Knowledge check 1.1 :-

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 is extremely large, and the number of predictors p is small.

Ans :-

In this case the better performance is flexible statistical learning comparing inflexible statistical learning method.

Leaving predictors p is extremely large, and the number of observation n is small.

Ans:- The performance of a flexiable statistical learing method is worst because of overfitting is very high

c) The relationship between the predictor and response is highly non-linear

Ans:- The Performance of a flexible statistical learing method is better because of normal distribution.

d) The varience of the error terms, i.e $\sigma^2 = var(\epsilon)$ is extremely high.

Ans:- The Performance of a flexiable statistcal learining is worse when the varience term is very high.

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.

Ans:-

* Regression problem:-

Because of this scenario is a continuous variable.

* 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

collect ~~data on~~
launched - for each products that were previous launched. For whether it was success or failure, price charged for the product, marketing budget, competition price and ten other variable.

Ans!-

* classification problem!-

Because of this scenario is a categorical variable.

* Prediction
* n = 20, P = 13

c) We are intrested in predicting the y. change in the USD/Euro exchange rate in relation to the weekly changes in the world stock market, Hence we collectly weekly data for all of 2012. For each week we record the y. change in the USD/Euro, The y change

in the US market, the Y. change in the british market, and the Y. change in the German market.

Ans!-

* Regression problem
* Prediction
* n=52, p=4

3) a) Describe three real-life application in which regression migth be useful. Describe the responce, as well as the prediction. Is the goal of each application inference or prediction. Explain your answers.

Ans!:

* Predicting the marks of a student based on the nuber of hours he/she put into the preparation.
* Predicting and inference the whether reports.
* Movie rating before directing the different region.

b) Describl three real-life application in which classification might be useful, describe the response as well as the prediction. Is the goal of each application inference or prediction? Explain

Ans:-

*Classify it is spam or not (Email)

* The goal in to identify images of single digit 0-9 correctly (handwritten digit Recognition)

* Identify disease or tissue types based on the gene Expression levels.

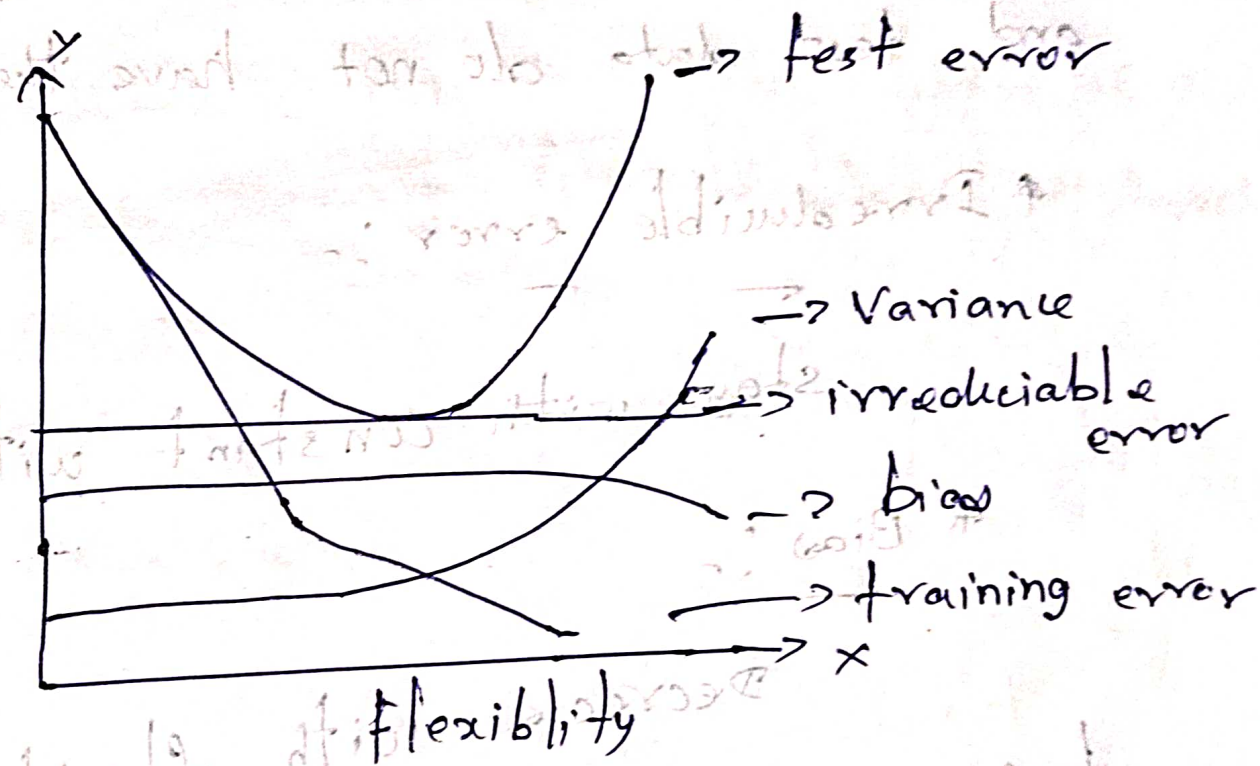
Knowledge Check 1.2:-

1) we how revisit the bias - variance decomposition

a) Provide a sketch of typical bias, variance, training error, test error, and bayes error curves, on a single plot, as we go from flexiable statistical learning method towards more flexiable approach. The x-axis should represent the amount of flexibility in the method.

and the y-axis should represent the values for each. There should be five make sure to label each one.

