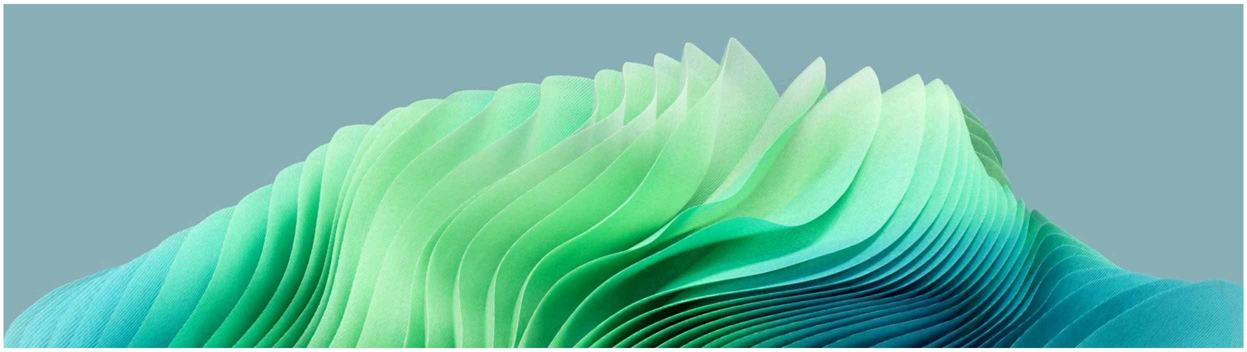


Conceptual Questions

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 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 (curse of dimensionality), making the requirement of more data exponentially high to cover the space, as it NEEDS to be covered because 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 overfitting.
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 models.

(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.

2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

This is a **regression problem** as CEO salary is a quantitative value, where we are more interested in **Inference**, here **n = 500 and p = 3** (profit, number of employees, industry), response variable= 1(CEO Salary)

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

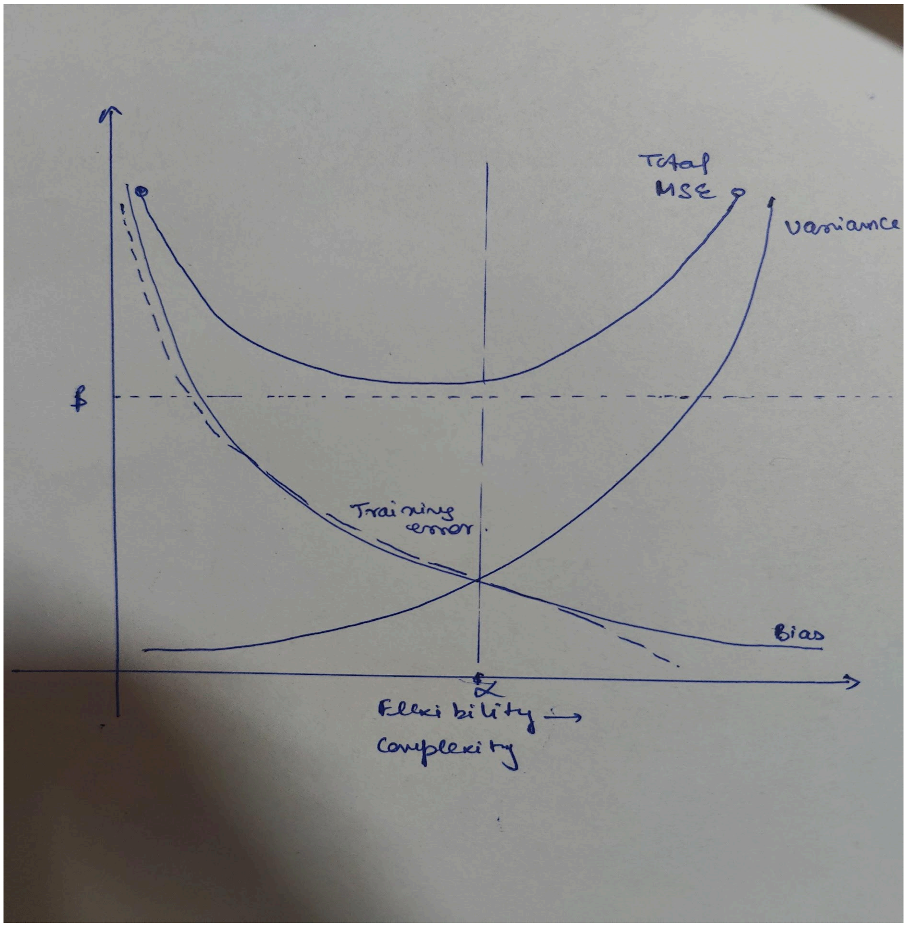
n=20, p=13, response variable =1 (Success or failure) ; this is also a **classification problem** but we are more interested in **prediction.**

