Name :- Parth Rathod

CWID :- A20458817

Course :- CSP-571 DPA

Sem :- FALL 2021

Part I

1.1]

Ans: $x_1 = (3, 4)$, $x_2 = (5, 12)$, $x_3 = (8, 15)$

$y_1 = 0$, $y_2 = 0$, $y_3 = 1$, $k = 2$, $x_0 = (0, 0)$.

↳

L_2 Norm of x_1 in R_3

$x_1 = \text{sqrt}(3^2 + 4^2) = \underline{5}$

Similarly,

$x_2 = \text{sqrt}(5^2 + 12^2) = \underline{13}$

$x_3 = \text{sqrt}(8^2 + 15^2) = \underline{17}$

Now to compute label of point x_0, we will see which 2 points are close.

Since x_1 and x_2 are close, we will assign their label

∴ Label of $x_0 = (0, 0)$ is $\underline{0}$

1.2]

Ans:-

$d = 15$, $n = 12,000$

N dimensional data will result in N x N covariance Matrix,

Hence in our case,

The dimensions of the covariance matrix = 15×15

A multivariate normal distribution is defined by two parameters,

- Mean Vector μ which is the expected value of distribution.
- Covariance Matrix Σ, which measures how dependent random variables are and how they change together.

The multivariate normal with normality 'd' has a joint probability given by,

$$p(x \mid \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(\frac{-1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

where,

μ = mean vector

Σ = Covariance matrix (size $d \times d$)

We denote multivariate normal distribution as,

$$N(\mu, \Sigma)$$

1.3] Total number of observations n.

Let the no. of folds be k.

Size of training set $= \frac{n}{k}(k-1)$

Size of validation set $= \frac{n}{k}$

All k folds have equal no. of observation. So each fold has $\frac{n}{k}$ terms.

If LOOCV is performed.

Size of each fold would be considered as single observation

$\therefore k = n$

Size of training set $= n-1$

Size of validation set $= 1$

1.4]

Ans:- For Bayes Decision Boundary to fall on point $x = 4.0$, the class covariances of both classes have to be equal. Otherwise, there will be an unequal separation between boundary with each class mean.

1.5]

Resampling with replacement will allow.

The sampling employed by bootstrap involves randomly selecting n observations with replacement, which means some observation can be selected multiple times while other observations are not included at all.

Total we can generate 4^4 sample data sets

1.6]

Ans:- A logit function is also called odds function.

In order to understand logit function, we need to understand what is an "odds".

ODDS is defined as probability of success / probability of failure.

Hence logit function / odds function is defined as, "the logarithm of the odds."

If we call parameter α, it is defined as

$$\text{logit}(\alpha) = \log\left(\frac{\alpha}{1-\alpha}\right)$$

The logistic function is inverse of the logit.

If we have value x, the logistic is:

$$\text{logistic}(x) = \frac{1}{1+e^{-x}}$$

The range of the odds function is "0" to "Infinity"

The rang of logistic function is "0" to "1"

Example says 20% chance of occuring.

$\therefore P(\text{occurring}) = 0.2$

$\therefore P(\text{not occurring}) = 0.8$

$\therefore$ ODDS $= P(\text{occurring}) / P(\text{not occurring}) = 0.2/0.8 = 0.25$

$\therefore$ The odds of occurring the event is 1 out of 4

1.7]

Ans:- Let the random variable be X

Let the sample size be n.

∴ Sample Mean is given by

$$u = \frac{1}{n} \sum_{i=1}^{n} x_i$$

∴ Sample Variance is given by

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - u)^2$$

Now, If we continue to replace missing values with sample mean u

$$x_j = u \text{ for some } j$$

$$\therefore x_j - u = 0$$

$$\therefore (x_j - u)^2 = 0$$

$$\therefore \sum_{i=1}^{n} (x_i - u)^2 \text{ reduces}$$

$$\therefore \frac{1}{n-1} \sum_{i=1}^{n} (x_i - u)^2 \text{ reduces}$$

$$\therefore \sigma^2 \text{ reduces}$$

Thus, Value of Sample Variance continue to decrease

For $n \to \infty$, we see $\sigma^2 \to 0$

Thus Sample Variance approach to Zero asymptotically.

1.8) When performing ridge regression, λ can vary between $[0, \infty)$

When $\lambda = 0$:

- The objective becomes same as simple linear regression.
- We will get the same coefficients as simple linear regression.
- No parameters are eliminated.

When $\lambda = \infty$:

- The coefficients will be zero. Because of infinite weightage on square of coefficients, anything less than zero will make objective infinite.
- All coefficients are eliminated.

When λ increases:

- The magnitude of λ will decide weightage given to different parts of objective.
- The coefficients will be somewhere between 0 and ones for simple linear regression.

Hence as the value of λ increases, the model complexity reduces.

