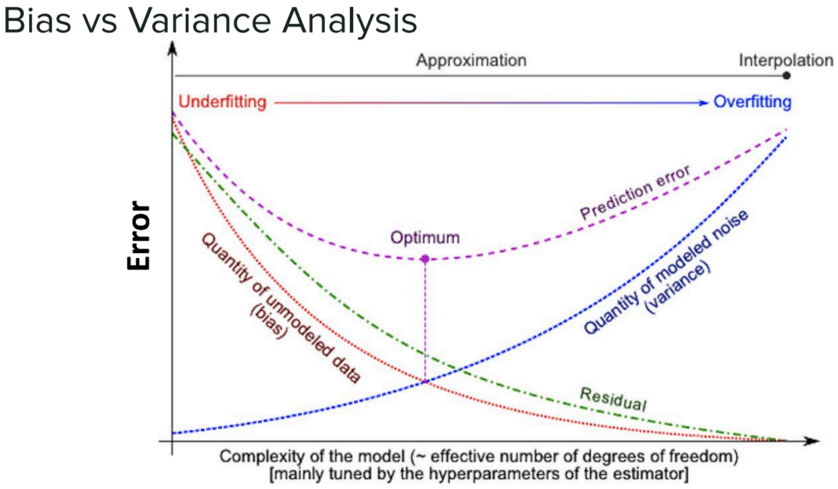
CSE 343 : Machine Learning
Assignment 1 Solutions

Section A

1. As we increase the complexity of the model , the model will be able to understand the underlying patterns easily thus reducing the bias but at the same time it will become more sensitive i.e will understand more from the noise around than the normal data thus increasing the variance of the model leading to overfitting of the data.



From the Lecture slides

[Image Reference : Tutorial 2 slides (in class)]

From this image we can see that as the complexity of the model increases (plotted on x axis), the variance increases and thus leads to overfitting of the data.

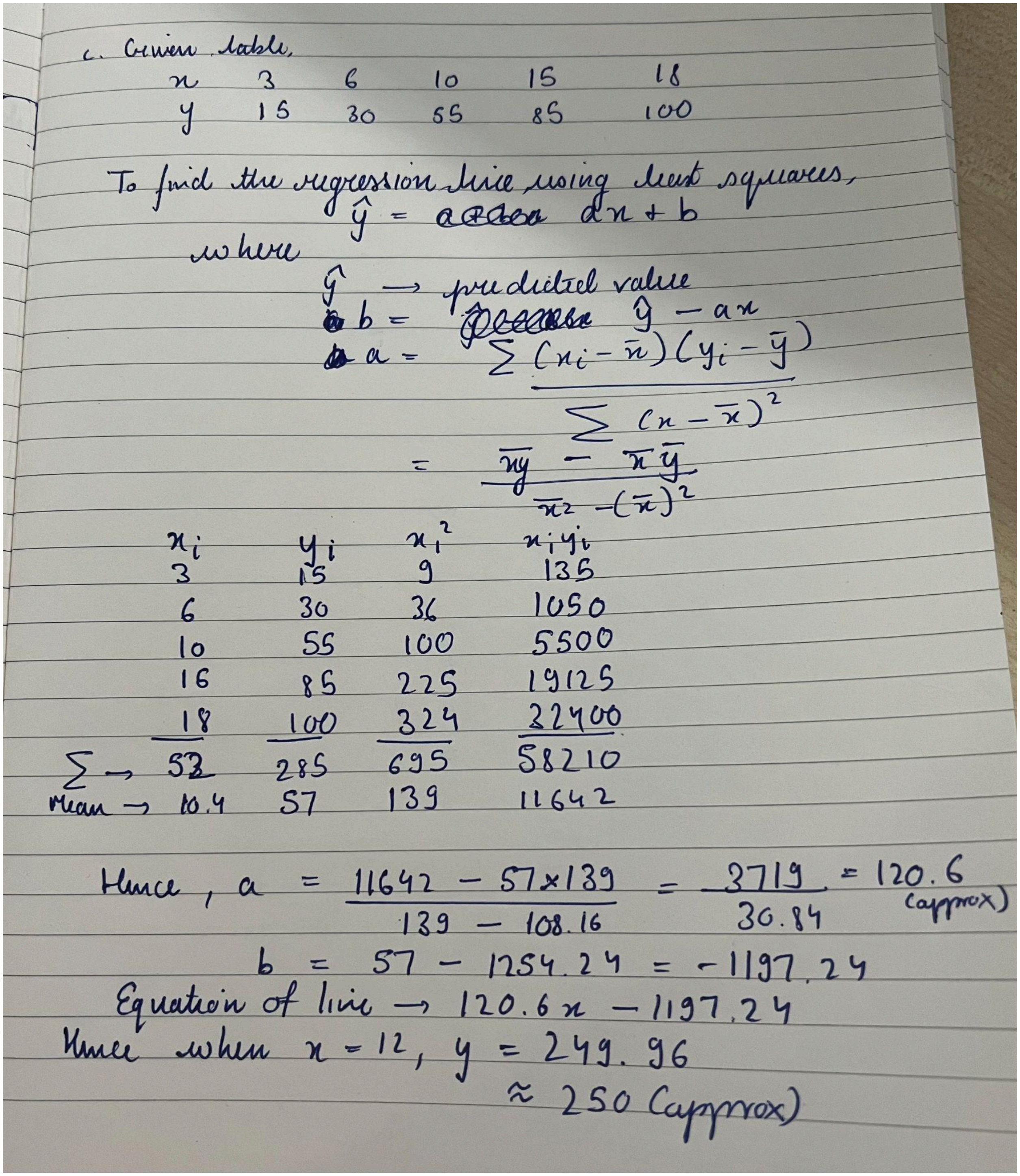
2. To measure the performance of the model, We can use two commonly used metrics of evaluation : Recall and Precision.
For spam emails :
True Positives (TP): 200 (spam emails correctly identified as spam)
False Positives (FP): 20 (legitimate emails incorrectly labeled as spam)
False Negatives (FN): 50 (spam emails incorrectly labeled as legitimate)
True Negatives (TN): 730 (legitimate emails correctly identified as legitimate)

