STAT 318/462: Data Mining

Xin Gao (43044879)

Assignment 1

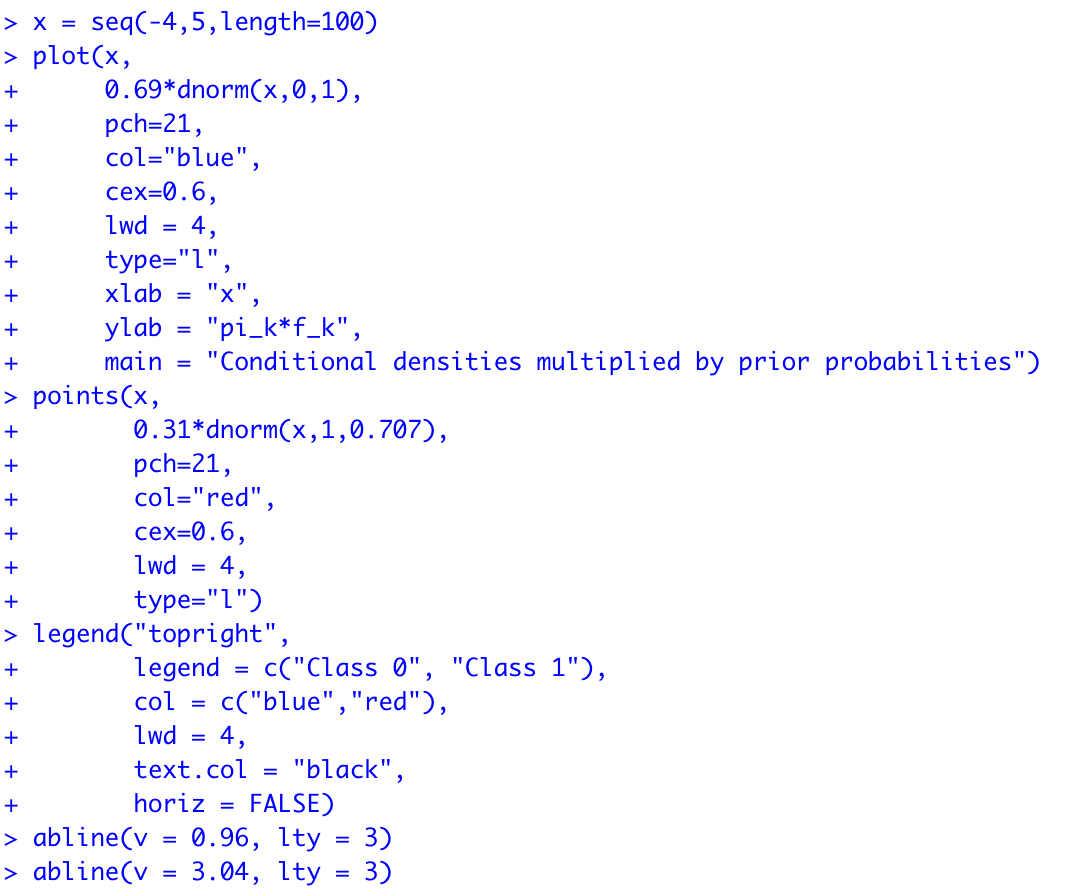
1. Describe one advantage and one disadvantage of flexible (versus a less flexible) approaches for regression. Under what conditions might a less flexible approach be preferred?