2.1] $F = \frac{(TSS - RSS)/p}{RSS/n-p-1}$

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad RSS = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

p : no. of features

n : no. of observations.

$$\therefore F = \frac{TSS - RSS}{p} \times \frac{n-p-1}{RSS}$$

$$: \frac{TSS - RSS}{RSS} \times \left(\frac{n-p-1}{p}\right)$$

Since 'n' is considerably large then,

$$F \propto \frac{1}{p}$$

$\therefore$ F-statistic is inversely proportional to p [no. of features]

$\therefore$ p increases F Decreases.

If $p > n$, then there are more coefficients β_j to estimate then observations from which to estimate them.

In this case we cannot even fit the multiple linear regression model using least squares.

So F statistic cannot be used.

2.2]

Ans.) KNN acts poorly when feature dimensionality p increases, because when feature dimension is increased the length of feature space increases.

) All the data is stored in each dimension uniformly and separated by linear parameters.

) As length of dimension grows the feature space is less dense and distance between k neighbour increases.

) As the distance between neighbours increases the space to find neighbours increases;

Hence KNN acts poorly when feature space dimensionality is increased.

) This is also referred as Curse of Dimensionality, i.e The size of data space grows exponentially with number of dimensions. This means that size of our data set must also grow exponentially in order to keep same density. If you don't, then data points starts getting farther and farther apart.

) This is a problem with KNN as it requires a point to be close in every single dimension. It needs all points to be close along every axis in data space. And each new axis added, by adding a new dimension, makes it harder and harder for two specific points to be close to each other in every axis.

Equation for $f(x_0)$ with respect to K and No in higher dimension.

$$\therefore \quad f(x_0) = \frac{K}{N_0} * \frac{1}{P}$$

- In above equation P is higher dimension.
- The dimension P is increased linearly and the No neighbour is increased exponentially.
- The region of K neighbour is separated in a distance that they don't have anything in common and are located far in higher dimension.

2.3]

$\hat{\beta}_1 = 3.92$

$S.E(\hat{\beta}_1) = 1.25$

95 % Confidence Interval

$\therefore 1-\alpha = 0.05$

$t_{\alpha/2} = 1.96$ For 95 % we get 1.96 from t table

Confidence Interval $= \hat{\beta}_1 \pm t_{\alpha/2(d)} SE(\hat{\beta}_1)$

$\therefore$ Lower Limit $= 3.92 - (1.96) \times (1.25)$

$= 1.47$

$\therefore$ Upper Limit $= 3.92 + (1.96)(1.25)$

$= 6.37$

95 % confidence interval of β_1 for population value (1.47, 6.37)

To test,

$H_0 : \beta_1 = 0$ $H_a : \beta_1 \neq 0$

Test Statistic,

$$t = \frac{\hat{\beta}_1 - \beta}{SE(\hat{\beta}_1)} = \frac{3.92 - 0}{1.25} = 3.136$$

$\therefore$ At 95 % confidence interval we would reject null hypothesis for this coefficient.

2.4] $n = 101$

Variance $(\sigma^2) = 3.75$

RSS $= 125$

Sample of a Variance is given by.

