yes"}) \times P(\text{play Tennis} = \text{"yes"})}{\cancel{P(X)}\ P(X/PT = \text{"y"})\,P(PT = \text{'y'}) + P(X/PT = \text{"n"})\,P(PT = \text{"n"})}$$

$P(\text{play Tennis} = \text{"no"}) = \frac{4}{10}$ $P(\text{play Tennis} = \text{"yes"}) = \frac{6}{10}$

$P(\text{Outlook} = \text{"Overcast"}/\text{play Tennis} = \text{"yes"}) = \frac{2}{2} = 1$

$P(\text{Outlook} = \text{"Overcast"}/\text{play Tennis} = \text{"no"}) = 0/2 = 0$

$P(\cancel{\text{outlook}}\ \text{humidity} = \text{"high"}/\text{play Tennis} = \text{"yes"}) = \frac{1}{4}$

$P(\text{humidity} = \text{"high"}/\text{play Tennis} = \text{"no"}) = \frac{3}{4}$

$P(\text{wind} = \text{"weak"} / \text{playTennis} = \text{"yes"}) = \frac{4}{5}$

$P(\text{wind} = \text{"weak"} / \text{playTennis} = \text{"no"}) = \frac{1}{5}$

$P(X / \text{playTennis} = \text{"yes"}) = \frac{4}{5} \times \frac{1}{4} \times 1 = \frac{1}{5}$

$P(X / \text{playTennis} = \text{"no"}) = 0 \times \frac{1}{5} \times \frac{3}{4} = 0$

$$P(\text{playTennis} = \text{"yes"} / X) = \frac{6/10 \times 1/5}{\frac{6}{10} \times \frac{1}{5} + \frac{4}{10} \times 0}$$

$$= \frac{6/50}{6/50}$$

$$= 1$$

$$P(\text{playTennis} = \text{"no"} / X) = \frac{0 \times \frac{4}{10}}{\frac{6}{10} \times \frac{1}{5} + \frac{4}{10} \times 0}$$

$$= 0$$

$\therefore$ The class label of the give sample would be playTennis = "yes".

9b. Yes, there is an error in such prediction. Here from the dataset we can see that If outlook = "overcast" then there is no such sample where playTennis = "no". So in that situation $P(\text{playTennis} = \text{"no"} / \text{outlook} = \text{"overcast"})$ is 0. So no matter what ~~the~~ value the other featureset has ~~it will~~ outlook = "overcast" will return class label no.

4

To deal with this problem there are several methodologies available. One of them is Laplace Correction.

In laplace correct we add a very small value to both numerator and denominator making the probability of the biased class a very small non zero value.

So let us say we have 1000 sample. And in class prediction comes up where the probability distribution become

$P_1 = \frac{549}{1000}$, $P_2 = \frac{300}{1000}$, $P_3 = \frac{151}{1000}$, $P_4 = \frac{0}{1000}$

After Laplace correction the probability distribution becomes.

$$P_1 = \frac{550}{1004} \qquad P_3 = \frac{152}{1004}$$

$$P_2 = \frac{301}{1004} \qquad P_4 = \frac{1}{1004}$$

So P_4 will not result to zero after Bayes probability prediction, rather it will return a value close to zero.

c. If any ~~featuree~~ in naive bayes classification has a continuous value. ~~Th~~ Rather using the product we use the gaussian normal distribution

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

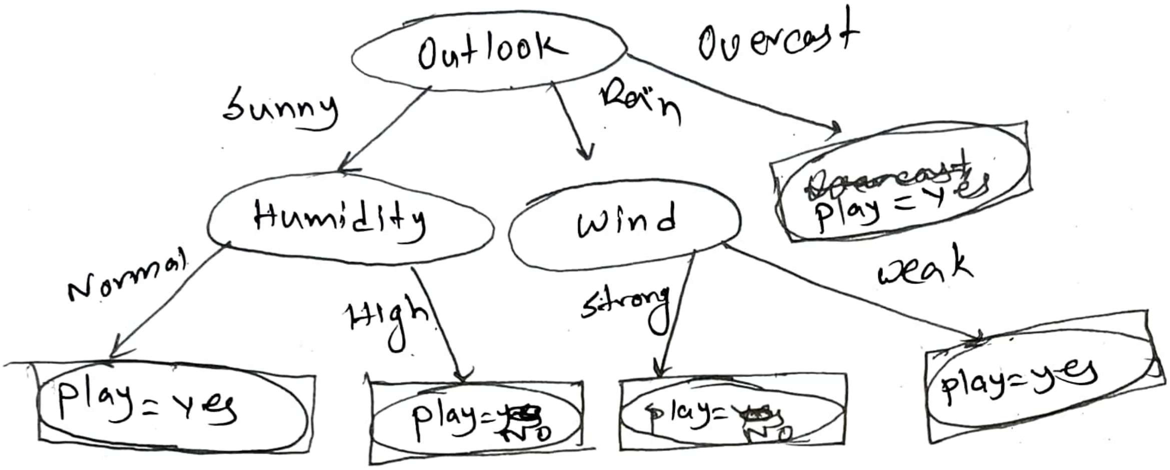
So for a sample x_k, $g(x_k, \mu_k, \sigma_k) = \frac{1 \times e^{\frac{(x-\mu_k)^2}{\sigma_k}}}{\sigma\sqrt{2\pi}}$

3

So $P(X/Y) = \frac{P(Y(X) \times P(X)}{\Sigma P(X)}$

So in ~~this~~ ~~case~~ ~~first~~ let us consider the x_k has a continuous ~~probal~~ distribution then ~~we~~ ~~use~~ instead of ~~co~~ the probability value we ~~to~~ use the gaussian normal ~~PDF~~ value.

