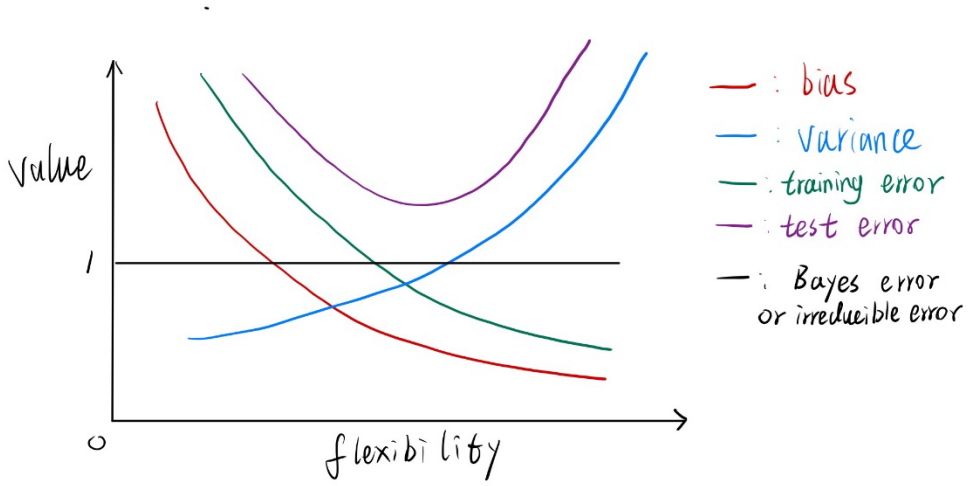
2.4 Exercises

3.

(a).



The relation among bias, variance, training error, test error, and Bayes/irreducible error curves with the flexibility of the model are showing above.

(b).

Bias: bias is the error from the difference of model and the function. A more flexible model will be more similar to real model, and the bias therefore decreases.

Variance: variance is the change of the model when the training set change. As the flexibility increases the variance will increase as well since the noise in the training set will be included in the model.

Training error: The training error is given by the average (squared) difference between the predictions of the model and the observations. As the flexibility increases this difference will decrease, because it will more similar to the real model.

Test error: for inflexible models, the bias is decreasing as flexibility increases until it reaches a minimum. Then the variance starts to increase and the test error starts increasing.

Bayes error / irreducible error: they are constant since by definition it does not depend on X and therefore on the flexibility of the model.

2.4 Exercises

7.

(a).

Euclidean distance:

d1 = sqrt(0^2+3^2+0^2) =3

d2 = sqrt(2^2+0^2+0^2) =2

d3 = sqrt(0^2+1^2+3^2) =sqrt(10)

d4 = sqrt(0^2+1^2+2^2) =sqrt(5)

d5 = sqrt((-1)^2+0^2+1^2) =sqrt(2)

d6 = sqrt(1^2+1^2+1^2) =sqrt(3)

(b).

For K=1, the nearest point is d5 and it is green, so the prediction is green,

(c).

For K=3, the nearest points are d5, d6 and d2, and the prediction is red, since d6 and d2 are red.

(d).

For highly non-linear boundaries, we would expect the best value of K to be small. Smaller values of K mean a flexible KNN model, and this will produce a decision boundary that is non-linear. A larger K would mean more data points are considered model and its decision boundary is closer to a linear shape.

3.7 Exercises

3.

(a). we can know that the model is:

$$\hat{Y} = \beta_0 + \beta_1 \times GPA + \beta_2 \times IQ + \beta_3 \times Gender + \beta_4 \times GPA \times IQ + \beta_5 \times GPA \times Gender = 50 + 20 \times GPA + 0.07 \times IQ + 35 \times Gender + 0.01 \times GPA \times IQ - 10 \times GPA \times Gender$$

So, we can know that iii is true. Only in the case of their GPA exceeds 3.5, males earn more on average than females.

(b).

According to the model, with IQ of 110 and a GPA of 4.0, Y = 137.1.

4.

(a). Cubic regression will have lower RSS. The cubic regression model is more flexible than the linear regression model since the cubic regression model can fit the real data better and get a lower training RSS.

(b). Linear regression will have lower RSS. Flexible models have less bias and higher variance. Since the true relationship between the predictor and the response is linear, the linear regression model will have low bias and a not-higher variance.

(c). Cubic regression will have lower RSS, and the reason is the same as (a).

(d). There is not enough information to tell. Because we don't know how far the real relationship is from linear, we don't know the bias of Cubic regression and Linear regression, in the result we cannot determine which model would have lower RSS.