Introduction to Statistical Learning: Chapter 2 Excersizes

Sean Rollins

September 8, 2021

1. (a) An inflexible method would be expected to fit better because with only a few predictors, the relationship is simpler.
   (b) A flexible method would be better, because methods with low flexibility fail at approximating a multivariate relationship.
   (c) A highly flexible method would be expected to be better because methods with low flexibility fail at modeling sophisticated, complex relationships.
   (d) One cannot derive whether a flexible method or an inflexible method would perform better because $\mathbf{Var}(\varepsilon)$ is not a function of flexibility.

2. (a) Regression, inference. $n = 500, p = 3$
   (b) Classification, prediction, $n = 20, p = 13$
   (c) Regression, prediction, $n = 52, p = 3$

3. (a) dont feel like it rn
   (b)

4. (a) Predicting presence of a disease, based on variables associted with bloodwork, predicting someone's answer on a true/false question, based on demographic predictors, and predicting the species of an animal based on color, weight, etc.
   (b) Predicting the price of a security in the future based on its historical price data, seeking the relationship between standardized test scores and the household income of each participant, and predicting the amount of rainfall in the near future based on factors such air pressure, temperature, etc.

(c) Dividing customers of a certain store into different groups based on frequency of purchase and average money spent, reallocating congressional districts based on data regarding the income, location and the demographic composition of a state's constituients, categorizing regions as urban, suburban or rural based on population density.

5. Flexible approaches are better, when it is known that relationship between the response variables and the predictors is known to be complex, either having many important predictors, or a complicated function of few predictors. In general, one sacrifices interpritability of the model, and it may be prone to overfitting. Flexible approaches are better, when it is known that the relationship between the response variables and predictors is less complex, possibly due to there being a small number of predictors. In general, one may sacrifice accuracy in the form of there being and increased bias, but one gains human interpretibility, which is important if one wishes to visualize or otherwise seek relationships between indivudual predictors and responses.

6. A parametric approach reduces the problem to computing a vector of parameters. Its efficacy(and efficacy relative to non-parametric methods) hinges on how similar the form chosen for $\hat{f}(X)$ is to $f(X)$. Non-parametric functions can fit a wide variety of shapes if the form of f is not known, but can be costly and be prone to overfitting(but generally less severe than guessing f wrong with parametric approaches).

7. (a) $O_1 = 3, O_2 = 2, O_3 = \sqrt{10}, O_4 = \sqrt{5}, O_5 = \sqrt{2}, O_6 = \sqrt{3}$

   (b) Our prediction is **Green** because O_5 is the 1st closest to the origin and is associated with **Green**.

   (c) Our prediction is **Red** because O_5, O_6, and O_2 are the 3 closest to the origin and two-thirds of them have **Red** as the associated response variable.

   (d) We should expect the best value for K to be small because smaller K introduces increased flexiblity, which gives complex decision boundaries, resulting in a suitable approximation of the Bayes decision boundary.