Solution CSC591 HW5

Name: Rutvik Kolhe
Unity ID: rkolhe
Student ID: 200258232

Question 1 a) : Time-30mins

Validation Set

- In this technique, we randomly select a subset of the data as training data and the remaining as test data or validation set.
- The model is fitted on the training set and the training data set points are used on the fitted model to for model assessment. The test error rate is measured by calculating the MSE and the model with the smallest MSE is considered to be the best
- Advantages:
  - This technique is simple and easy to implement as the splitting of training and validation set is done randomly
- Disadvantages:
  - Since the splitting is done on a random basis, for every new sample, we get a different result of MSE and these results are highly variable. This variability fails to allow us to compare different models and choose the best.
  - Secondly, only a subset of samples that are present in the training data are chosen for training the model. When, the model is trained on fewer number of samples, it tends to overestimate the test error rate for the model on the entire data set.

Leave-One-Out Cross Validation (LOOCV)

- This is similar to the validation set approach but overcomes the disadvantages of the former method.
- Here, a single data point is selected for validation and all the remaining points (n-1) are used for training. The resulting fitted model is tested on single point and the corresponding MSE is calculated.
- This process is repeated almost n-times where at-least every data point is considered for testing. This decreases the bias as compared to the validation set approach.
- Advantages:
  - As mentioned above, we use (n-1) samples for training the data set as opposed to half of the samples in validation set approach. This does not overestimate the test error.
  - LOOCV technique produces the same result even after repeated iterations. There is no randomness in the results.
- Disadvantages:
  - LOOCV trains (n-1) samples , n-times, hence the computing time and resources required for training the data is large. This makes is expensive to implement LOOCV

k-Fold Cross Validation

- In this approach, we divide the entire dataset into k- partitions or folds. Out of these k-folds, one part is used for testing/validation while the others are used for training.
- This process is repeated, taking the different folds for testing at-least once. Lastly, the average of all the MSE’s is calculated

- Advantages:
    - This is computationally better than LOOCV, since the data is trained only k-times as opposed to n-times in LOOCV approach. Usually the value of k is 5 or 10.
- Disadvantages:
    - LOOCV provides better bias reduction as compared to k-fold validation set as the number of samples used for training are more (almost n) than k-fold approach, where the number depends on the value of k.

Question 1 b): Time-30mins

In the validation set approach, almost half the data set samples are used for training, hence it tends to overestimate the test error. Whereas, the LOOCV technique uses (n-1) samples for training. The dataset is trained n-times, thereby providing unbiased estimates. Accordingly, the k-fold approach, therefore provides intermediate bias estimate. We can conclude that LOOCV is better than k-fold and validation set approach for cross validation. However, LOOCV has a higher variance than k-fold cross validation. Since it provides outputs of n-fitted models which are correlated to each other. The outputs are highly correlated to each other since they are trained on almost similar samples as compared to k-fold outputs which are less correlated.

As we know, LOOCV is a special case of k-fold approach where the value of k is equal to n. Hence, the bias-variance trade-off is associated with the choice of the value of k. Typically, the value of k is chosen to be 5 or 10 so that there is neither excessive variance nor an excessive bias associated with the test values.

For classification, logistic and higher order polynomial models are used to define a decision boundary on the dataset. Cross validation is used to select the model which fits the best. Usually, 10- fold CV technique is used to compute the test error which is a result of fitting 10 different logistic models. The 10 fold CV provides the most accurate estimates for test error rate. If we use KNN classification approach, the 10-fold CV estimates the error rate that is very close to the best value of K (number of nearest neighbors).

Question 1 c): Time-15mins

Confusion Matrix is used to explain the performance of the classification models. Once the true values are known, we can understand the performance of the model by comparing the predicted values with the true values. The terminologies are: (Note: The outcomes being positive/ negative correspond to the fact that the outcomes belong to a particular class (positive) or do not belong to that particular class(negative) respectively)

- True Positive (TP)- The number of outcomes which are estimated to be positive and have a true positive value
- False Negative (FN)- The number of outcomes that are predicated to be negative but have a positive true value
- False Positive (FP)- The number of outcomes that are predicated to be positive but have a negative true value
- True Negative (TN)- The number of outcomes that are predicated to be negative and also have a negative true value
- Overall Accuracy – It describes how accurately the model has predicted the true values of positives and negatives to be correct. It id calculated using the formula- TP + TN / (Total number of values)

- Individual class Accuracy – It tells us how accurate is the model in predicting the mapping of outcomes to their particular output class. In other words, it tells us the ratio of the number of values that were estimated to belong to the particular class to the number of values that actually belong to that class.
- Precision (P) – It describes the ratio of the number of outcomes that were estimated correctly to that of the total number of outcomes predicted for that class. Formula – [TP / TP + FP]
- Recall (R)– It describes the ratio of the number of outcomes that truly belong to the class to the number of outcomes that are estimated to belong to that class. Formula- [TP / TP + FN]
- F-measure- It is defined as the weighted average of the precision and recall . Formula- [2RP / R+P]