(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market

n=52, p=3, response variable = 1(%change in USD/Euro exchange rate); we are interested in **prediction** but this is a r**egression problem** as we are predicting a continuous value rather than a class.

3. We now revisit the bias-variance decomposition.

(a) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.



There are 6 parts in this graph;

1. **Bias**2 :** - the curved continuous line which is first high then lowers down
2. **Variance :** contrary to Bias, it starts of low and ends up at higher value
3. **Irreducible error Var(e)**- displayed by the dashed horizontal line
4. **Test error/ MSE** : Continuous curve whose value goes down, then goes up at the end, it does not go down below irreducible error
5. **Training error:** displayed by dashed curve line which starts of high and then as flexibility increases it goes down
6. **Optimal complexity:** Denoted by the dashed vertical line where MSE curve at its lowest point.

(b) Explain why each of the five curves has the shape displayed in part (a).

Flex

We will go each curve one by one

1. **Bias**2-** Bias is defined as the difference between the curve estimated by the chosen statistical method and the actual true function (which inherently is a theoretical ideal function), Therefore whenever the flexibility or complexity of the statistical model increases, the model tends to incorporate more values, random noises of the dataset , aka it learns the

dataset more closely, which makes it resemble true function more hence reducing the Bias (this can also lead to overfitting)

2. **Variance-** This is defined as the sensitivity of the function or curve when different datasets are used, meaning , in ideal situations we would want variance to be as low as possible , we would want all the curves obtained from different datasets so be similar so as to obtain one generalized function that would represent all the training , validation and test dataset , but if introducing new values changes the curve a lot? That means there is a lot of variance in the statistical model.
    a. **Now looking at the diagram**, when the flexibility of the model is very low, the variance is very low, simultaneously when the complexity increases, so does variance as with more flexibility it is able to memorize complex patterns of the training dataset, so much that it can learn even minor intricacies in patterns, therefore introducing slightly different data points would then yield significantly different resulting model, leading to a high variance
3. **Irreducible Error -** This is the error that we can not control, which exists in real world, when considering the variance tradeoff equation which is **MSE (Mean square error) = var(x) + Bias^2 + Var(e),** we ideally want variance (curve sensitivity) and Bias(our model- true f) to be as low as possible , and since both the values are positive , the lowest they can ideally go is 0. Making MSE = var(e) ; or var(e) being the lowest value MSE can go , now when you look at the diagram , the MSE curve never goes bellow the dashed line.
4. **Test error/MSE- The error on Test dataset,** This curve is the combination of all the variance , bias and irreducible error curve , notice how it never goes below the irreducible error? Thats explained in the above para
5. **Training Error -** This is the average error on the training data, which starts of high on low complexity but when the model becomes more flexible , it learn the training data , as well as the noise more effectively, reducing the error and the bias to an inherent 0
6. **Optimal complexity -** This is the point where MSE curve is at the lowest

4. You will now think of some real-life applications for statistical learning.

(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

(c) Describe three real-life applications in which cluster analysis might be useful.