FDS - Theory Assignment 5

Ans 1-

(Validation Set Method)

a) Holdout Method → Simplest type of cross-validation where the dataset is seperated into 2 sets ; ie training and testing set. Model is prepared using the training set and predicts the output values for the data in testing set.

Advantage:-

i) It takes very less time to compute.

Disadvantage:-

i) This method can have a high variance.

ii) The error estimates can vary heavily (ie high variance) and depend heavily on which data points end up in the training set and which end in the test set.

~~Leave one out Method →~~

K-fold cross Validation Method → Dataset is divided into k subsets and the holdout method is repeated k times. Each iteration uses k-1 subsets as the training ~~sub~~set and the remaining one set as the test dataset. The average error across all k iterations is computed.

Advantage :-

i) Variance is reduced significantly since all the data points get to be in the training set k-1 times as well as in the testing set once.

Disadvantage:-

i> The method is computation intensive since it takes k iterations (compared to Holdout Method) to compute the result. Thus, it takes long time to run and prepare the model

Leave-one-out Cross Validation → It is k-folds method's extreme case where k equals N. where N is the number of data points in the set.

Advantage:-

i> It gives very good estimates since it has maximum coverage of data points and thus has minimum variance.

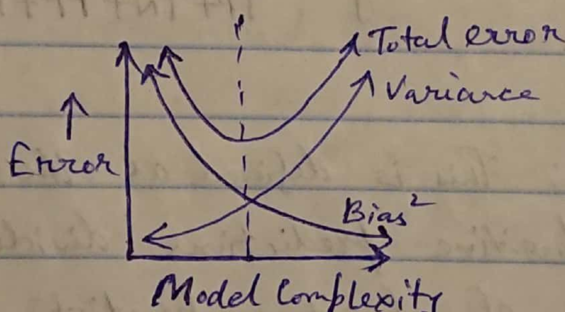
Disadvantage:-

i> This method is highly compute intensive since it is repeated N times.

Reference: https://www.cs.cmu.edu/~schneide/tut5/node42.html

b> The prediction error for an output ~~refers to the~~ is basically a result of error due to bias and error due to variance. Variance error is the error due to variability of the model prediction for a given data point. Bias error is the error obtained as the difference between the average prediction of the model and the correct value. ie. if the model building process is repeated, then this error is due to randomness in the underlying dataset, resulting in a range of predictions.

The tradeoff is that if we try reducing one type of error (bias OR variance), it results in the increase in the other type of error. For example: Ordinary Least Square method has high variance. Therefore, they are transformed ~~[illegible]~~ with the help of methods such as Ridge regression and LASSO to reduce variance, but in this process introduce variance.



Reference: https://classeval.wordpress.com/introduction/basic-evaluation-measures/

c) True positive : When actual and predicted value have been correctly identified as ~~give the same~~ positive. ~~class label.~~

False negative : When the predicted value is falsely (wrongly) identified as negative when the actual value is positive

False positive: When the predicted value is incorrectly identified as positive when the actual value is negative.

True negative: When the predicted value is correctly identified as negative value ie. predicted and actual value both are negative

Overall Accuracy: This is defined as the number of correct predictions divided by total number in the dataset.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: This is defined as the total number of correct positive predictions divided by the total number of positive predictions.

$$Precision: \frac{TP}{TP + FP}$$

Recall: This is defined as the number of correct positive predictions divided by the total number of positives.

$$Recall: \frac{TP}{TP + FN}$$

F-measure: This is defined as the harmonic mean of precision and recall.

$$\text{F-measure} = \frac{2 \times recall \times precision}{recall + precision}$$

Ans 2-a) Problems in the approach:

i) During the evaluation phase, there is no check for the class values if there is a conflict (due to complete match of 2 or more row values) where all the attributes match whereas the class labels are different.

ii) There is no mention of check of multicollinearity between the features that are selected.

The correct approach would be to find out the correlation matrix and the p-values to understand all those features that are significant and ~~take~~ ~~abs of~~ then removing all the features that can cause multicolinearity issue and only keeping one of them. Also, removing all the ~~insg~~ insignificant features. After a subset of features is selected, checking for any case where there might be a conflict due to ~~similar~~ same attribute values and then removing the conflict.

b)

x	y	
1	0	
2	0	← Train Set
3	0	
4	1	
5	1	← Test Set

$$\text{Mean} = \frac{10}{4} = 2.5$$

Hence, Test Set predicted = 1 (Since $5 > 2.5$)

Hence, $MSE_5 = 1$

x	y	
1	0	
2	0	
3	0	
4	1	← Test Set
5	1	

$$\text{Mean} = \frac{1+2+3+5}{4} = 2.75$$

Hence, Test Set predicted = 1 (Since, $4 > 2.75$)

Hence, $MSE_4 = 1$

x	y	
1	0	
2	0	
3	0	← Test Set
4	1	
5	1	

$$\text{Mean} = \frac{1+2+4+5}{4} = 3$$

Test Set Predicted $= \cancel{1}0$ (Since $3 = 3$)

Hence, $MSE_3 = 1$

x	y	
1	0	
2	0	← Test Set
3	0	
4	$\cancel{0}1$	
5	1	

$$\text{Mean} = \frac{1+3+4+5}{4} = \frac{13}{4} = 3.25$$

Test Set Predicted $= 0$ (Since, $2 < 3.75$)