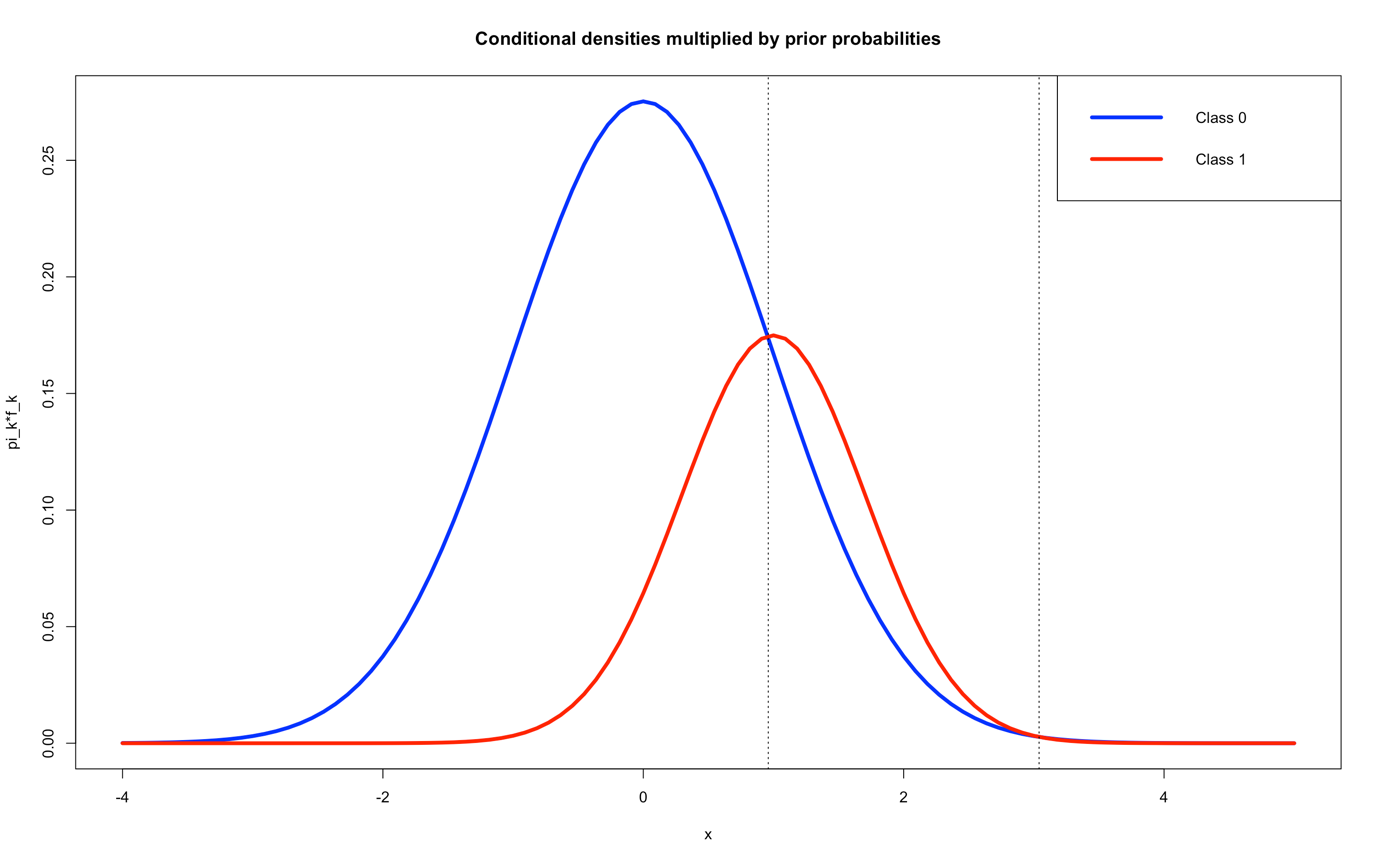
Advantage: More flexible approaches for regression can generate a much wider range of possible shapes to estimate function form, and obtain more accurate predictions.

Disadvantage: A very large number of observations is required, and easy to overfit which means following the errors, or noise, too closely. Very flexible approaches require a very large number of observations, and can lead to such complicated estimates that it is difficult to understand how any individual predictor is associated with the response, also harder to interpret.

