Machine learing.

Knowledge check 1:1 :-

1. For each of parts (a) through (d), indicate whether wether we would generally expect the performance of a flexible statistical learning method. Justity your answer.

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

Ans:-

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

b) the number of predittors p is extremely large, and the number of observations n is small.

Ans:-

The performance of a flexible statistical learning method is worst because of overfitting is very high.

c) The releationship between the prredictors and response is hishly non-liern.

Ans:-

The performance of a flexible statistical learning method is better because of normal distribu

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

Ans:-

The performance of a flexible statistrical learning is worse when the varience term is very high.

2) Explain whether each scenario is a classification or Regression problems, and indicate whether we are most intrested in Intorence or prediction. Finally provide n and P.

a) We collect a set of data on the top 500 frims in the us for each Forms we Record, profits, number of Employes, and the CEO salary. We are intrested in understanding which factor affect CEO salary.

Ans :-

Regression problems. Because of this scenario is a continious variable.

* Interence

* $n = 500$, $P = 4$

b) We are considaring launching a new product and wish to know whether it will be success or failure. We collect data on 20 similar products

The were previash launched for each product we have beorded whether it was a success or failure, price charged for the product, marketting budget conpetio price and ~~the~~ ten other varicabels.

Ans:-

* clasitication problem, because of this scenario is a categorical variable.

* prediction.

* n = 20, P = 13.

c) we are intrested in the predicting % . chang in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collectly weekly data for all of 2012 For each week we record the % in the USD/ Euro chang in the use Market, and the % change in the German market.

Ans :-

* Regression problem.
* prediction
* n = 52. p = 4

3) a) Discribe 3 real - lite Application in which Regression might be useful. Describe The Response as well as the predictors is the goal of each application Explain interence or predication your answer.

Ans :

* predicting the marks of a student Based on the number of hourse he/she put into preparation.

* predicting interence the interenec Reports.

* Movie Rating betore ask, Directi The Ditteomce region.

b) Discribe the real-lite applicatio in which classitication might be useful. Discribe the respons as well as predictors. Is the goal ot each application interence or prediction ? Explai your answer.

Ans :-

* classity it is spam or not Email.

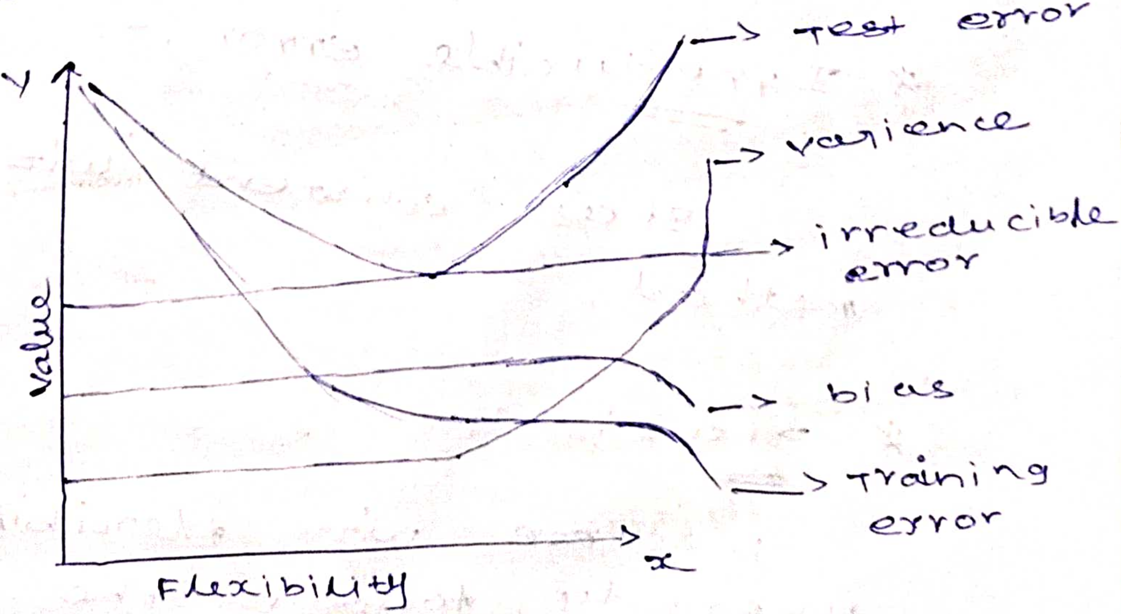
* The goal is Identity image ot single digit 0-9 correctly (hand written digit Recogonitio

* Identity disease or tissue type based on the goan expression Levels.

knowledge check 1.2

1. we now revisit the bias-varience decomposition.

a) provide a sketch of typical (sauariel) bias, varience, training error, test error, and payers errror curres, on a single plot, as we go from less flexible statistical learining method towards more flexible approches the x-axis should represent the amound of flexible in the method, and the y-axis values should resprent the values for each curres. there should be five cwrres, makes sure to lable each on.



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

Ans:-

* Treaing Error:- (80 % or 70 %)

Decreass with flexcibity - possible to be better follow the data with more flexible more.

* Test error: (20 % or 30 %)

Decreases and then increas with flexcibility error increase because model is following nois of data in traning set and test Data Do not have the same noise.

* Irreducible error:-

stay constent with the method.

* Bias:-

Decreee with flexcibitity more the data

more. appropriately fit the data.

* Increase with flexibitity more unsteddy. followwer the data more.

2. what are the advantages and disadvantages of a very flexible approch for regression or classification under what cricumstancess might a more flexible approch be preferred to a less flexible apprach? when might a less flexible approch be preferred?

Ans:-

Advantage:-

* Non Linear data

* Less bias

* variable intractions.

Disadvantage:-

* Lack in interpretability.

* high in varience.

Why take more or less flexible option?

more flexible when bts of data and many differnat groups and chose less flexible when few data poing for Both Regression and classification approches

3) Describe the Difference between a parametric and a non-paramtri statistical learning approch what are the advantage of a parametric approach to regression or classfication? what are its disadvatages?

Ans :-

Parametric method :-

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

Non - parametric method :-

This method do not assume anything about to funection when trying to estimate the fit of the data.

Advantages :-

* Linear
* easy to intreepret
* easy to detit

Disadvantage :-

* non - Linear
* no easy to intrepret and

ditit.

4) The table below provides a traning data get continning six observation, three predictors and one qualitative response variable.

	x_1	x_2	x_3	Y
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 $x_1 = x_2 = x_3 = 0$ usig k nearest heighbors

a) compute the Evelidean disadva between each observation and the test point, $(x_1, x_2, x_3) = (0,0,0$

Ans :-

$$1 = \sqrt{(0-0)^2 + (8-0)^2 + (0-0)^2} = \sqrt{9} = 3$$

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

$$3 = \sqrt{(0-0)^2 + (1-0)^2 + (2-0)^2} = \sqrt{10} = 3.16$$

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

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

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

b) what is out prediction with $k = 1$? why?

Ans:-

Green

c) what is out prediction with $K = 3$? why?

Ans:-

Red.

d) It the Buyers desision boundary in this problem is highly non lienear, then would we expect the best value for k to be Large on small?

why?

Ans:

small value.