

knowledge check 1.1

1) For each of parts (a) through (d), indicate whether we would 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 this extremely large and the number of predictors P is small.

Answer :-

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

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

Answer:-

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

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

Answer :-

The performance of the flexible statistical learning method is better because of normal distribution.

d) The varients of error terms. i.e $\sigma^2 = \text{Var}(E)$ is extremely high.

Answer :-

The performance of a flexible statistical learning is worst when the varients item 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 firms we record, profits, number of Employees, and the CEO Salary. we are interested in understanding which factor affect CEO Salary.

Answer :-

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 success or 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.

Answer :-

* classification problem, because of this scenario is a categorical variable

* prediction.

* $n = 20, P = 13$.

c) we are interesting in the predicting % 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 % in the USD/Euro change in the US Market, and the % change in the German market.

Answer :-

* Regression problem.

* prediction

* $n=52$, $D=4$

3) a) Describe 3 real-life Application 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.

Answer:-

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

* predicting inference the inference Reports

* Movie Rating before asking Directing the different region.

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

Answer :-

* classify it is Spam or not Email.

* The goal is Identify images of Single digit 0-9 correctly (handwritten digit Recognition).

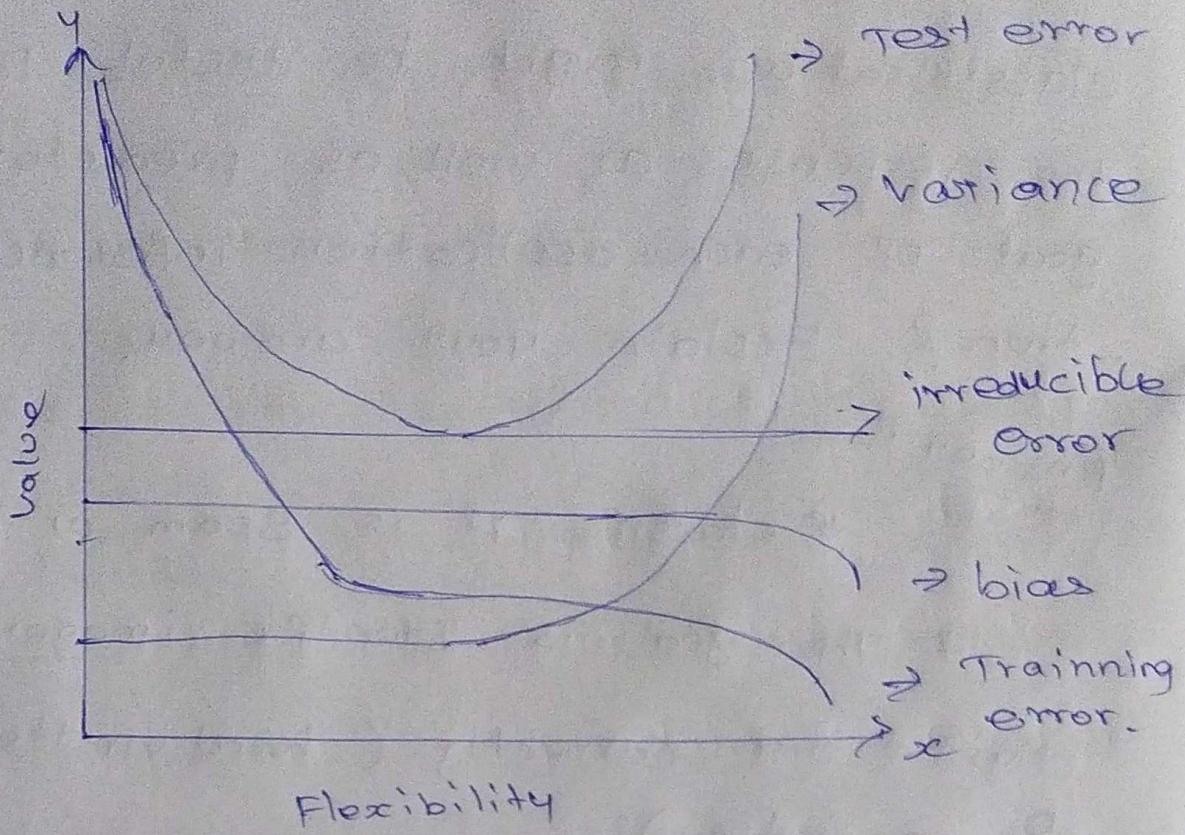
* Identify disease or tissue type based on the gene expression levels.

Knowledge check 1.2

1. we now revisit the bias-variance decomposition

a) provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes error curves, on a single plot, as we go from less flexible statistical learning method 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 curves, there should be five curves, make sure to label each one.

Answer:-



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

Answer:

* Training Error: (80% or 40%)

Decreases with flexibility - possible to be better follow the data with more flexible more.

* Test error: (20% or 30%)

Decreases and then increases with flexibility error increase because model is following noise of data in training set and test data do not have the same noise.

* Irreducible error:-

Stay constant with the method.

* Bias:-

Decrease with flexibility because more likely to appropriately fit the data.

* Variance:-

Increase with flexibility more Unstable, follows the data more.

2. what are the advantages and disadvantages of a very 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?

Answer:-

Advantage:-

- * Non Linear data
- * Less bias
- * Variable interactions.

Disadvantages :-

- * Lack in interpretability.
- * high in variance.

Why take more or less flexible option?

more flexible when lots of data
and many different groups and chose
less flexible when few data points for
Both the Regression and classification approach

3) describe the Difference 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?

Answer :-

parametric method :-

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

Non-parametric method:-

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

Advantages :-

- * Linear
- * easy to interpret
- * easy to define

Disadvantage :-

- * non-linear
- * no easy to interpret and define.

4) The table below provides a training data set containing 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 $x_1 = x_2 = x_3 = 0$ using k nearest neighbors.

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

Answer :-

$$x_0 \Rightarrow 0, 0, 0$$

$$1 = \sqrt{(0-0)^2 + (8-0)^2 + (0-0)^2} = \sqrt{64} = 8.$$

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

$$3 = \sqrt{(0-0)^2 + (1-0)^2 + (2-0)^2} = \sqrt{5} = 2.23.$$

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

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

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

b) what is our prediction with $K=1$? why?

Answer:-

Green

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

Answer:-

Red.

d) If the Buyers decision boundary in this problem is highly non linear, then would we expect the best value for K to be large or small? why?

Answer:-

Small value.