$MSE_2 = 1$

x	y	
1	0	← Test set
2	0	
3	0	
4	1	
5	1	

$$\text{Mean} = \frac{2+3+4+5}{4} = \frac{14}{4} = \frac{7}{2} = 3.5$$

Hence, Test Set predicted $= 0$ (Since $1 < 3.5$)

$$MSE_1 = 1$$

$$C.V = \frac{\sum MSE}{n} \times 100 = \frac{(1+1+1+1+1)}{5} \times 100\,\% = 100\,\%$$

c) Accuracy Assessment

i)

	Predicted	
Observed	TP = 5	FN = 5
	FP = 1	TN = 4

	Classification Prediction	Ground Truth
TP ←	1	1
TN ←	2	2
TP ←	1	1
TP ←	1	1
TP ←	1	1
FN ←	2	1
FN ←	2	1
TN ←	2	2
FN ←	2	1
TN ←	2	2
FN ←	2	1
TN ←	2	2
FN ←	2	1
TP ←	1	1
FP ←	1	2

ii) $$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{9}{15}$$

iii) $$\text{Precision}(p) = \frac{TP}{TP+FP} = \frac{5}{6}$$

iv) $$\text{Recall}(r) = \frac{TP}{TP+FN} = \frac{5}{10} = \frac{1}{2}$$

v) $$\text{F-measure} = \frac{2rp}{r+p} = \frac{2\times\left(\frac{5}{6}\right)\left(\frac{1}{2}\right)}{\frac{5}{6}+\frac{1}{2}} = \frac{\frac{5}{6}}{\frac{5+3}{6}} = \frac{5}{8}$$

Ans 3-

Bayesian Interval for a normal prior distribution and normal population is given by,

$$\mu^* - z_{\alpha/2}\sigma^* < \mu < \mu^* + z_{\alpha/2}\sigma^*$$

Where, $\mu^* = \dfrac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2}$, $\sigma^* = \sqrt{\dfrac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}}$

Here, n = sample size = 10

$\bar{x}$ = sample average = 9

μ_0, σ_0^2 = normal prior mean & variance = 8, 0.2

σ^2 = sample variance = 0.64

Computing $\mu^* = \dfrac{(10\times 9\times 0.2) + (8\times 0.8^2)}{(10\times 0.2) + 0.8^2} = \dfrac{(90\times 0.2) + (8\times 0.64)}{(10\times 0.2) + 0.64}$

$$= \frac{23.12}{2.64} = 8.75$$

Computing $\sigma^* = \sqrt{\dfrac{0.2\times 0.8^2}{(10\times 0.2) + 0.8^2}} = \sqrt{\dfrac{0.128}{2.64}} = 0.220$

Hence,

$$\mu^* - z_{\alpha/2}\sigma^* = 8.75 - (1.96\times 0.220) = 8.3188$$

$$\mu^* + z_{\alpha/2}\sigma^* = 8.75 + (1.96\times 0.220) = 9.1812$$

Hence, the 95% Bayesian Interval is: $\boxed{[8.3188, 9.1812]}$

Ans 6-

a) $P(A=0, B=1, C=0, D=1, E=0, F=1)$

$= P(A=0)\cdot P(B=1)\cdot P(C=0|A=0)\cdot P(D=1|A=0, B=1)\cdot\cdot$
$\quad P(E=0|C=0, D=1)\cdot P(F=1|E=0)$

$= 0.4 \times 0.4 \times 0.8 \times 0.7 \times 0.4 \times 0.9$

$= 0.16 \times 0.56 \times 0.36$

$= 0.032256$

b) i) $P(A,B,C,D,E,F) = P(A)\cdot P(B)\cdot P(C|B)\cdot P(D|A,C)\cdot P(E|B,D,F)$
$\cdot P(F|A)$

ii) To prove $P(B,D|A,C) = P(B|A,C) * P(D|A,C)$

Taking LHS,

$P(B,D|A,C) = P(B|D,A,C)\cdot P(D|A,C)$

Solving $P(B|D,A,C)$

$$= \frac{P(B,D,A,C)}{P(D,A,C)}$$

$$= \frac{\sum_{EF} P(A,B,C,D,E,F)}{\sum_{BEF} P(A,B,C,D,E;F)}$$

$$= \frac{P(A)\cdot P(B)\cdot P(C|B)\cdot P(D|A,C)\sum_F P(F|A)\cdot\sum_{E,F} P(E|B,D,F)}{P(A)\cdot P(D|A,C)\cdot\sum_B P(B)\cdot P(C|B)\sum_F P(F|A)\sum_{E,F} P(E|B,D,F)}$$

$$= \frac{P(B) \cdot P(C|B)}{P(C)}$$

$$= \frac{P(B,C)}{P(C)} = P(B|C) = P(B) \quad \{\text{from Bayesian Net}\}$$

Therefore,

$$P(B,D|A,C) = P(B) * P(D|A,C)$$

Now, taking RH.S i.e. $P(B|A,C) * P(D|A,C)$

$$P(B|A,C) = \frac{P(B,A,C)}{P(A,C)}$$

$$= \frac{P(A) \cdot P(B) \cdot P(C|B)}{P(A) \cdot P(C|B)}$$

$$= P(B)$$

Hence, RHS becomes

$$= P(B) * P(D|A,C)$$

Hence, LHS = RHS.

Hence, proved.