Question 2 a): Time-20mins

The problem that I found out in the above example was the ordering in which the feature selection and cross validation takes place. In the given example, only 10 features are selected out of 2000 based on all the 100 provided samples. Intuitively, while feature selection took place, all the samples had been observed, after which 10 samples had been chosen. Now, if we split the test and training samples after feature selection, the problem arises while testing the logistic model as the samples had already been analyzed during the feature selection and therefore the test samples are not completely new. This would impact the results obtained after testing. Also, the given example does not consider the mutual correlation between the selected features and would fail to capture the non-linear relationship between the features and the predicted variable.
Hence, my solution would be to perform feature selection on the training data after cross validation, so that the test samples are completely independent of the feature selection analysis. Also, I observed that the number samples obtained are only 100. This would be a reason for anomalous prediction, as ideally the number of samples should atleast be 20-30 times the number of features. Hence, we atleast need 250-300 samples for a 10 feature dataset.

Question 2 b): Time-20mins

Given:

X	Y
1	0
2	0
3	0
4	1
5	1

For iteration 1:

Validation set = {1}
Training set = {2,3,4,5}

Mean = 2+3+4+5 / 4 = 14/4 = 3.5

Fitting model on training set

X	Y’
1	0
2	0
3	0
4	1
5	1

Using validation set to test model;
According to the model, Y’-value for 1 = 0
Therefore, MSE1 = 0

For iteration 2:

Validation set = {2}
Training set = {2,3,4,5}

Mean = 1+3+4+5 / 4 = 13/4 = 3.25

X	Y
1	0
2	0
3	0
4	1
5	1

Using validation set to test model;
According to the model, Y’-value for 2 = 0
Therefore, MSE2 = 0

For iteration 3:
Validation set = {3}
Training set = {1,2,4,5}

Mean = 1+2+4+5 / 4 = 12/4 = 3

X	Y
1	0
2	0
3	0
4	1

5	1

Using validation set to test model;
According to the model, Y’-value for 3 = 0
Therefore, MSE3 = 0

For iteration 4:

Validation set = {4}
Training set = {1,2,3,5}

Mean = 1+2+3+5 / 4 = 11/4 = 2.75

X	Y
1	0
2	0
3	1
4	1
5	1

Using validation set to test model;
According to the model, Y’-value for 4 = 1
Therefore, MSE4 = 0

For iteration 5:

Validation set = {5}
Training set = {1,2,3,4}

Mean = 1+2+3+4 / 4 = 10/4 = 2.5

X	Y
1	0
2	0
3	1
4	1
5	1

Using validation set to test model;
According to the model, Y’-value for 5 = 1
Therefore, MSE5 = 0

Hence, the average of MSE’s = 0

The actual and predicted labels for the data set using LOOCV are as follows:

X	Y	Y’
1	0	0
2	0	0
3	0	0
4	1	1
5	1	1

Since there is no variation in the predicted and actual values of Y- Accuracy = 100%

Question 2 c): Time-20mins

From the given table, we calculate the following values:
Let positive = 1 negative = 2
True Positive: 5
False Positive: 1
True Negative: 4
False Negative: 5

		Predicted		Total
		1	2	
Actual	1	TP : 5	FN : 5	10
	2	FP : 1	TN : 4	5
Total		6	9	15

Overall Accuracy = TP + TN / (n) = 5+4/15 = 0.6
Precision = TP / TP + FP = 5 / 6 = 0.833
Recall = TP / TP + FN = 5 / 10 = 0.5
F-Measure = 2 *RP / R + P = 2(0.5)(0.833) / 0.5 + 0.833 = 0.6249

Answers-

i	Overall Accuracy	0.6
ii	Precision	0.833
iii	Recall	0.5
iv	F-Measure	0.6249

Question 3) Time-30mins
Ref: Probability and Statistics for Engineers and Scientists- Walpole, Myers, Myers, Ye

From the given data

prior distribution

$\bar{x} = 9$ $\qquad$ $\mu_0 = 8$

$n = 10$ $\qquad$ $\sigma_0^2 = 0.2$

$\sigma = 0.8$

Now, we find the posterior distribution values : μ^*, σ^*

$$\mu^* = \frac{\sigma_0^2}{\sigma_0^2 + \frac{\sigma^2}{n}} \bar{x} + \frac{\sigma^2/n}{\sigma_0^2 + \frac{\sigma^2}{n}} \mu_0$$

$$= \frac{0.2}{0.2 + \frac{(0.8)^2}{10}} \times 9 + \frac{(0.8)^2/10}{0.2 + \frac{(0.8)^2}{10}} \times 8$$

$$= \frac{1.8}{0.264} + \frac{0.512}{0.264}$$

$$\mu^* = \underline{\underline{8.7575}}$$

$$\sigma^* = \sqrt{\frac{\sigma_0^2 \sigma^2}{n \sigma_0^2 + \sigma^2}}$$

$$= \sqrt{\frac{0.2 \times (0.8)^2}{(10)(0.2) + (0.8)^2}}$$

$\sigma^* = \underline{0.2202}$

Since we need to find 95% Bayesian interval for μ:

We use the formula:

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

Here

$$z_{\alpha/2} = z_{0.025} = 1.96$$

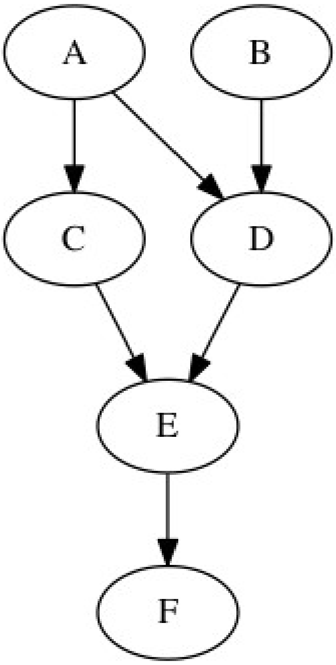
∴ Bayesian interval for μ is:

$$8.7575 - (0.2202)1.96 < \mu < 8.7575 + (0.2202)\,1.96$$

$$\Rightarrow \quad 8.3259 < \mu < \underline{9.1891}$$

Question 6 a) Time-15mins

Given:



From the given data, we can calculate the marginal probabilities and conditional probabilities as follows:

A	P(A)
0	0.6
1	0.4

B	P(B)
0	0.4
1	0.6

A	P(C | A)
0	0.2
1	0.7

A	B	P(D | A,B)
0	0	0.3
0	1	0.7
1	0	0.6
1	1	0.3

C	D	P(E | C,D)
0	0	0.8
0	1	0.6
1	0	0.2
1	1	0.5

E	P(F | E)
0	0.9
1	0.6

According to Bayesian networks,

$P(A, B, C, D, E, F) = P(A)\, P(B)\, P(C| A)\, P(D \mid A,B)\, P(E \mid C,D)\, P(F \mid E)$

Now,

$P(A = 0) = 0.4$

$P(B = 1) = 0.4$

$P(C = 0| A = 0) = 1- 0.2 = 0.8$

$P(D = 1 \mid A = 0, B = 1) = 0.7$

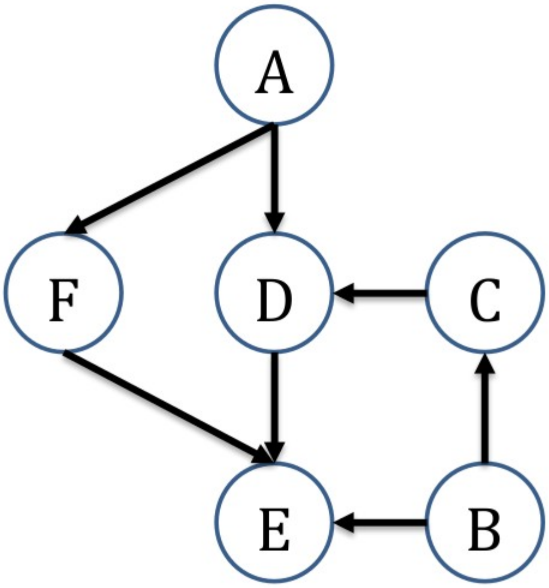
$P(E = 0 \mid C = 0, D = 1) = 1- 0.6 = 0.4$

$P(F = 1| E = 0) = 0.9$

Therefore,

$P(A, B, C, D, E, F) = 0.4*0.4*0.8*0.7*0.4*0.9 =$ **0.032256**

Question 6 b): Time-15mins



To find: $P(A, B, C, D ,E ,F)$

According to the Chain Rule:

$P(A, B, C, D ,E ,F) = P(A\mid B,C...F)\ P(B,C...F)\ P(B \mid C...F)\ P(C,...F).....P(F)$
Since, we have information from the Bayesian network, the above equation can be reduced to the following:

$P(A, B, C, D ,E ,F) = \mathbf{P(A) * P(B) * P(C \mid B) * P(D \mid A,C) * P(E \mid F, D, B) * P(F \mid A)}$

ii)
For B and to be conditionally independent, given A and C we need:

To prove: $P (B, D \mid A, C) = P(B \mid A, C) * P(D \mid A, C)$

We can simplify the above equation as follows:

$P (B, D \mid A, C) = [P(B, A, C) / P(A, C)]* P(D \mid A, C)$

$= [P(A) * P(B) P(C \mid B) / P(A) * P(C \mid B)] * P(D \mid A, C)$ {Derived above in (i)}

$= P(B) * P(D \mid A,C)$

To prove: $P (B, D \mid A, C) = P(B) * P(D \mid A,C)$

Now, we solve for LHS:

$$P(B, D \mid A, C) = P(B \mid D, A, C) * P(D \mid A, C)$$

$$= P(A, B, C, D) / P(D, A, C) * P(D \mid A, C)$$

$$= [\, P(A) * P(B) * P(C \mid B) * P(D \mid A, C) / P(A) * P(C \mid B) * P(D \mid A, C) \,] * P(D \mid A, C)$$

{Derived above in (i)}

$$= P(B) * P(D \mid A, C)$$

Hence, proven that B and D are conditionally independent given A and C