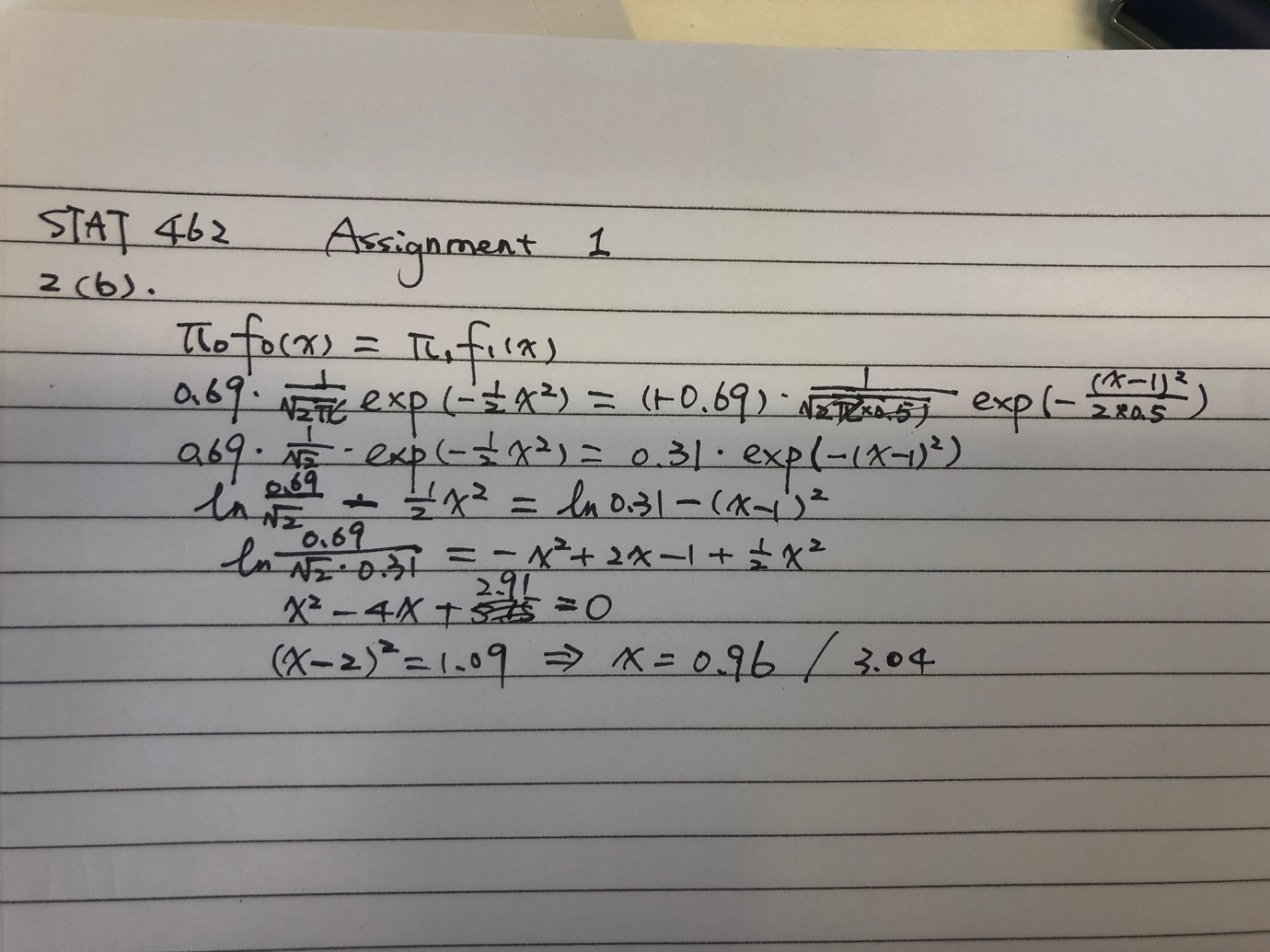
When inference is the goal, there are clear advantages to using simple and relatively inflexible statistical learning methods. In some settings, however, we are only interested in prediction, and the interpretability of the predictive model is simply not of interest, we might expect that it will be best to use the most flexible model available.







