Daniel Rance
dra142
49144736

# STAT 318/462: Data Mining Assignment 1

1. ***(4 marks) Describe one advantage and one disadvantage of flexible (versus a less flexible) approaches for regression. Under what conditions might a less flexible approach be preferred?***

   Fitting a more flexible model can be useful in getting closer to the true unknown form of f. If it is unclear which functional form f is following, a more flexible model will allow us to fit many different possible functional forms for f. However, too much flexibility can lead to a phenomenon known as *overfitting the data* where our estimated f does not follow the true unknown form of f. Rather, it follows the errors (or noise) too closely. Another disadvantage of a more flexible model is that, in general, fitting a more flexible model requires estimating a greater number of parameters. A less flexible approach may be preferred in conditions where we have strong indications that f is following a specific functional form, i.e., linear. Another time we might prefer a more restrictive model is when we are working with an inference problem and our main desire is interpretability. The more flexible a model is, the more complicated the estimate of f is such that it is difficult to understand how any individual predictor is associated with the response variable. Less flexible models are also preferred when there is a small amount of training data (small n).

2. ***(6 marks) Consider a binary classification problem Y ∈{0, 1} with one predictor X. The prior probability of being in class 0 is PR (Y = 0) = π0 = 0.69 and the density function for X in class 0 is a standard normal***

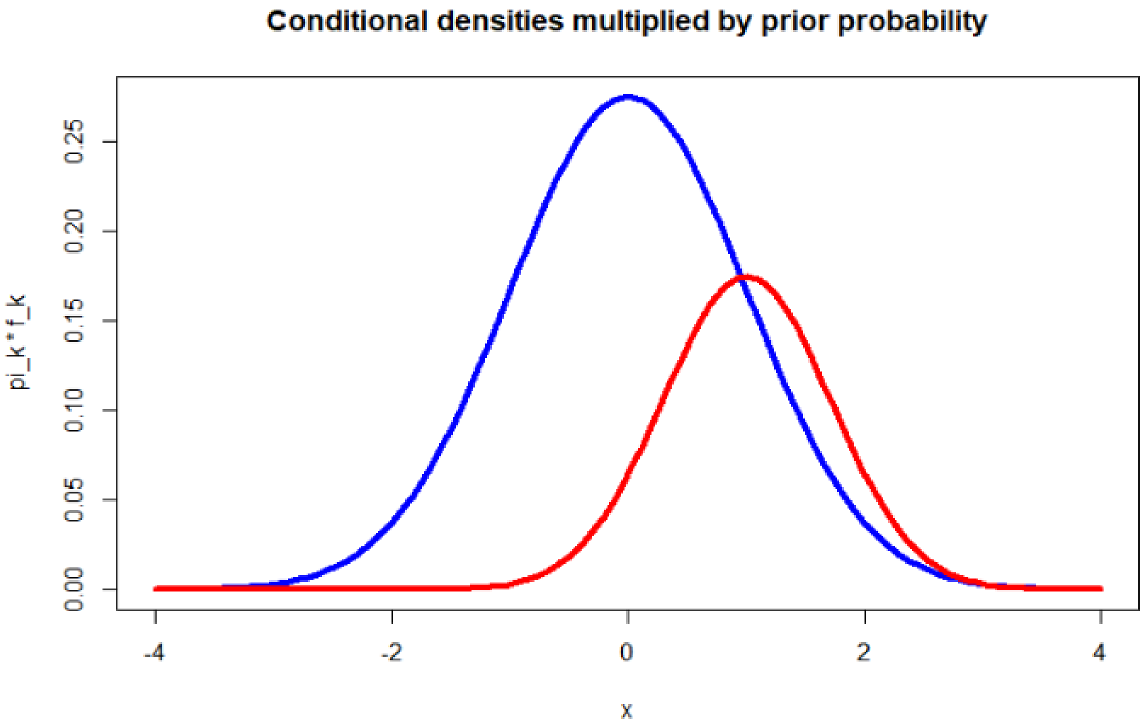
$$f_0(x) = \mathrm{Normal}(0, 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right).$$

