Name :- Soumitra Pandit

Subject :- DS502

HOMEWORK ASSIGNMENT 1

Ex 2.4

Q1.)

a.) The sample size is large and the number of predictors is small. This set up would be ideal for a more inflexible statistical learning method. We are actually considering two variables when we make that assessment. The first is the sample size n. Now, in case of a smaller sample size, a flexible model would overfit and the variance would be off the charts. However, as we move towards a larger sample size, a decently flexible model would estimate the real function better than an inflexible or parametric method.

The second variable is the number of predictors p. This is a bit of a tricky situation. Having a low number of predictors increases the bias that is inherently present in the system. In such instances, when the bias-variance tradeoff is lopsided in favour of bias, over comitting to the given data by using a more flexible model can be dicey as we can easily disrupt the tradeoff. As the ideal falls in minimising both the variance and tradeoff, and given that the inherent bias of this setting is quite high, we should opt for an approach which nulls the variance. Hence a linear / parametric or another kind of a non flexible approach would reign supreme here as it would reach the ideal bias-variance tradeoff point. Flexible models, on the other hand, can overfit and then the model would have both - a high bias due to the low number of predictors and a high variance too.

P.T.O

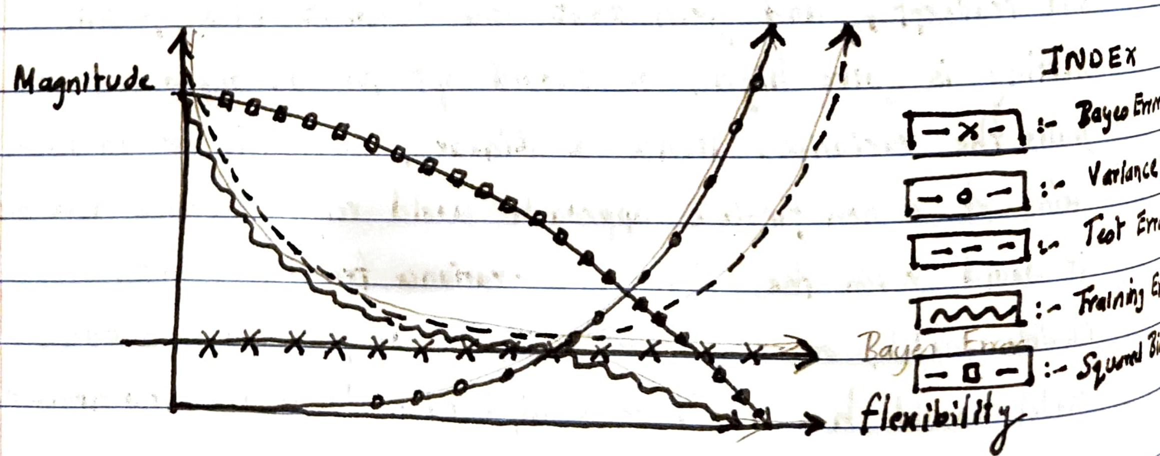
Ex 2.4

Q1.) b.) Even though the number of Predictors is large, we are unaware of their statistical significance and their capacity to account for the variance. Had all of the predictors been statistically significant with amazing R^2 values, a flexible model could have been considered as the inherent bias is minimal and all we are trying to do with the model (in that case) is reduce the variance. However, as that is not the case, using a flexible model would be risky. Also, the number of data points observed is low so an inflexible model would be the safest bet.

Q1) c.) As the relationship is highly nonlinear, an inflexible model would fail by a large margin as it would drive the bias through the roof. What we need is a flexible model that fits but doesn't overcommit.

Q1.) d.) Nothing we can do about that short of getting more predictors which are statistically significant I suppose.
But once again, an inflexible approach might be the way to go as any attempt to reduce bias might result in overfitting.

Q3.)



Q3)

b.) ① Baye's Error. :- Is independent of flexibility so it will remain constant. Now it does not have to be constant w.r.t to predictors but the flexibility of the model does not affect it.

② Variance :- Variance will increase as the mode gets more flexible. And I doubt that the relationship will be linear. Hence I have chosen an exponential curve. In all honesty though, I suspect that the degree of the exponent would be a function of the number of predictors; as the number of predictors increases, the ~~data is~~ model will fit more "tightly" with the data depending on its flexibility. This however is just a primary hunch.

③ Bias Squared :- The bias will decrease as the mode fits more snugly and once again, I have chosen an ~~logarithmic~~ exponential decay curve to depict that.

