ML Assignment - 1

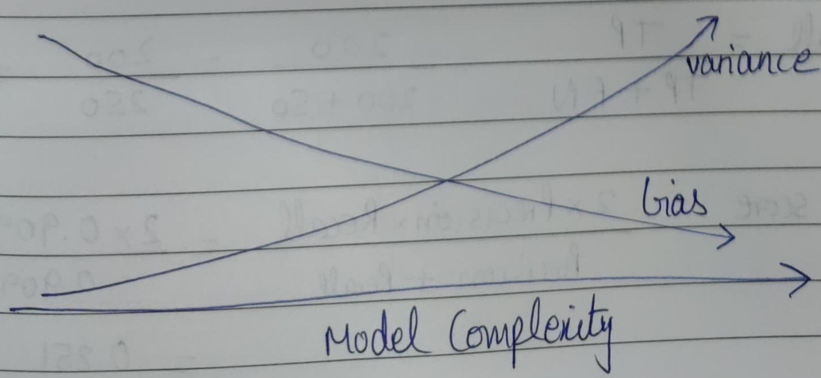
Theory

Ans 1 (a) At low model complexity :-

- Model ~~bias~~ has high bias and low variance
- The model is too simple and makes inaccurate predictions.
- Training error and test error are both high.
- This is known as an underfitting model.

As model complexity increases :-

- The model starts to fit the data well. If the complexity increases too much, the model learns the noise of the data i.e. fits the training data too well.
- At this point, the bias of the model decreases and variance increases.
- ~~The mod~~ Training error decreases but test error increases.



Page No.:
Date:

(b) True Positive (TP) : Spam emails correctly identified as spam.

False Negative (FN) : Spam emails incorrectly classified as legitimate

True Negative (TN) : Legitimate emails correctly classified as legitimate.

False Positive (FP) : Legitimate emails incorrectly flagged as spam.

TP = 200
FN = 50
TN = 730
FP = 20

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{200 + 730}{200 + 730 + 50 + 20}$$

$$= \frac{930}{1000} = 0.93$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{200}{200 + 20} = \frac{200}{220} = 0.909$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{200}{200 + 50} = \frac{200}{250} = 0.8$$

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times 0.909 \times 0.8}{0.909 + 0.8}$$

$$= 0.851$$

~~Specifity~~

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{730}{730 + 20} = \frac{730}{750} = 0.973$$

Model performs well with high accuracy and specificity i.e. it correctly handles legitimate emails most of the time.

High precision shows that most of the emails marked as spam are indeed spam.

80% recall indicates that 20% of spam emails are misclassified as legitimate.

(c) Using minimization of mean square error :- for linear regression

$$y = mx + b$$

$$m = \frac{\sum_{i=1}^{n} (\bar{x} - x_i)(\bar{y} - y_i)}{\sum_{i=1}^{n} (\bar{x} - x_i)^2}$$

n = no. of training samples

$$b = \bar{y} - m\bar{x}$$

$$\bar{x} = \frac{3+6+10+15+18}{5} = 10.4$$

$$\bar{y} = \frac{15+30+55+85+100}{5} = 57$$

$$m = \frac{(10.4-3)(57-15)+(10.4-6)(57-30)+(10.4-10)(57-55)+(10.4-15)(57-85)+(10.4-18)(57-100)}{(10.4-3)^2+(10.4-6)^2+(10.4-10)^2+(10.4-15)^2+(10.4-18)^2}$$

$$m = \boxed{5.78}$$

$$b = \bar{y} - m\bar{x} = 57 - (5.78)(10.4)$$

$$= 57 - 60.112 =$$

$$= \boxed{-3.112}$$

$$y = 5.78x - 3.112$$

For $x = 12$,

$$y = 5.78 \times 12 - 3.112$$

$$= \boxed{66.248}$$

(d) Toy example :-

Training data :-

X	Y
1	1
2	2
3	3
4	4

$f_1(x) \rightarrow$ complex model

$$f_1(x) = x^4 - 10x^3 + 35x^2 - 50x + 24$$

Predictions :-

X	Y	$\hat{f}_1(x)$
1	1	1
2	2	2
3	3	3
4	4	4

Perfectly fits the training data,

empirical risk (loss) on training set = 0

$f_2(x) \rightarrow$ Simple model

$f_2(x) = 2x$

X	Y	$\hat{f_2}(x)$
1	1	2
2	2	4
3	3	6
4	4	8

The empirical risk (loss) for $f_2(x)$ on training set is more than $f_1(x)$

If $L \rightarrow$ mean square loss, then $L(\hat{f_2}(x), y) = 7.5$

Unseen data $\rightarrow$

X	Y
5	5
6	6

$\hat{f_1}(x)$ and $\hat{f_2}(x)$

X	Y	$\hat{f_1}(x)$	$\hat{f_2}(x)$
5	5	15	10
6	6	38	12

Clearly, $\hat{f_2}(x)$ generalizes better on unseen data than $f_1(x)$.

$L(\hat{f_1}(x), y) = 562 \quad > \quad L(\hat{f_2}(x), y) = 30.5$