**INF552 HW2**

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

Better. With the larger dataset, the flexible model can extract extra information to fit the original data. It can fit closer and reduce the overfitting problem.

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

Worse. The flexible model may cause overfitting with small number of observations.

**(c) The relationship between the predictors and response is highly non-linear.**

Better. With more degrees of freedom, a flexible model can better describe the non-linear relationships between predictors and response. The inflexible model is not appropriate for non-linear relationships.

**(d) The variance of the error terms, i.e. σ2 = Var(e), is extremely high.**

Worse. High variance data will cause noise in the flexible model, which will cause error and a worse performance. While the inflexible model is less sensitive to 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 Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.**

Distance of Obs1 = sqrt[(0 - 0)^2 + (3 - 0)^2 + (0 - 0)^2] = 3

Distance of Obs 2 = sqrt[(2 - 0)^2 + (0 - 0)^2 + (0 - 0)^2] = 2

Distance of Obs 3 = sqrt[(0 - 0)^2 + (1 - 0)^2 + (3 - 0)^2] = sqrt[0 + 1 + 9] = sqrt[10] = 3.16

Distance of Obs 4 = sqrt[(0 - 0)^2 + (1 - 0)^2 + (2 - 0)^2] = sqrt[1 + 4] = sqrt[5] = 2.24

Distance of Obs 5 = sqrt[(-1 - 0)^2 + (0 - 0)^2 + (1 - 0)^2] = sqrt[1 + 1] = sqrt[2] = 1.41

Distance of Obs 6 = sqrt[(1 - 0)^2 + (1 - 0)^2 + (1 - 0)^2] = sqrt[1 + 1 + 1] = sqrt[3] = 1.73

**(b) What is our prediction with K = 1? Why?**

The nearest neighbor to test point (0, 0, 0) is Obs 5 (-1, 0, 1). When k = 1, we predict that the test point (0,0,0) will also be Green.

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

The most three nearest points to test point (0, 0, 0) are Obs 5, Obs 6 and Obs 2. The Observation of them are Green, Red, Red, we choose the majority class, that is Red.

**(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?**

As K becomes larger, the boundary becomes inflexible due to the high variance. So the best value for K would be small.