④ Test Error :- Test error will go down till we hit min (variance, bias) and then pick up again as the model starts to overfit.

⑤ Training Error :- Will go towards zero asymptotically, even surpassing the Bayesian Error line as the model overfits to the training data.

Q6.) A parametric method starts with an ~~is~~ unweighted model and then "adds weights" as the data is analysed. A non parametric model is not really a model at all, at least in the beginning. Eventually a model forms from the data informed by the smoothness required but there is no "initial shape".

P.T.O

⇒ Continued.

The advantages to regression or classification are simple - there is no real threat of overfitting. And the disadvantages, somewhat drawing from the previous lines is that the bias can never really disappear unless and until our parametric function happens to be of the same class as the function we are trying to model. So advantage is there is very low variance but the disadvantage is that the bias can be significant.

Ex 3.7

Q1.) Table 3.4 Demonstrates a linear Regression with multiple Coefficients. The P-Values determine if the null hypothesis can be rejected safely or not. For the given three predictors, the null hypothesis basically states that there is no correlation between ~~the~~ a given predictor and the response. The p-value is the probability of getting the values that have actually been observed given that the Null Hypothesis is true. Thus a very low P-value disproves the null hypothesis. (Usually, very low is anything less than 0.05).

↳ As can be seen in Table 3.4 T.V and Radio have P-Values lower than 0.0001, which indicates a very strong statistically significant relationship between these two predictors and the response. We will therefore reject the null hypothesis and go for the alternate hypothesis which says that there is a statistically significant relationship between T.V and Sales, and Radio and Sales.

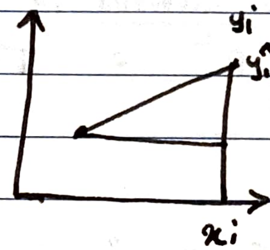
↳ Furthermore, it can also be seen that newspaper has a high P-Value ~~of~~ (0.86) which validates the Null hypothesis making the predictor moot in our further enquiry.

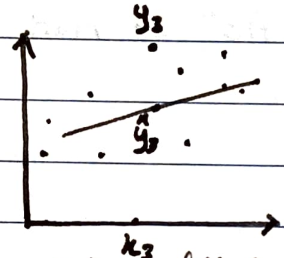
Q5.) $\hat{y_i} = x_i \hat{\beta}$

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i y_i\right) / \left(\sum_{i'=1}^{n} x_{i'}^2\right)$$

How does this work exactly?

$$\hat{\beta} = \frac{\hat{y_i}}{x_i}$$





$\hat{y_3} = x_3 \beta$

So wait how is this line getting fitted exactly?

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i y_i\right) / \left(\sum_{i'=1}^{n} x_i^2\right)$$

$$\hat{y_i} = \left(\frac{x_1 y_1 + x_2 y_2 + \cdots + x_{n-1} y_{n-1} + x_n y_n}{x_1^2 + x_2^2 + \cdots + x_{n-1}^2 + x_n^2}\right) x_i$$

And somehow, we want to get it in the form where y_i ~~is the only independent~~ is in summation. Okay.

$$\hat{y_i} = \frac{(x_1 y_1 + x_2 y_2 + \cdots + x_n y_n) x_i}{x_1^2 + x_2^2 + \cdots + x_n^2}$$

Now we can split this.

let $d = x_1^2 + x_2^2 + \cdots + x_n^2$

$$\therefore \hat{y_i} = \left(\frac{x_1 y_1}{d} + \frac{x_2 y_2}{d} + \cdots \frac{x_n y_n}{d}\right) x_i$$

$$\therefore \hat{y_i} = \left(\frac{x_1 x_i}{d}\right) y_1 + \left(\frac{x_2 x_i}{d}\right) y_2 + \cdots + \frac{(x_n x_i)}{d} y_n$$

$$\therefore a_{i'} = \left(\frac{x_{i'} x_i}{x_1^2 + x_2^2 + \cdots + x_n^2}\right)$$

Q6) Using (3.4) show that the least squares line always passes through the point $(\bar{x}, \bar{y})$

→ This is a cool problem. let me think.

Well, if $\bar{x}, \bar{y}$ belong to the regression line, then they should satisfy the equation of the line. i.e

$$\bar{y} = \hat{\beta_1}\bar{x} + \beta_0$$

lets see if thats true

$$\bar{y} = \hat{\beta_1}\bar{x} + (\bar{y} - \hat{\beta_1}\bar{x})$$

$$\bar{y} = \bar{y}$$

Thus proved. Quite Neat.