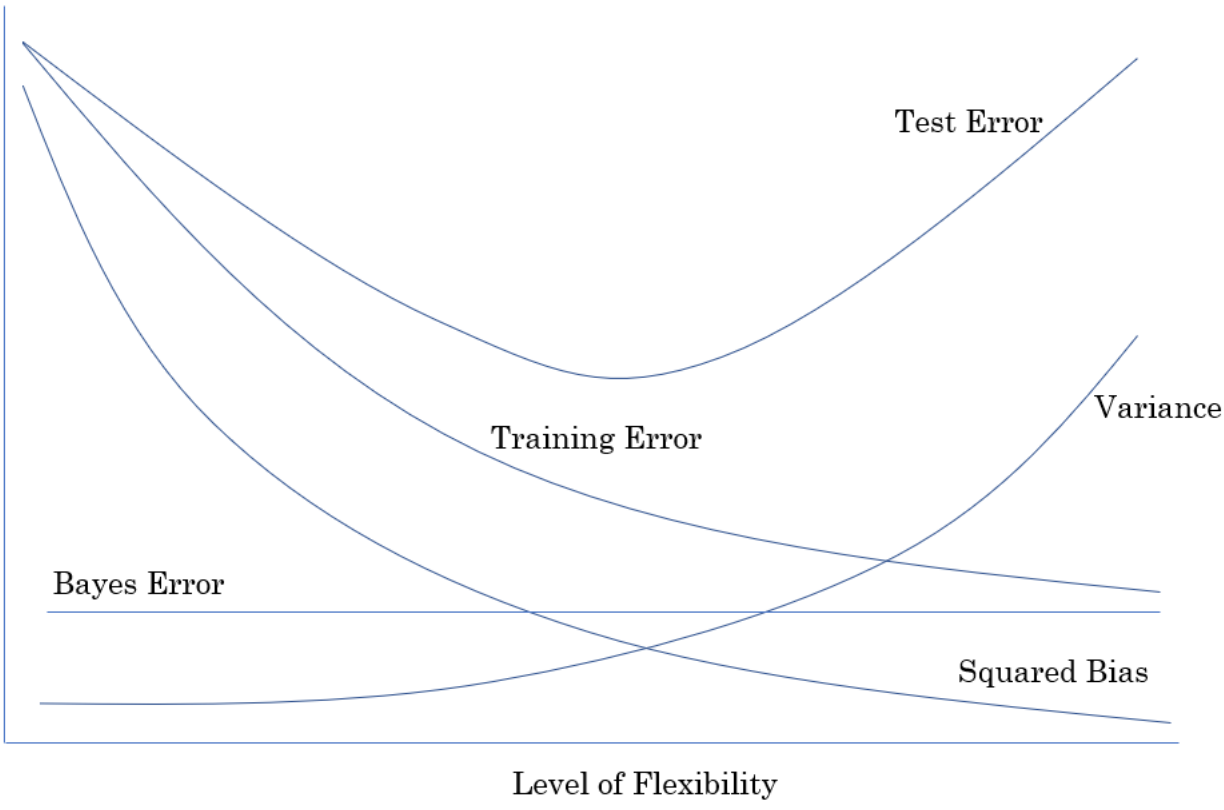
Introduction to Statistical Learning
Chapter 2 Exercise

Penghao Chen

February 14, 2019

1. (a) Flexible model would perform better.
   (b) Inflexible model would perform better.
   (c) Flexible model would perform better.
   (d) Inflexible model would perform better.
2. (a) Regression problem. Inference is our priority. n = 500 and p = 3.
   (b) Classification problem. Prediction is our priority. n = 20 and p = 13.
   (c) Regression problem. Prediction is our priority. n = 52 and p = 3.
3. (a)



(b) Omitted.

4. (a) 1. Hand-written character recognition.
   2. Email spam filtering.
   3. Cancer diagnosis.

   (b) 1. Predicting future value of the stock market.
   2. Determining which advertising channel contributes the most to sales.
   3. Predicting the probability of raining.

   (c) 1. Market segment analysis.
   2. Human genetic clustering analysis.
   3. Social network analysis.

5. Advantages of the flexible approach: 1. Low bias 2. Good for nonlinear models

   Disadvantages of the flexible approach: 1. Prone to overfitting problems. High variance. 2. Requires lots of data. Subject to the curse of dimensionality. 3. Less interpretable.

   Cases suitable for flexible approach: When you have huge amount of data. You are focusing on the accuracy of the predictions and are less concerned with the interpretability of the model.

   Cases suitable for inflexible approach: When you have limited amount of data. You are focusing on the inference problem and are concerned with the interpretability of the model.

6. Parametric approach has stronger assumptions about the underlying model. It is easier to train since the problem is reduced to estimating a few parameters.

   Non-parametric approach has no assumption about the underlying model and is harder to train. It requires a large amount of data to make the model stable.

   Advantages of parametric approach: Easier to train. Requires relatively less data. Stable outputs once well-trained. Better interpretability.

   Disadvantages of parametric approach: Higher bias if the assumption is drastically different from the true model. Or higher variance if more flexible model is used.

7. (a)
   $D(Obs_1, Obs_0) = 3$
   $D(Obs_2, Obs_0) = 2$
   $D(Obs_3, Obs_0) = \sqrt{10}$
   $D(Obs_4, Obs_0) = \sqrt{5}$
   $D(Obs_5, Obs_0) = \sqrt{2}$

$D(Obs_6, Obs_0) = \sqrt{3}$

(b) Green

(c) Red

(d) Small