CONCEPTUAL QUESTIONS

1. 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.

Better – Since we have enough data to fit the model, and also it will overcome overfitting.

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

Worse – since the number of observations is small, highly flexible data may lead to overfitting.

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

Better – Highly flexible method will be able to better fit the non linear relationship.

(d) The variance of the error terms, i.e. σ2 = Var(), is extremely high.

Worse – More flexible methods give more variance.

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.

Regression , inference

N = 500, p = 4

(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.

Classification ,prediction

N = 20,p = 14

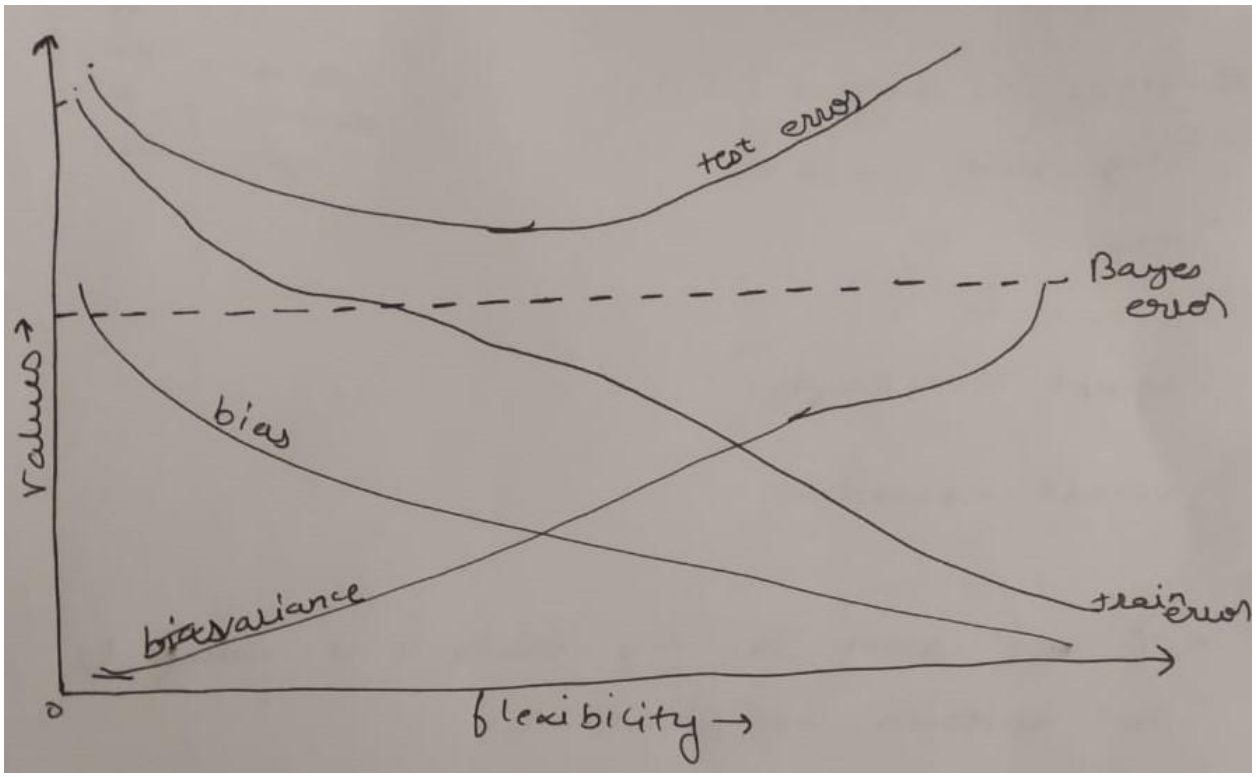
(c) We are interesting in predicting the % change in the US dollar in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the dollar, the % change in the US market, the % change in the British market, and the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

Regression ,prediction

N = 52(number of weeks in 2012), p = 4

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.



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

Variance – Variance has low value for lower flexibility and increases with increase in flexibility.

Bias – Bias has high value for lower flexibility and decreases with increasing flexibility.

Bayes Error – It is constant irreducible error.

Train error – As the flexibility increases the model fits in the patterns in the training data set.

Test error – First it decreases with the bias, this is the phase where model is fitting the relationship, after that it increases due to overfitting.

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.

- **Email Spam Detection – Given an email we want to classifify it as spam or not spam. The goal of the application is prediction.**
- **Cats and Dogs classification – Given an image, we want to classify whether the image is of a cat or a dog. The goal here is prediction.**
- **Detecting cancer, or some other diseases, given reports and history of the patient. Here we are looking for both prediction and inference.**

(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.

- **Prediction employees salary given one's profile. This can be used for both prediction as well as inference.**
- **Getting to know house prices, given features of the house and how the price is related to the different features available. This also can be used for both prediction and reference.**
- **Predicting test score of students given their past performances. This situation is quite useful for the pandemic situation as the colleges are passing the situations without exam, it can be used to decide the grades.**

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

- **Clustering articles in groups like sports, politics, and science.**
- **To Find the simlilarities in a data of patients.**

- **For any data we can apply clustering to know the subgroups in the data which are more related to one another.**

5. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

More flexible approaches are better able to fit the non linear relationship, and usually provide better results than prediction when used for prediction

We will favor flexible approaches over less flexible in the cases when our sole aim is prediction.

A less flexible approach is suitable for inference problems. When we are interested in knowing how the target is related to the given labels.

6. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

In parametric approach we approximate f (true function) by assuming some kind of relationship between the dependent and independent variables. Whereas in non-parametric approach we don't have any such assumption about the relationship between dependent and independent variables.

The advantages in parametric search are that it reduces the space for the search to approximate f. Assuming a linear relationship our search is now reduced to finding the appropriate p+1 coefficients (p is number of independent variables). Finding p+1 coefficient is way easier than finding a p dimensional function.

The disadvantage of using it is in case if the assumed relationship differs a lot from true relationship, then the estimates produced will be poor.

7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Suppose we wish to use this data set to make a prediction for Y when X1 = X2 = X3 = 0 using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.

First – 3

Second –2

Third – 3.16

Fourth – 2.23

Fifth – 1.414

Sixth – 1.73

(b) What is our prediction with K = 1? Why?

For K=1, the nearest neighbor is fifth data point, So the prediction is GREEN

(c) What is our prediction with K = 3? Why?

For K=3,The three nearest neighbor are fifth,sixth, and second data point. Two of the three data points have label as Red, So, the prediction will be RED.

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

For highly non-linear boundary we would choose a large K, Because as we increase the value of k, the non linearity also increases.