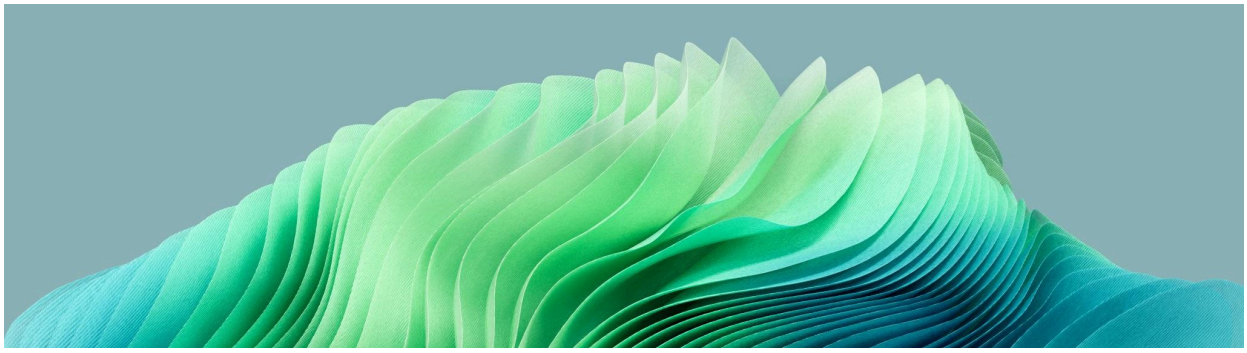

Conceptual Questions

Chapter 2 : Statistical Learning

Q1. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(a) The sample size n is extremely large, and the number of predictors p is small.

When n>>p, that means we have ample observation and the dimensions are very low meaning we won't be having the curse of dimensionality. A highly flexible model comes with high variance but this limitation is constrained and effectively managed , as the sheer amount of data helps the model to recognize generalized patterns that will help predict future values and not stray away from those to

random noise, stopping the case of overfitting. Hence, a more flexible model is a better than inflexible

(b) The number of predictors p is extremely large, and the number of observations n is small.

When p>>n, it generally leads to the case of overfitting in highly flexible models, as with more p the flexible method takes the form of polynomials with high dimensions like x^100,or x^300. With so many features the data is spread across very sparse or thin, making the requirement of more data exponentially high to cover the space, the space NEEDS to be covered as highly flexible models will learn not only the generalized patterns but also all the random noise patterns making them work exceptionally well on training data but fail on validation and test data, also called as curse of dimensionality
Methods relying on distances like K-nearest neighbors (KNN) also fail because of the thinly spread sparse data
Hence, when p>>n a highly flexible model leads to overfitting and curse of dimensionality making them a worse fit than inflexible statistical learning model.

(c) The relationship between the predictors and response is highly non-linear.

Highly flexible models are generally non-linear. They have more parameters and can therefore learn more complex, non-linear relationships in the data. This would make flexible models generally a better fit than inflexible methods.

(d) The variance of the error terms, i.e. sig = Var(e), is extremely high

MSE (Mean square error) = var(x) + Bias^2 + Var(e)

Here variance = E(x-mean)sq making this a positive value, Bias = E(fcap(x)-f(x)) hence squaring it ensures a non negative value , considering these we can say that the lowest value of MSE will always be Var(e) which is the irreducible error or noise that is out of our control.
Now according to our question, the variance var(e) is extremely high, making the lowerbound value of MSE extremely high, which we do not want as we always tend to make MSE value as low as possible. Moreover, flexible models already have high var(x) and now with high var(e) the overall MSE would be extremely high for test or validation dataset hence, Flexible statistical learning model will be worse than inflexible statistical learning model.