$$\text{Variance } (\sigma^2) = \frac{\Sigma (x - \bar{x})^2}{n-1}$$

$$\therefore \Sigma (x - \bar{x})^2 = (n-1)\sigma^2$$

$$\therefore \Sigma (x - \bar{x})^2 = 100 \times 3.75$$

$$\therefore \Sigma (x - \bar{x})^2 = 375$$

$$\therefore \Sigma (x - \bar{x})^2 = TSS = 375$$

$$\therefore R^2 = 1 - \frac{RSS}{TSS}$$

$$\therefore R^2 = 1 - \frac{125}{375} = 1 - 0.33$$

$$\therefore R^2 = \underline{0.666}$$

Since Covariance asses the fit of model

$$\therefore Cov(x, y)^2 = R^2$$ For Single Predictor

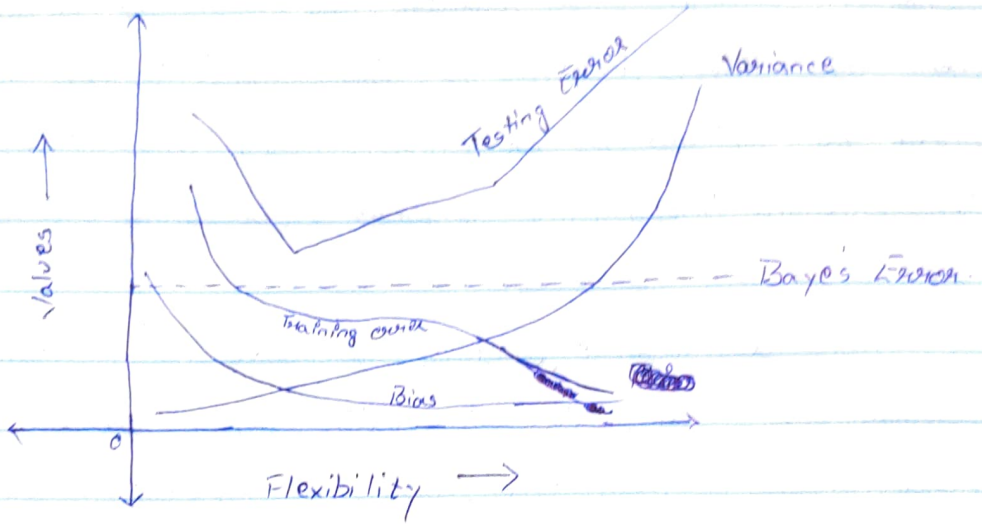
$$\therefore Cov(x, y) = \sqrt{0.666} = \sqrt{0.67}$$

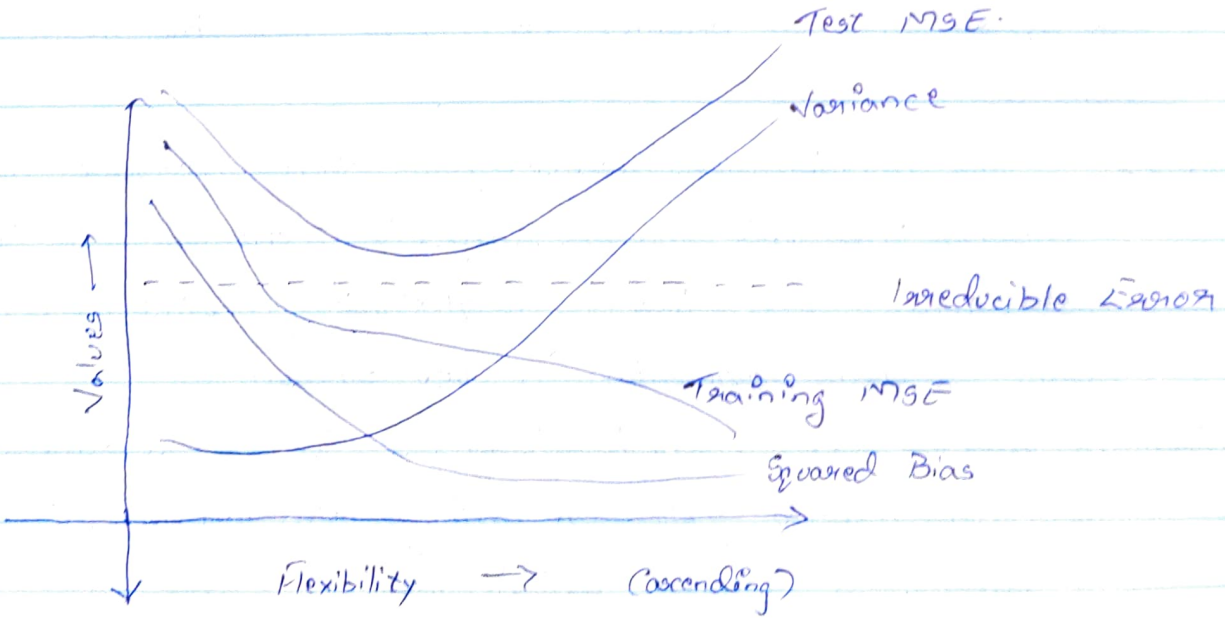
$$\therefore Cov(x, y) = \underline{0.8185}$$

Part III

17

Ans





- Bayes Error :- This is also referred as Irreducible Error. It remains constant for increase in flexibility as this is independent of model. As this is from specific set of data, they cannot be altered by modelling features. Thus it represents a parallel straight line usually assumed to be standard normal with $\mu = 0$ & $\sigma = 1$.
- Variance :- It defines a feasible increase in curve as flexibility increases from its inital value. The main reason for this curve characteristic is that model with high flexibility will yield better or improved variance. With increase in flexibility, the system function becomes less robust there by increase in variance.
- Bias :- This monotonically decreases while there is increase in flexibility. Generally, bias is inversely proportionall to flexibility. Bias decreases since increase in flexibility generates more complex system function representing a real problem.
- Training Error :- As flexibility increases, this error decreases as the model is being fitted on training data set. At the end, it attains a optimal constant value.
- Test Error :- This error will always be higher than variance as discussed above, because variance cannot be predicted by model as it is model independent. This is a place where the flexibility is starting to overfit the training data. Hence curve lies below test error.

Bonus Question:

1] Turing Award.

2] Pic2Recipie [looks at photos of food and predicts ingredients and suggests similar recipies].

3] $432,500

4] Google's Deepmind

5] University of Liverpool

6] GPT-3

7] Singapore