HW2INF552 :

1. **ISLR 2.4.1**

(i). The sample size n is extremely large, and the number of predictors p is small.

Answer : *Better.* In this scenario a flexible method would outperform the inflexible method. Flexible method would be able to extract more information from the large vale of n, and this would be very helpful in overcoming the problem of overfitting.

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

Answer: *Worse.* Since the number of observations are small, a flexible method would have high chances of incurring overfitting and incurring noise.

(iii). The relationship between the predictors and response is highly non-linear.
Answer :*Better. Inflexible method are not efficient in responding to non-linear relationship.*

(iv). The variance of the error terms, i.e. σ2 = Var(e), is extremely high.
Answer : *Worse. As variance is high, it implies that the sample has considerable amount of noise in the relationship. Hence, an inflexible method would be less likely to overfit the data. <noise>.*

2. ISLR 2.4.7
(i) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.

Answer: Observation 1 has Euclidean Distance sqrt[(0 - 0)^2 + (3 - 0)^2 + (0 - 0)^2] = 3.
Observation 2 has Euclidean Distance sqrt[(2 - 0)^2 + (0 - 0)^2 + (0 - 0)^2] = 2.
Observation 3 has Euclidean Distance sqrt[(0 - 0)^2 + (1 - 0)^2 + (3 - 0)^2] = sqrt[0 + 1 + 9] = sqrt[10] = ~3.16.
Observation 4 has Euclidean Distance sqrt[(0 - 0)^2 + (1 - 0)^2 + (2 - 0)^2] = sqrt[1 + 4] = sqrt[5] = ~2.24.
Observation 5 has Euclidean Distance sqrt[(-1 - 0)^2 + (0 - 0)^2 + (1 - 0)^2] = sqrt[1 + 1] = sqrt[2] = ~1.41.
Observation 6 has Euclidean Distance sqrt[(1 - 0)^2 + (1 - 0)^2 + (1 - 0)^2] = sqrt[1 + 1 + 1] = sqrt[3] = ~1.73.

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

Answer : Prediction with k = 1 will be Green. Since the nearest neighbor to test point (0, 0, 0) is Obs 5 (-1, 0, 1), and observation 5 has a code of Green. We predict that the test point will be green too. Also, by the calculation :

$P(Y=\text{Red}|X=x0)=11\sum i\in\mathcal{N}0I(yi=\text{Red})=I(y5=\text{Red})=0$

$\mathbf{P(Y=Green|X=x0)=11\sum i\in ?0I(yi=Green)=I(y5=Green)=1}$

So prediction is green.

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

Answer: **Prediction with k = 3 will be Red.** The nearest three neighbors to test point (0, 0, 0) are Obs 5, Obs 6 (with distance ~1.73), and Obs 2 (with distance 2). Since two out of three are red, we predict that test point will be red. So by calculation :

$P(Y=\text{Red}|X=x0)=13\sum i\in\mathcal{N}0I(yi=\text{Red})=13(1+0+1)=2/3$

$P(Y=\text{Green}|X=x0)=13\sum i\in\mathcal{N}0I(yi=\text{Green})=13(0+1+0)=1/3$

(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?

Answer : The best value for the k will be small. With a large value of K, the boundary becomes smoother i.e. inflexible (linear), so we would expect the best value for k to be small.