Hence , Recall = TP / (TP + FN) = 200 / (200 + 50) = 0.80
And Precision = TP / (TP + FP) = 200 / (200 + 20) = 0.9091

For legitimate emails :

True Positives (TP): 730 (legitimate emails correctly identified)
False Positives (FP): 50 (spam emails incorrectly labeled as legitimate)
False Negatives (FN): 20 (legitimate emails incorrectly labeled as spam)
True Negatives (TN): 200 (spam emails correctly identified as spam)
Hence , Recall = TP / (TP + FN) = 730 / (730+ 20) = 0.973
And Precision = TP / (TP + FP) = 730 / (730 + 50) = 0.9358

3. The expected value for y is 250 when x = 12. I have attached all the working in the picture below (did the implementation and working using pen and paper)



c. Given table,

x	3	6	10	15	18
y	15	30	55	85	100

To find the regression line using least squares,

$$\hat{y} = ax + b$$

where

$\hat{y} \rightarrow$ predicted value

$b = \hat{y} - ax$

$$a = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x - \bar{x})^2}$$

$$= \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x2} - (\bar{x})^2}$$

	x_i	y_i	x_i^2	$x_i y_i$
	3	15	9	135
	6	30	36	1050
	10	55	100	5500
	16	85	225	19125
	18	100	324	32400
$\Sigma \rightarrow$	53	285	695	58210
Mean →	10.4	57	139	11642

Hence , $a = \frac{11642 - 57 \times 139}{139 - 108.16} = \frac{3719}{30.84} = 120.6$ (approx)

$b = 57 - 1254.24 = -1197.24$

Equation of line → $120.6x - 1197.24$

Hence when $x = 12$, $y = 249.96$

≈ 250 (approx)

4. The above situation where the empirical loss on f1 might be lower but it may not generalize better describes the situation of overfitting the model. For example, Let us take a simple classification model to predict the tumors as malignant and benign. There might be a possibility that f1 is a complex model with many parameters, such as a high-degree polynomial regression or a deep neural network with multiple layers. Due to its complexity, f1 can fit the training data very well, potentially even memorizing it leading to a low empirical loss. On the other hand, f2 might be a simpler model. While f1 might excel on the training data, it's more susceptible to overfitting.In contrast, f2 might have a slightly higher empirical loss on the training set, but its simplicity can help it avoid overfitting and generalize better to new data.