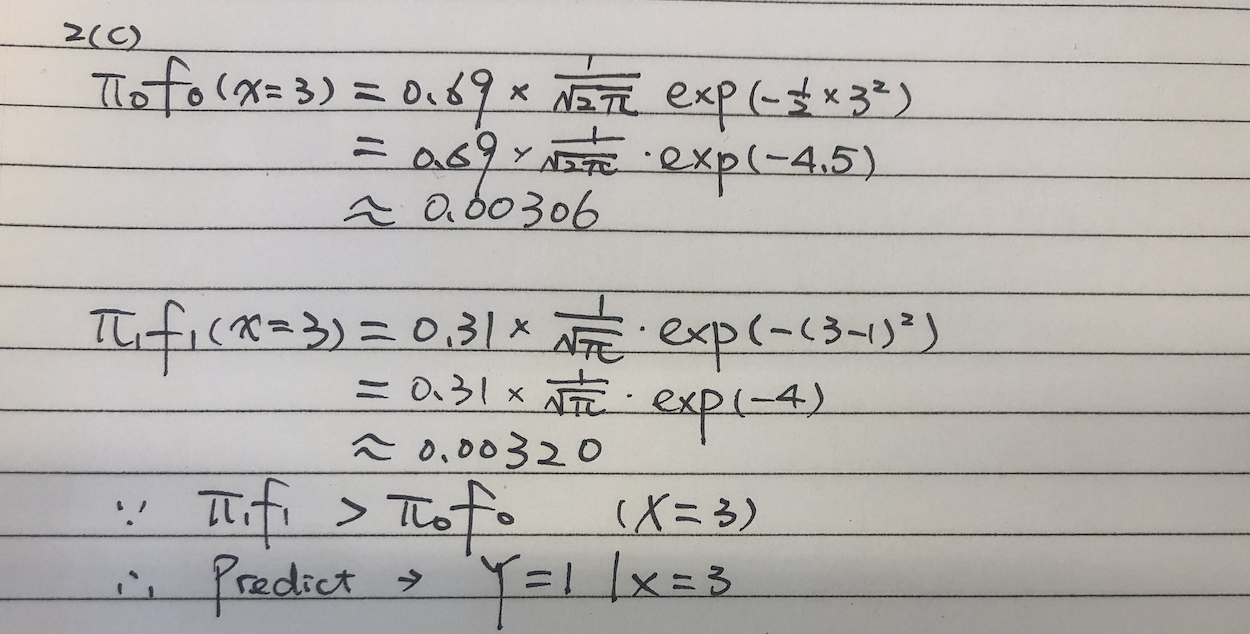
b)



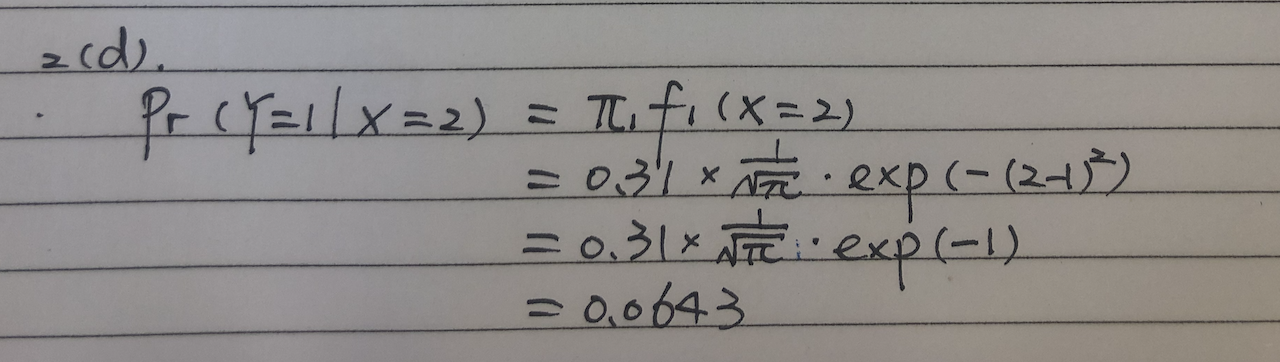
c)

According to Bayes classifier, when 0.96<X<3.04, Pr(Y=1|X) is greater. So predict that the observation X=3 will fall in class 1.

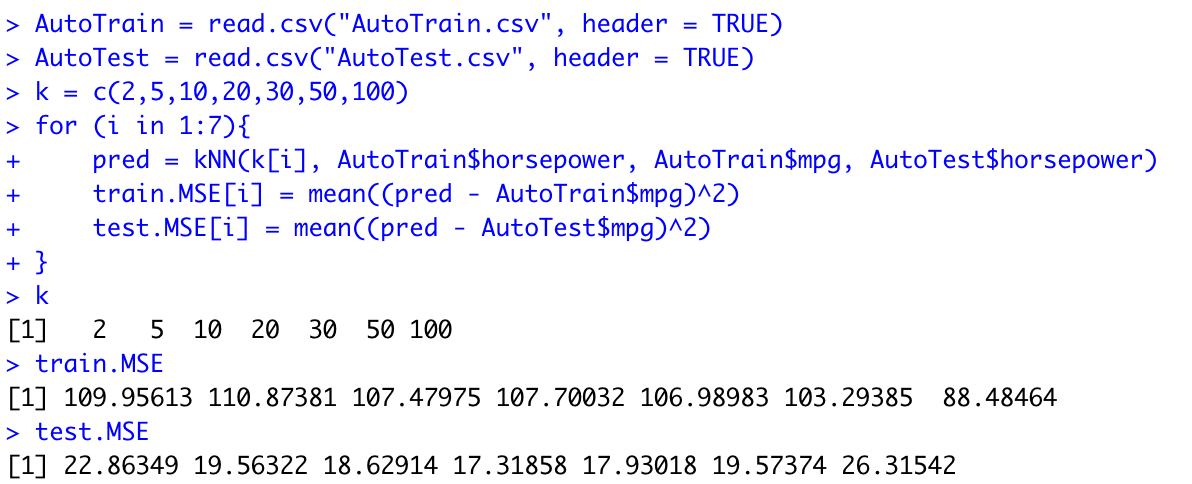
Justify the prediction:



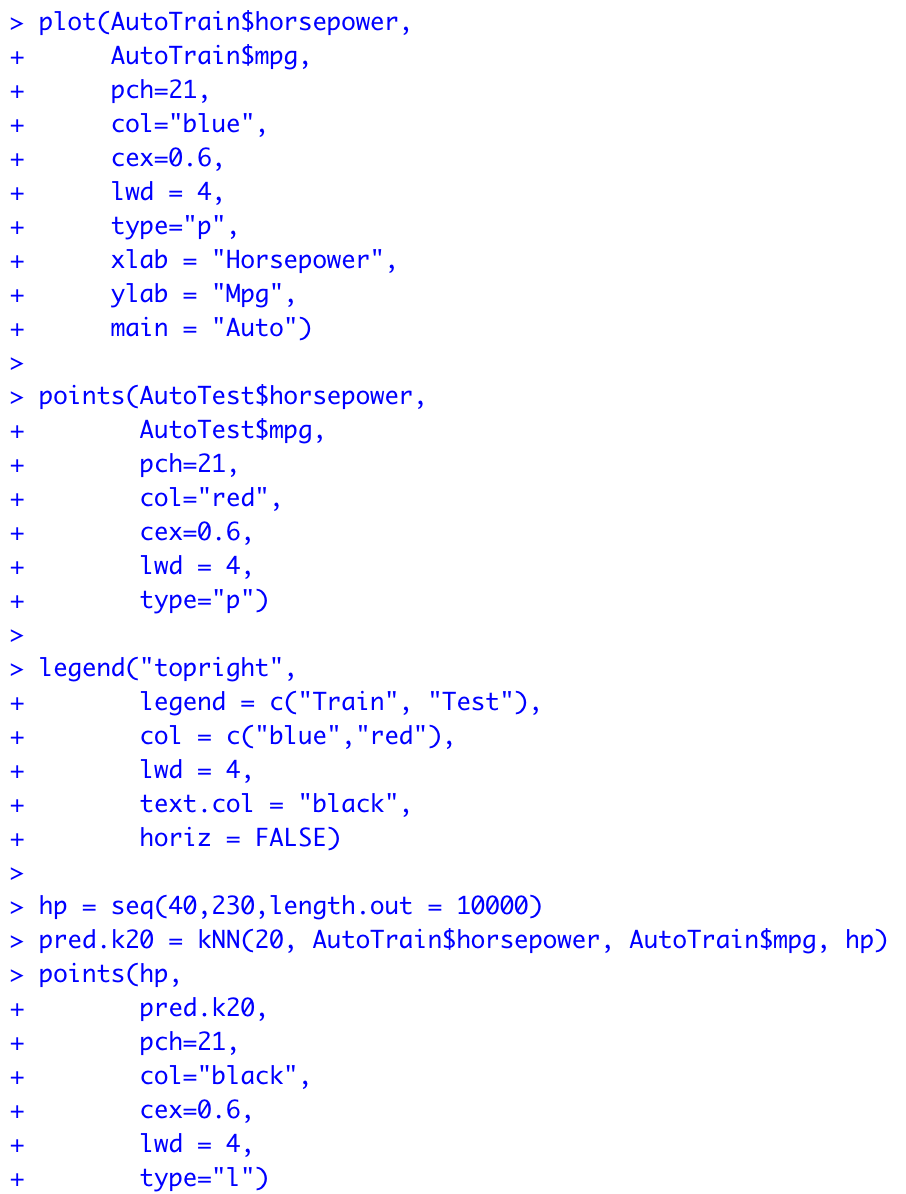
d)

