ENEL/ENSE 865: Applied Machine Learning

Professor: Dr. Abdul Bais

Instructional Designer/TA: Dr. Muhammad Hamza Asad

Submitted by:

Name: Nandish Bakulkumar Bhatt

Student ID: 200441204

Assignment:1

Q: 1 Given that,

Fuc^n:

$$J(w, w_o) = (y - Hw - w_o I)^T (y - Hw - w_o I) + \lambda w^T w.$$

→ Assume $\bar{x} = 0$, so the input data has been centered.

Optimizing the function, we get –

$$J(w, w_o) = \frac{1}{N} \sum_{i=1}^{N} (y_i - Hw - w_o I)^T (y_i - Hw - w_o I) + \lambda w^T w \quad (1)$$

→ For finding, gradient J, ∇J, as there are 2 variables. So, we will do partial derivative – First differentiating w.r.t. w_o,

$$\frac{\partial J(w, w_o)}{\partial w_o} = 0$$ [∵ Equating to zero for closed form solution =]

$$\therefore -2 \frac{1}{N} \left(\sum y_i - \hat{w}_o I \right) = 0$$

$$\therefore -2 \left(\frac{1}{N} \sum y_i - \hat{w}_o I \right) = 0.$$

$$\frac{1}{N} \sum y_i = \hat{w}_o I$$

$$\therefore \hat{W}_0 = \frac{\sum y_i}{N}$$

$$\therefore \boxed{\hat{W}_0 = \overline{y}}$$

→ Now, partial differentiating eq^n (1) w.r.t. W,

$$\frac{\partial J(w, w_0)}{\partial w} = 0 \quad \left[\because \text{Equating to zero for closed form solution}\right]$$

$$-2H^T(y - H\hat{w}) + 2\lambda I\hat{w} = 0$$

$$-H^T(y - H\hat{w}) + \lambda I\hat{w} = 0$$

$$-H^T y + H^T H\hat{w} + \lambda I\hat{w} = 0$$

$$\therefore \hat{w}(H^T H + \lambda I) = H^T y$$

$$\therefore \boxed{\hat{w} = (H^T H + \lambda I)^{-1} H^T y}$$

Hence, proved.

Q : 2

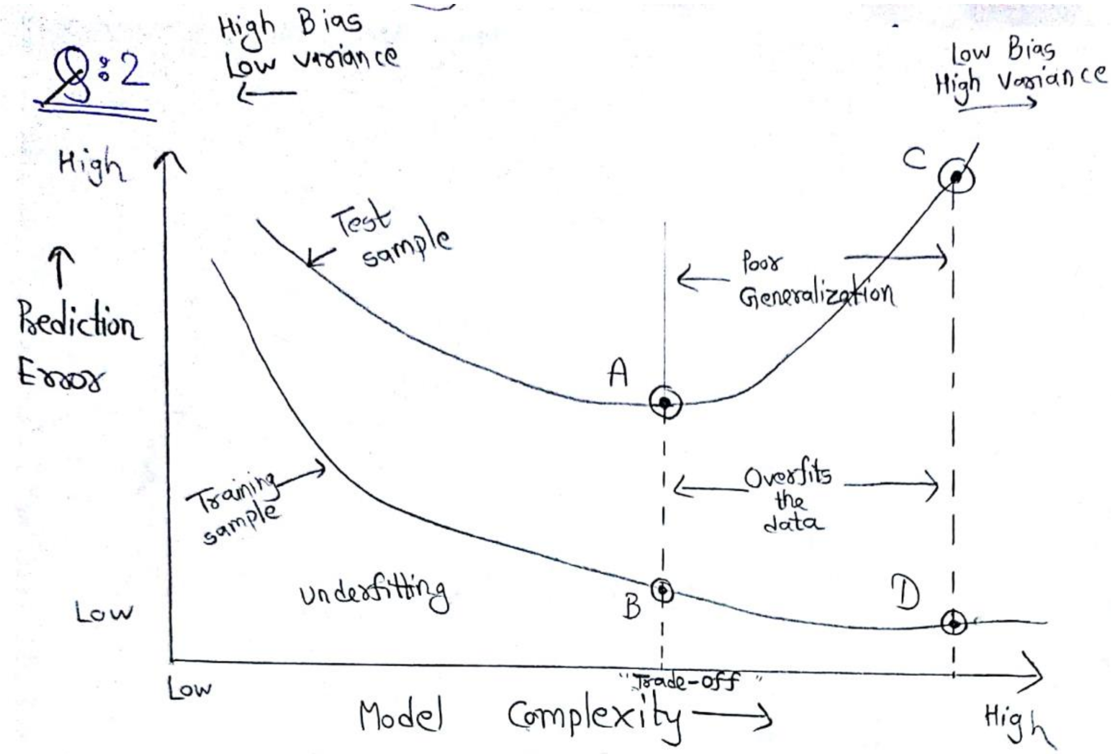


Fig (1)

→ Now, before explaining, first I will explain some of the terms which are there in the graph.

(1) Model Complexity :- Model complexity indicates the order of polynomial equation of model.

- If order increases, then model complexity also increases.
- So, Quadratic or cubic model is more complex than linear model.

(2) Prediction Error :- It is on the y-axis and is dependent variable.

- Prediction Error indicates the error of the training and testing data set.

(3) Bias :- Bias is a measure of flexibility of our model to capture the true model.

$$\text{Bias} = f_{w(\text{true})}(x) - f_{\bar{w}}(x)$$

$\bar{w}$ ↓ mean model.