   ***The density function for X in class 1 is also normal, but with μ = 1 and σ 2 = 0.5***

$$f_1(x) = \mathrm{Normal}(1, 0.5) = \frac{1}{\sqrt{\pi}} \exp\left(-(x-1)^2\right).$$

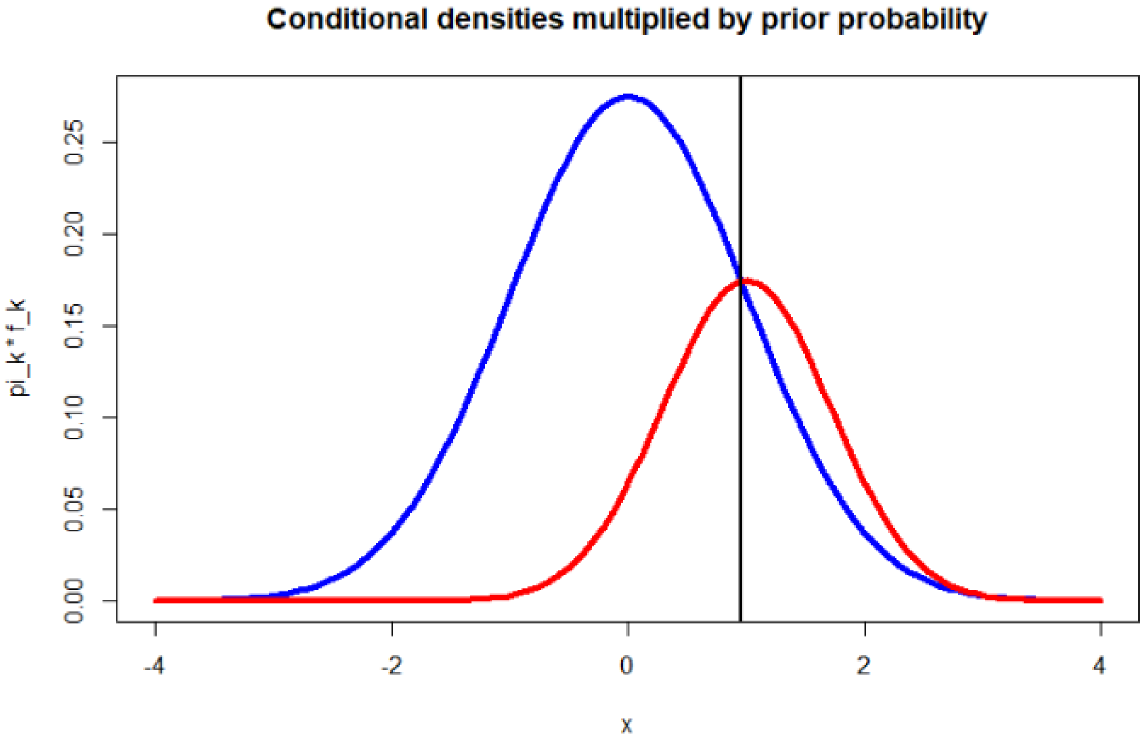
Daniel Rance
dra142
49144736

### (a) Plot π0f0(x) and π1f1(x) in the same figure.



### (b) Find the Bayes decision boundary (Hint: π0f0(x) = π1f1(x) on the boundary).

Note: the 'BDB' key present in the legend refers to Bayes decision boundary



### *(c) Using Bayes classifier, classify the observation X = 3. Justify your prediction.*

In this classification problem, our Bayes Decision Boundary is located at roughly X = 0.954. It will be referred to as X = 0.954 from here. Where X is greater than 0.954 we would consider the probability of Y = Class 1, given X, to be greater than 50%. Similarly, where X is less than 0.954, we would consider the probability of Y = Class 0, given X, to be greater than 50%.

