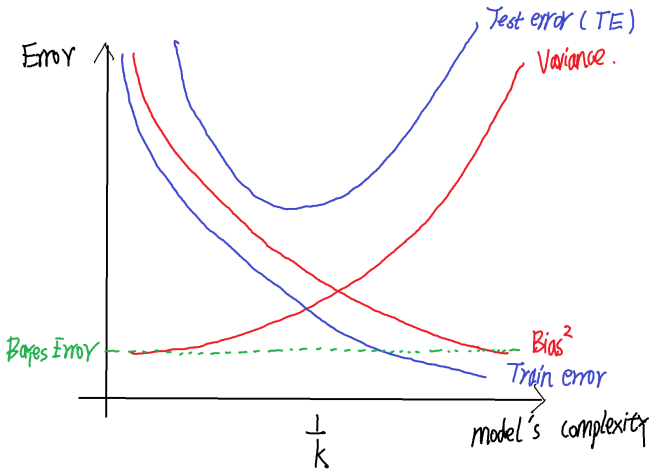
HW1

Yuan Cheng, Yurui Chang, Yuxin Cui

February 2021

1 Q1



2 Q2

a) Better. With larger sample size and smaller number of predictors, we can conclude that the points will be closer to each other so the flexible method will fit the data better and closer and for an inflexible method the error will be higher (performance will be worse)

b) Worse. With smaller number of predictors and smaller sample size, the distance between points will be larger so that the neighbors of x0 may not be very similar to x0 which will result misclassification and over fitting. As a result, an inflexible method will perform better.

c) Better. When the relationship between predictors and response is highly

nonlinear, such distribution of data is complicated so we need a flexible method to fit the data.

d) Worse. When the model is flexible enough, the errors will also be fitted and when variance of error is large, the flexible method's fitting will lead to larger testing error so we would prefer to use inflexible methods.

3 Q3

a) Dist1 $= \sqrt{(0-0)^2 + (0-3)^2 + (0-0)^2} = 3$

Dist2 $= \sqrt{(0-2)^2 + (0-0)^2 + (0-0)^2} = 2$

Dist3 $= \sqrt{(0-0)^2 + (0-1)^2 + (0-3)^2} = \sqrt{10} = 3.1622$

Dist4 $= \sqrt{(0-1)^2 + (0-1)^2 + (0-2)^2} = \sqrt{5} = 2.236$

Dist5 $= \sqrt{(0+1)^2 + (0-0)^2 + (0-1)^2} = \sqrt{2} = 1.414$

Dist6 $= \sqrt{(0-1)^2 + (0-1)^2 + (0-2)^2} = \sqrt{6} = 2.4495$

From the distances calculated above, we can conclude that when K = 3, we can choose the first three closest observations which are 5, 2, and 4. According to the given table, the prediction will be Green.

b) When the sample size is big enough and the Bayes classifier is highly nonlinear, we can conclude that the distribution of the data is very complicated so we intend to use some more flexible methods and for KNN, for example, we want to make sure the parameter K is small enough to ensure the model complex enough.