(4) Variance :- Variance means variation from the expected fit or how scattered the data is.

→ Now, as shown in fig(i), there are two curves for training set data and testing set data.

→ As the model complexity increases, the training ~~set~~ error decreases after one point and testing error ~~incre~~ first decreases and after some point, it will start to increase again after that.

→ If we consider points "C" and "D" as shown in fig(i), then at those points, training error is less, but testing error is more. Thus, by this, we can say that model is good only, when training set is given. When testing data is given, it is not predicting the predicted output for the test data, and thus the test error increases more. Thus, training data accuracy is high, however testing data accuracy is low. This is known as Overfitting.

→ Now, if we take the lowest point in the test sample i.e., point "A", then corresponding to point "A", we have point "B" related to the training sample. Now, in this case, test error is low, ~~error~~ but training error is not low. Still, training error is at optimal point, not that much high. This point is called "Trade - Off" point.

→ Now, on the left hand side of Trade -off point, the model complexity is low, and prediction error is high for both training and testing set and thus we can say that region has High Bias and Low Variance. It is called Underfitting.

→ On the right hand side of Trade-off point, as the model complexity increases, the prediction error for training set is decreasing, but the prediction error for test set is increasing and thus it yields in Poor Generalization and it overfits the data so called overfitting.

Q:3

3(i)
→ Answering to the relation of test and train error to the three sources of error. (noise, bias and variance).

→ The

$$\text{Actual error} = \underbrace{\sigma^2}_{\text{noise}} + \underbrace{[\text{Bias}]^2}_{\text{model flexibility}} + \underbrace{\text{Variance}}_{\text{sensitivity to dataset}}$$

(i) Noise: It is irreducible error caused.
e.g., Predicting the market price, but due to some negotiation, the price of house might be decreased.

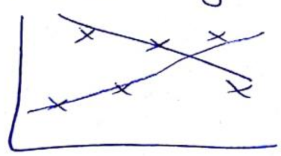
- standard deviation for this $= \sigma^2 = E(\varepsilon_i^2)$

(ii) Bias: $\text{Bias} = f_{w(\text{true})}(x) - f_{\bar{w}}(x)$

↓ mean model.

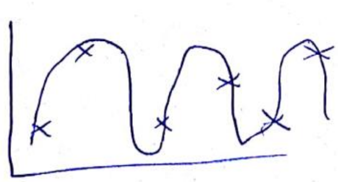
- It is a measure of flexibility of our model to capture the true model.
- Less-complex model yields low flexibility.

e.g., Linear Fit.



← High Bias
Less flexible, so low variance

15th order



← Less Bias
More flexible, so high variance

(iii) Variance:- Variance means variation from expected fit.

→ If variance is high, then flexibility is high, so high sensitivity to datasheet.

→ Now, relating to how these each data is changing with the increase in the number of datapoints is as explained below:-

→ As shown in figure of Q:3, the Mean Square Error (MSE) on test set by different models of different degrees vs N is plotted.

→ Now, the test error level comprises of 2 terms-

(a) Noise Floor: An irreducible component that all models incur, due to intrinsic variability of generating process.

(b) Structural Error: It depends on differences between the generating model ("truth") and the model.

→ Now, it is observed from fig, that, of degree 1, 2, 10, 25 let's assume, models be M-1, M-2, M-10, M-25.

→ Structual error for M-2 and M-25 is zero, as both are able to follow true generating method. Also, M-1 's structural error is noticeable, as the level occurs high above noise floor.

→ Moreover, test error will go to zero faster for simpler models, as there are very less parameters to estimate.. So, particularly for finite training set, there is difference b/w varibles that we estimate, and actual variables. This is called approximation error and it tends to zero as N tends to zero.

→ Now, from fig- (b) & (c), it is evident that with the increase in size of training set, the training set and testing data set approaches to be same, when size of training set exceeds than 80.

→ Whereas in fig (a), the training and testing set's MSE never become equal. So, it is less flexible and thus variance is low. and bias is high. It is underfitting data.

→ In Fig - (d), the training and testing set's MSE takes two different and it does not overlap any component. So, it is actually overfitting and thus it has Low Bias and high - variance.

→ In fig (b) i.e., when degree = 2, it is falling under best fit i.e., in Bias-variance trade off region.

Reference: Machine Learning - A Probabilistic Perspective. by Kevin P Murphy.

Q:4

→ The intercept (w_0) term in L1 and L2 regularization will not impact the complexity of model.

→ But, w_0 is only impacting the height of the function, which is not associated to overfitting. Thus, it should not be penalized.

→ Also, the input data features in Lasso as well as Ridge regression are normalised, then
- w_0 is not dependent on λ.
- Mean = 0
- Standard deviation = 1.

→ Thus, the cost function in L1 and L2 is modified to -

↳ L1 / Lasso Regression,

$$cost(w, w_0) = \frac{1}{N} \sum_{i=1}^{N} \left[y_i - (w_0 + wH(x_i)) \right]^2 + \lambda \|w\|_1 \quad (1)$$

↳ L2 / Ridge Regression,

$$cost(w, w_0) = \frac{1}{N} \sum_{i=1}^{N} \left[y_i - (w_0 + w^T H(x_i)) \right]^2 + \lambda w^T w \quad (2)$$

→ From above eq^n (1) and (2), it is observed that w_0 is not affected by value of λ.

→ Hence, input feature parameters and intercept (w_0) are normalized and does not affect on regularization by above two equations. Thus, intercept (w_0) is independent of λ.