We can calculate the probabilities for Y=0|X=3 & Y=1|X=3 as:

***Pr(Y=0|X=3) = $\pi_0 f_0$(x=3) = 0.003057975***
***Pr(Y=1|X=3) = $\pi_1 f_1$(x=3) = 0.003203383***
***Pr(Y=1|X=3) > Pr(Y=0|X=3)***

As the probability of Y being in class 1 given that X=3 is slightly higher than that of Y being in class 0, the classification of X=3 would be considered class 1. Therefore, we can conclude that the bayes classifier would classify X=3 as class 1.

### *(d) What is the probability that an observation with X = 2 is in class 1?*

The probability of Y = 1, given an X of 2 can be found using the following equation.

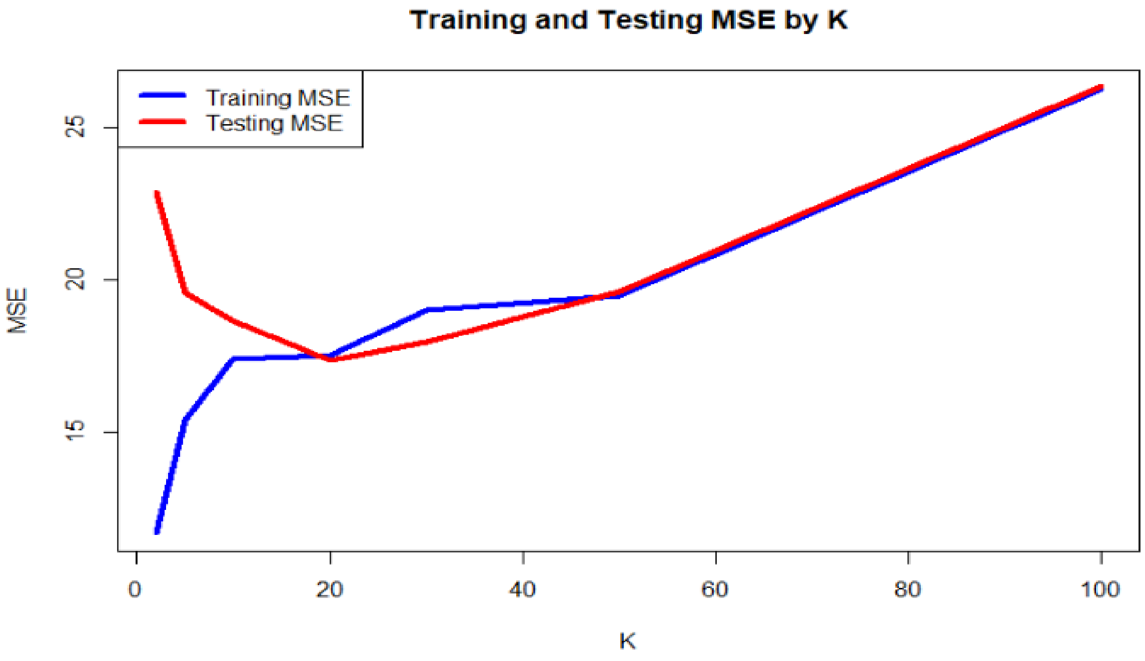
$$\Pr(Y = 1|X = 2) = \frac{\pi_1 f_{1(X=2)}}{\pi_0 f_0 + \pi_1 f_1}$$
$$\Pr(Y = 1|X = 2) = 0.6333126$$

Therefore the probability that an observation X=2 is in class 1 is $\mathbf{0.6333126}$.

