ISL - Chapter 2 Exercises

Liz Muehlmann

7/26/2021

These questions (in grey) are copied from the Intro to Statistical Learning textbook. ## Chapter 2: Statistical Learning ### conceptual 1. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(a) The sample size n is extremely large, and the number of predictors p is small.
(b) The number of predictors p is extremely large, and the number of observations n is small.
(c) The relationship between the predictors and response is highly non-linear.
(d) The variance of the error terms, i.e. $\sigma^2 = Var(\epsilon)$, is extremely high.

a. **Better.** Inflexible models reduce estimating f to estimating one set of parameters. (p. 21)
b. **Worse.** Flexible models require a large number of observations to fit. Using a flexible model when the number of observations is small would lead to over-fitting (p. 23).
c. **Better.** An inflexible model would not capture the true form of f because they assume the relationship between the predictors and response is linear. (p.22)
d. **Worse.** A flexible model would likely overfit the data because it tries to reduce the noise.

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

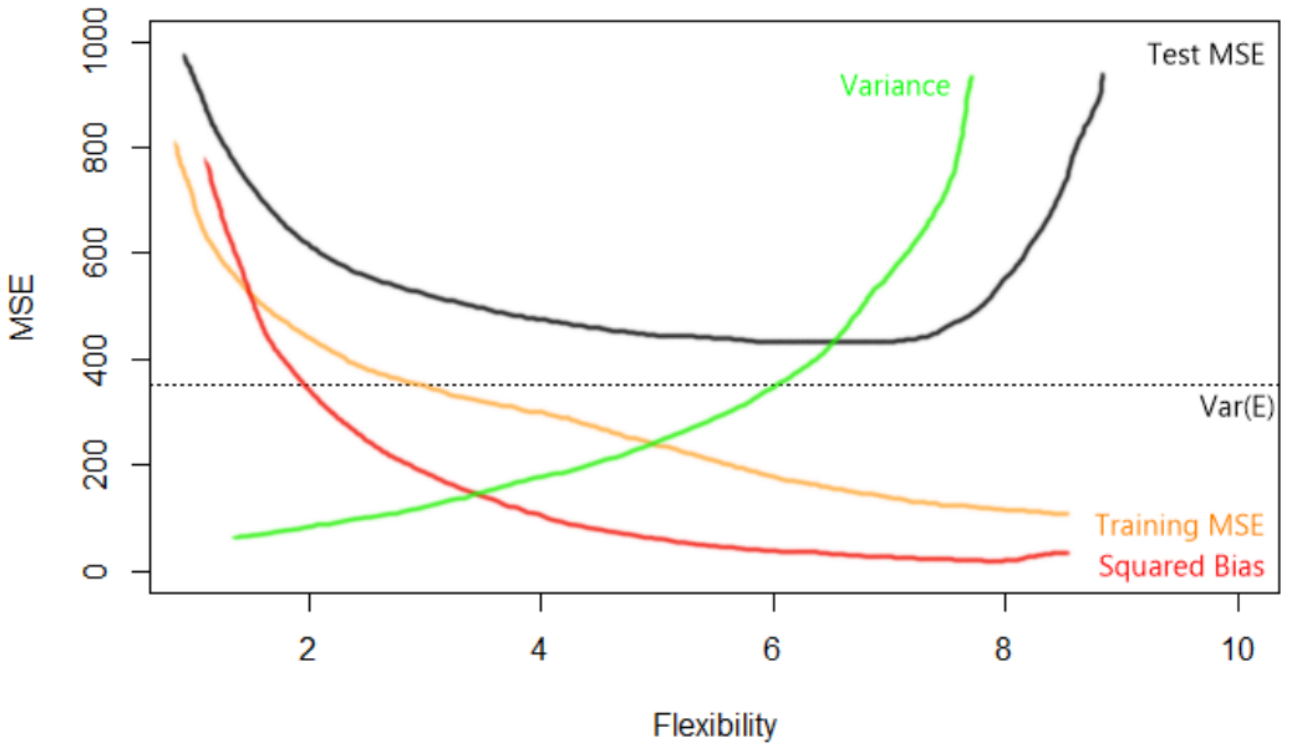
(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.
(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.
(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

a. Regression & inference.The predictors are quantitative. We want to know what affect certain factors have on the CEO's salary, not predict it.
   n = 500
   p = profit, number of employees, industry, CEO salary
b. Classification & prediction. The variable of interest (success/failure) is binary.
   n = 20
   p = success / failure, price charged for the product, marketing budget, competition price, other variables.
c. Regression & prediction. The predictors are quantitative. We want to predict percent change in USD/Euro exchange rate by using weekly changes in the work stock markets.
   n = 52
   p = % change in the USD/Euro, the % change in the US Market, the % change in the British market, and

the % change in the German market.