6.a)



Outlook	Atumidity	Wind	Play Tennis CV = ~~yes~~	play Tennis predicted class
Sunny	Normal	Strong	Yes	Yes
~~Rain~~ Overcast	Normal	Strong	No	Yes
Rain	High	Strong	Yes	NO
Sunny	High	Weak	NO	NO
Rain	High	Strong	NO	NO

Confusion Matrix.

		Predicted Class Yes	Predicted Class No
Trues Class	Yes	True positive 1	False Negative 1
Trues Class	No	False positive 1	False Negative True 2

$$\text{Precision} = \frac{TP}{TP + FP}$$

4

$$= \frac{1}{1+1} = 0.5$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{1}{2} = 0.5$$

$$F_\beta \text{ score} = \frac{1+\beta^2}{\beta^2} \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

$$F_1 \text{ score} = \frac{2 \times 0.5 \times 0.5}{0.5 + 0.5}$$

$$= 0.5$$

b)

i. Holdout Method: In holdout method from the dataset ~~for~~ a percentage of the dataset is taken for training a the rest of the dataset is used for validation.

This selection is completely random and repeated ~~for~~ a number of times.

And ~~mean~~ noramally the mean of the all accuracies are reported with an error aproximation, i.e the standard deviation.

Training Dataset

~~Cross~~

↓

Random selection → $Accuracy_1$

↓

shuffle

↓

Random selection → $Accuracy_2$

⋮

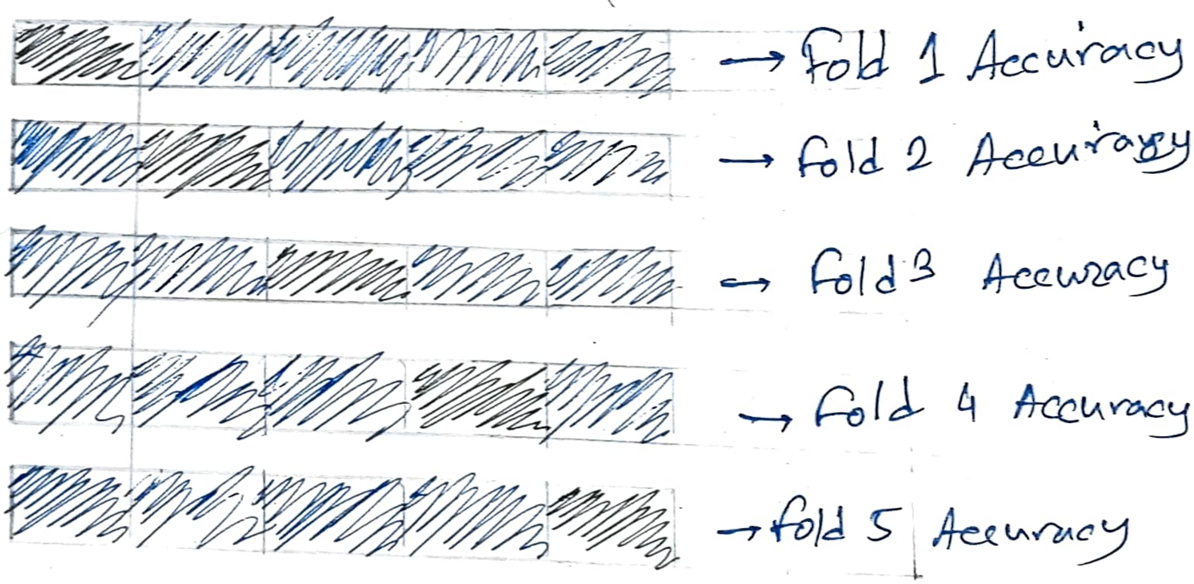
↓

Random selection → $Accuracy_i$

Accuracy = mean ($Accuracy_1$, $Accuracy_2$ ... $Accuracy_i$)
+ std_deviation ($Accuracy_1$, $Accuracy_2$ ... $Accuracy_i$)

Cross validation: There are mainly two types of cross validation present. i. K fold cross validation and startified cross validation.

In cross validation the dataset is divided into K folds. And the metrics is calculated K times. ~~First~~ Each of the fold is used as a validator for once, and rest of the folds are used for training



Accuracy = mean (Folds Accuracy) + std_deviation (Accuracy)

Bootstrap Method: ~~In~~ The bootstrap method works by the priniciple "resampling by replacement". Suppose there are d instances present in a bootstrap. One percentage of the bootstrap is used for training and the rest of the portion which couldnot make it to the boot strap is used for testing. The normal

3

bootstrap is 0.368, as $1-\frac{1}{d} \approx e^{-1} \approx 0.368$

The accuracy is calculated as follows.

$$Acc = (0.368\ Acc_{train\_set} + 0.632 \times Acc_{test\_set}.)$$

---

e. Main Purpose of Ensemble Classifier.

i. ~~Ensemble classifier~~ Use of ensemble classifier is fruitfull in various ways. For example a hypothesis is trained over a training set and the accuracy is calculated. But there is a possibility that the training set has some unseen instances. So using a ensemble classifier outputs to a set of accuracy. and predicted class lables. Depending on that weighted ~~class~~ majority the correct class lablel is choosen from a set of class labels.

Adaboostion: Adaboost is boosting algorithm used in ensemble classifiers. In ~~boosting~~ Adaboost several instances of classifiers are taken. ~~And then depending on that~~ ~~instan~~ And if a new instance is introduced that instance is give priority over all. And during traing there exist some particular threshold. If the new ~~value~~ instance works worse compared to that threshold the instance is discarde. the threshold is determined using null hypothesis or t testing

$$t = \frac{err(M_1) - err(M_2)}{\sqrt{var(M_1) - var(M_2)}}$$

2½