Ans!:-



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

Ans!:-

* Training Error!-(80% or 70%)

Decress with flexibity - possiable to better follow the data with more

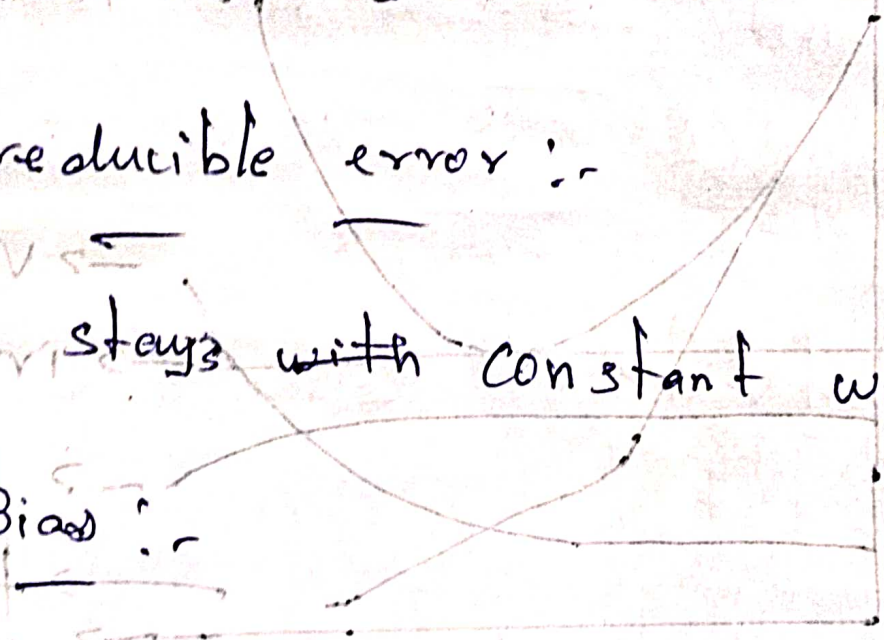
* Test Error :- (20% or 30%)

Decress and the increases with flexibility, error increses because model is following of data in training set and and test data do not have the same noi

* Irreducible error :-

stays constant with the metho

* Bias :-



Decreases with flexibility because more libely to appropriately fit the data.

* Variance :-

Increases with flexibity more wobbly, follows the data more.

2) What are the advantages and disavantages of a very flexiable approach for regression or classification under what circumstances might a more flexiable approach be prefrred loss flexiable approach? when might a less flexiable approach be preferred?

Ans :-

Advantage :-

* Non - Linear data
* Less biad
* Variable Intractions

Disadvantage :-

* Lack in interpretability
* high in Varience.

Why take a more or less flexible option?

Chose more flexiable when lots of data and many different

groups and chose less flexible when few data points for both regression and classification approaches.

3) 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? What are its disadvantages?

Ans:-

Parametric method:-

This method make an assumption about the function of the model and that it is linear.

Non-parametric method:-

This methods do not assumes anything about the function when trying to estimate the fit of the data.

Advantages :-

* Linear
* easy to intrepret
* eeay to defit

Disadvantages :-

* non-Linear
* No easy to intrepret and defit.

4) The table below provides a training data set containing six observations, three predicators and one qualitative response Variable.

	x_1	x_2	x_3	
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

suppose we wish to use this data set to make a prediction for Y when $x_1 = x_2 = x_3 = 0$ using K-nearest neighbors

a) Compute the Euclidean distance between each observation and the test point, $(x_1, x_2, x_3) = (0, 0, 0)$

Ans:-

$x_0 = 0, 0, 0$

1 $\Rightarrow \sqrt{(0-0)^2 + (3-0)^2 + (0-0)^2} = \sqrt{9} = 3$

2 $\Rightarrow \sqrt{(0-0)^2 + (0-0)^2 + (0-0)^2} = \sqrt{11} = 2$

3 $\Rightarrow \sqrt{(0-0)^2 + (0-0)^2 + (0-0)^2} = \sqrt{10} = 3.16$

4 $\Rightarrow \sqrt{(0-0)^2 + (1-0)^2 + (2-0)^2} = \sqrt{5} = 2.23$

5 $\Rightarrow \sqrt{(-1-0)^2 + (0-0)^2 + (1-0)^2} = \sqrt{2} = 1.41$

6 $\Rightarrow \sqrt{(1-0)^2 + (1-0)^2 + (1-0)^2} = \sqrt{3} = 1.73$

b) What is our prediction with $k=1$? Why?

Ans:-

Green

c) What is our prediction with $k=3$? Why?

Ans:-

d) If the Bayes decison boundry in this Problem is highly non-Linear, then would we expext the best Value for k to be large or small? why?

Ans!:-

Small Value