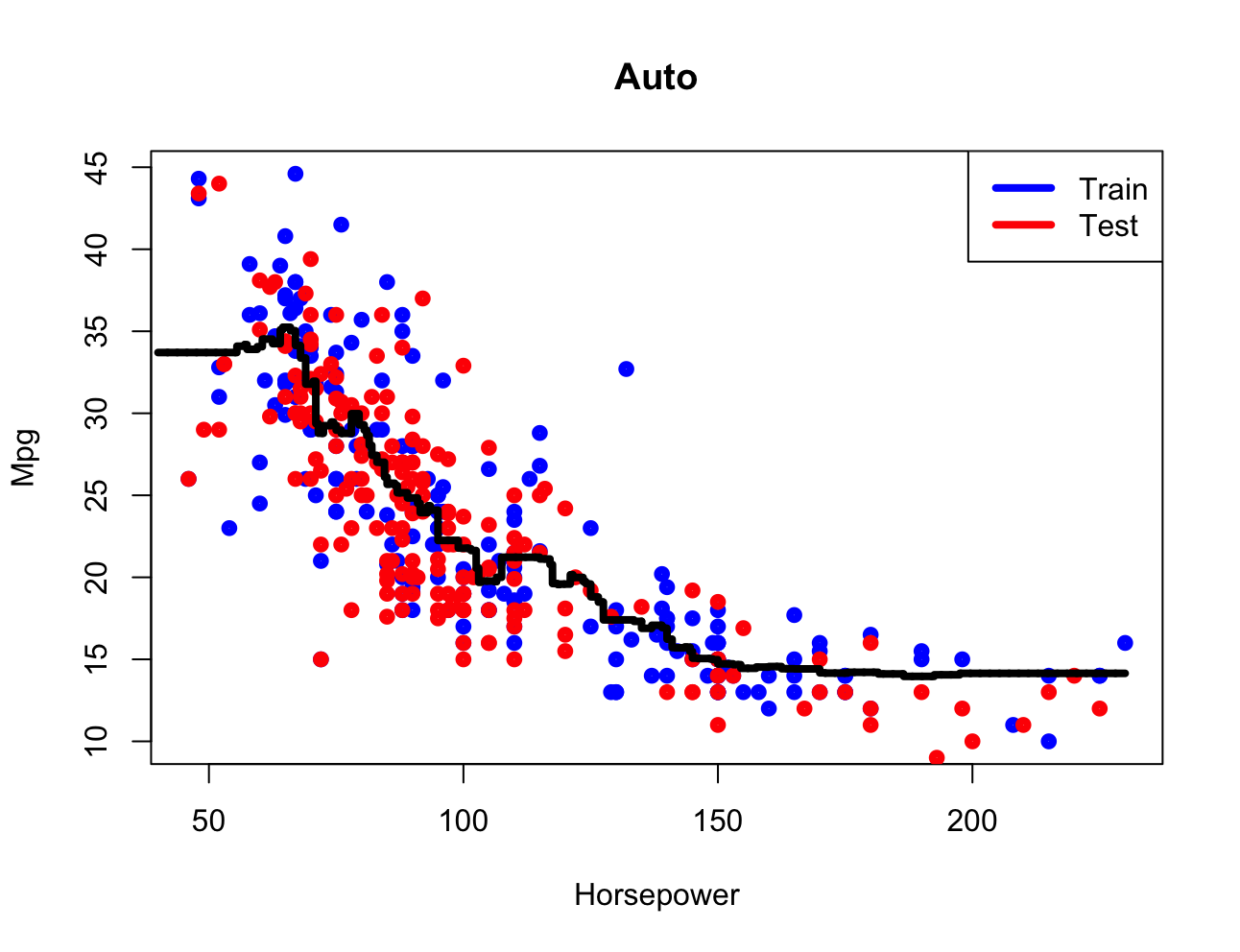






1. K= 20 performed best, because the test.MSE (17.31858) is the smallest when k=20





1. Bias-variance trade-off: when K grows, the method becomes less flexible which corresponds to a low-variance but high-bias classifier.