**QUESTION 3 CONT. ON NEXT PAGE**

3. ***(8 marks) In this question, you will fit kNN regression models to the Auto data set to predict Y = mpg using X = horsepower. This data has been divided into training and testing sets: AutoTrain.csv and AutoTest.csv (download these sets from Learn). The kNN() R function on Learn should be used to answer this question (you need to run the kNN code before calling the function)***

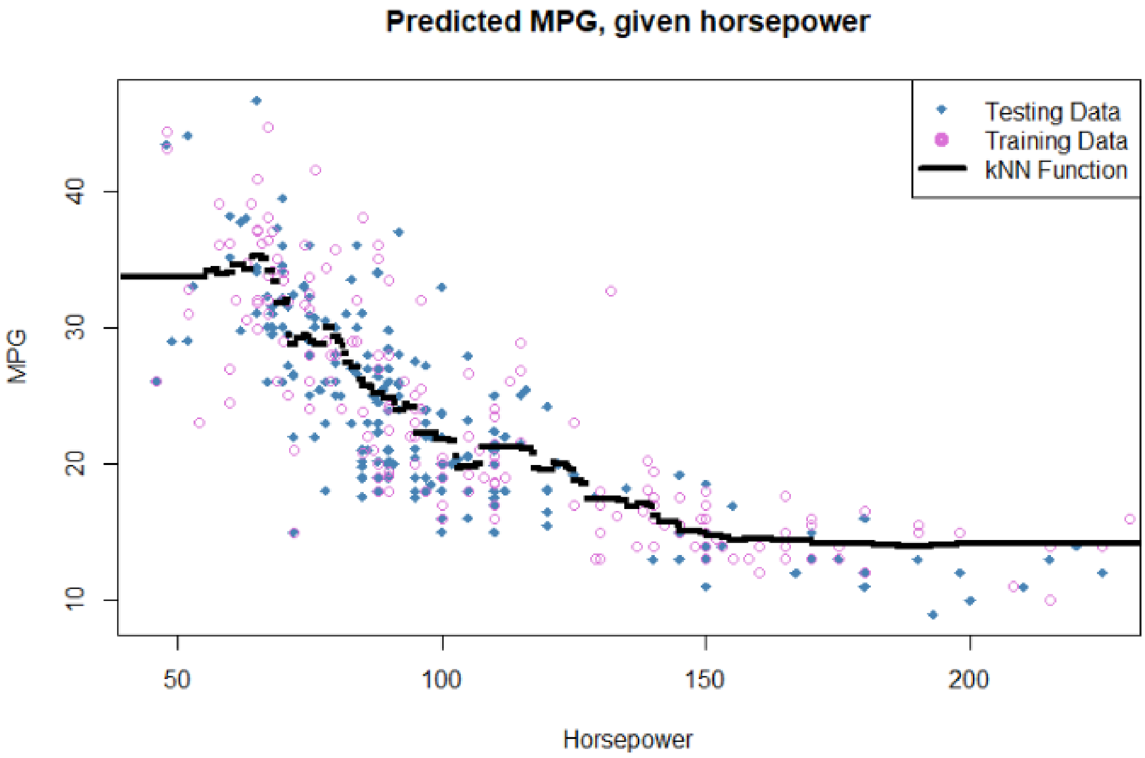
***(a) Perform kNN regression with k = 2, 5, 10, 20, 30, 50 and 100, (learning from the training data) and compute the training and testing MSE for each value of k.***



***(b) Which value of k performed best? Explain.***

In this classification problem, we would say that K = 20 was our best performing K value as it was at this value where the Testing MSE was at its lowest.

### *(c) Plot the training data, testing data and the best kNN model in the same figure. (The points() function is useful to plot the kNN model because it is discontinuous.)*



### *(d) Describe the bias-variance trade-off for kNN regression.*

Choosing the value of K is an important factor in the bias variance trade-off for kNN regression. At one end of the scale, we have an example where K is very large (relative to n). Let's say K=N. In this example, all responses given would be equal to the sample mean. Here we have a situation where our bias is very high (all responses skewed to a particular point) but our variance is very low (very little/no spread in response data. At the other end of the scale, we have a K that is very small (relative to n). Let's say K=1. In this example, all responses given have the possibility of being unique and, if the x predictor had been seen before in the training data, the response variable would be equal to the predicted variable. Here we have a situation where our bias is very low (data not skewed to any particular point) but our variance is very high (very high spread of response data). Finding the balance between these two extremes is the bias variance trade-off.