3. We now revisit the bias-variance decomposition.

(a) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.
(b) Explain why each of the five curves has the shape displayed in part (a).
(a)



(b) **Squared Bias:** Bias is reduced as flexibility increases because the method can better fit the data.
Variance: As flexibility increases, variance increases as the method becomes over-fit.
Training Error (Training MSE): Decreases as more flexibility means the method can closely fit the training data.
Test Error (Test MSE): Reduces to an optimum point because increased flexibility means better fit. Any further increase in optimization would lead to over-fitting the data.
Bayes (Irreducible Error): $Var(\epsilon)$ is the irreducible error. Test predictions cannot be better than this, so it is a straight line.

4. You will now think of some real-life applications for statistical learning.

(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.
(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.
(c) Describe three real-life applications in which cluster analysis might be useful.

a. Classification is useful when the outcomes need to be classified into groups (binary or multiclass)

Classification would be useful when determining the majority vote share by gender in the latest election
Response: vote share
Predictor: gender
Inference

predicting what type of legislation is most likely to be passed by a given by looking at legislation that has passed/failed in the last year
Response: success / failure of a piece of legislation
Predictor: legislation type
Prediction

using ownership data to classify the format used on local radio stations in a given market.
Response: radio station format
Predictor: ownership info and market area
Prediction

b. Regression is useful the response is quantitative.

determining the level political knowledge of people with access to news and those without.
Response: level of political knowledge
Predictor: access to news
Inference

How much a house will cost in a given market by looking at location, features, school ranking.
Response: cost of the house
Predictor: location, features, school ranking Prediction

Whether appeals to certain issues affects voter turnout.
Response: Turnout
Predictor: appeals to certain issues.
Inference

c. Cluster analysis is good when we do not have a target response.

determining voting preferences of individuals using their intersectional identities.
based on certain characteristics like income, age, party ID classify people into groups of voters
based on characteristics of the area classify radio stations into different format types.

5. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

Flexible models may be overfit and lack easy interpretation. However, if the true form of f is non-linear, flexible models will come closer to capturing its true form. A more flexible model would be preferred if the true form of f is non-linear, if there is a large number of observations and a small number of predictors, when we do not make assumptions about the estimated function, and when variance is normal. These models typically have low bias, but higher variance because the data is overfit.

6. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

Non-parametric statistical learning requires no assumption of the functional form of f and are better suited for real-life problems because these methods do not assume a linear relationship between the predictors and the response. However, they require a large number of observations and small number of predictors. If the functional form of f is linear, then these methods will be accurate, but lack clear interpretability.

Parametric statistical learning assumes that the predictors are linearly related to the response. Thus, it makes strong assumptions about the true from of f. Parametric approaches are usually more easily interpreted and usually better when we want to make inferences. Furthermore, since these methods assume a linear relationship between the predictors and responses, a smaller number of parameters is needed.

7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X1	X2	X3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	−1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.
(b) What is our prediction with K = 1? Why?
(c) What is our prediction with K = 3? Why?
(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

// # 7 answer taken from https://onmee.github.io/assets/docs/ISLR/Statistical-Learning.pdf (https://onmee.github.io/assets/docs/ISLR/Statistical-Learning.pdf) (a) The Euclidean distance is the straight line distance between two points. This can be calculated using the Pythagorean theorem.

For 3D space we have:

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}$$

Using the above formula, we get the following distances:
d(1, test) = 3
d(2, test) = 2

$d(3, test) = 3.16$
$d(4, test) = 2.24$
$d(5, test) = 1.41$
$d(6, test) = 1.73$

b. Green: as nearest single observation is green.
c. Red: as nearest three observations are green, red and red.The probability of the test point belonging to red is 2/3 and green is 1/3. Therefore, the prediction is red.
d. For highly non-linear boundaries, we would expect the best value of K to be small. Smaller values of K result in a more flexible KNN model, and this will produce a decision boundary that is non-linear. A larger K would mean more data points are considered by the KNN model and this means its decision boundary is closer to a linear shape.