PRATEEK PARHI USC ID: 7461350213

(2.4.1) For each part (a) through (d) indicate wether 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?

Ans (a) Flexible Method works Better in this case than the inflexible method. The increase in Sample Size although might increase the variance of individual observation but overall variance around the Sample mean and population mean decrease. Also, the flexible Method will learn extra information from the data and will perform well. (NO Overfit)

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

Ans (b) Inflexible Method works better than the flexible Method. Because the Sample Size is small the overall variance between Sample Mean and Population mean is high. So, using a flexible Model may lead to Overfitting. Hence, we prefer an Inflexible Method.

(C) The relationship between predictors and response is highly non linear?

Ans (C) As we have prior Knowledge that the relation between predictor and response is highly non-linear therefore a flexible model would be usefull here. Flexible Model works better because the degree of freedom is high in this case and it also fits the data better than the Inflexible Model.

(d) The variance of error term, i.e $\sigma^2 = Var()$, is extremely high?

Ans(d) Inflexible model works better in ~~these~~ this case than the flexible model because the Noise (High error) will cause the flexible Model to overfit the data. Inflexible model will not be affected that much by the noise

(2.4.7) The table below provides a training data Set containing Six observations, three predictors, and one qualitative response variable.

Obs	X1	X2	X3	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 X1 = X2 = X3 = 0 using K-nearest Neighbors

(a) compute the Eucledian distance between each observation and the test point, X1 = X2 = X3 = 0

Ans(a)

Observation	Distance of observation
1	$\sqrt{(0-0)^2+(3-0)^2+(0-0)^2} = 3$
2	$\sqrt{(2-0)^2+(0-0)^2+(0-0)^2} = 2$
3	$\sqrt{(0-0)^2+(1-0)^2+(3-0)^2} = 3.162$
4	$\sqrt{(0-0)^2+(1-0)^2+(2-0)^2} = 2.246$
5	$\sqrt{(-1-0)^2+(0-0)^2+(1-0)^2} = 1.414$
6	$\sqrt{(1-0)^2+(1-0)^2+(1-0)^2} = 1.732$

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

Ans (b) observation 5 → $(-1,0,1)$ is the nearest to test data $x_1=0$, $x_2=0$ & $x_3=0$. So predicton with $K=1$ assigns $Y(x1, x2, x3)$ = Green. Thus green color is assigned to $x_1=0$, $x_2=0$, $x_3=0$

(C) what is our Prediction with $k=3$? Why?

Ans(C) The three ($K=3$) nearest neighbors of $x_1=0$, $x_2=0$ & $x_3=0$ are:

- observation 5: $(-1,0,1)$; Green
- observation 2: $(-2,0,0)$; Red
- observation 6: $(-1,1,1)$; Red

Since the Majority label among these three is Red, So $Y(x1, x2, x3)$ = Red. Thus we assign Red color to x_1, x_2, x_3 test data.

(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?

Ans(d) Since the bayes decision boundary is highly non linear so the high variance in the data will cause overfitting when k increases. Therefore best value of K in this case is expected to be Small.