An Introduction to Statistical Learning: Chapter 2

Due on December 29,2015 at 3:10pm

G. James et.al.

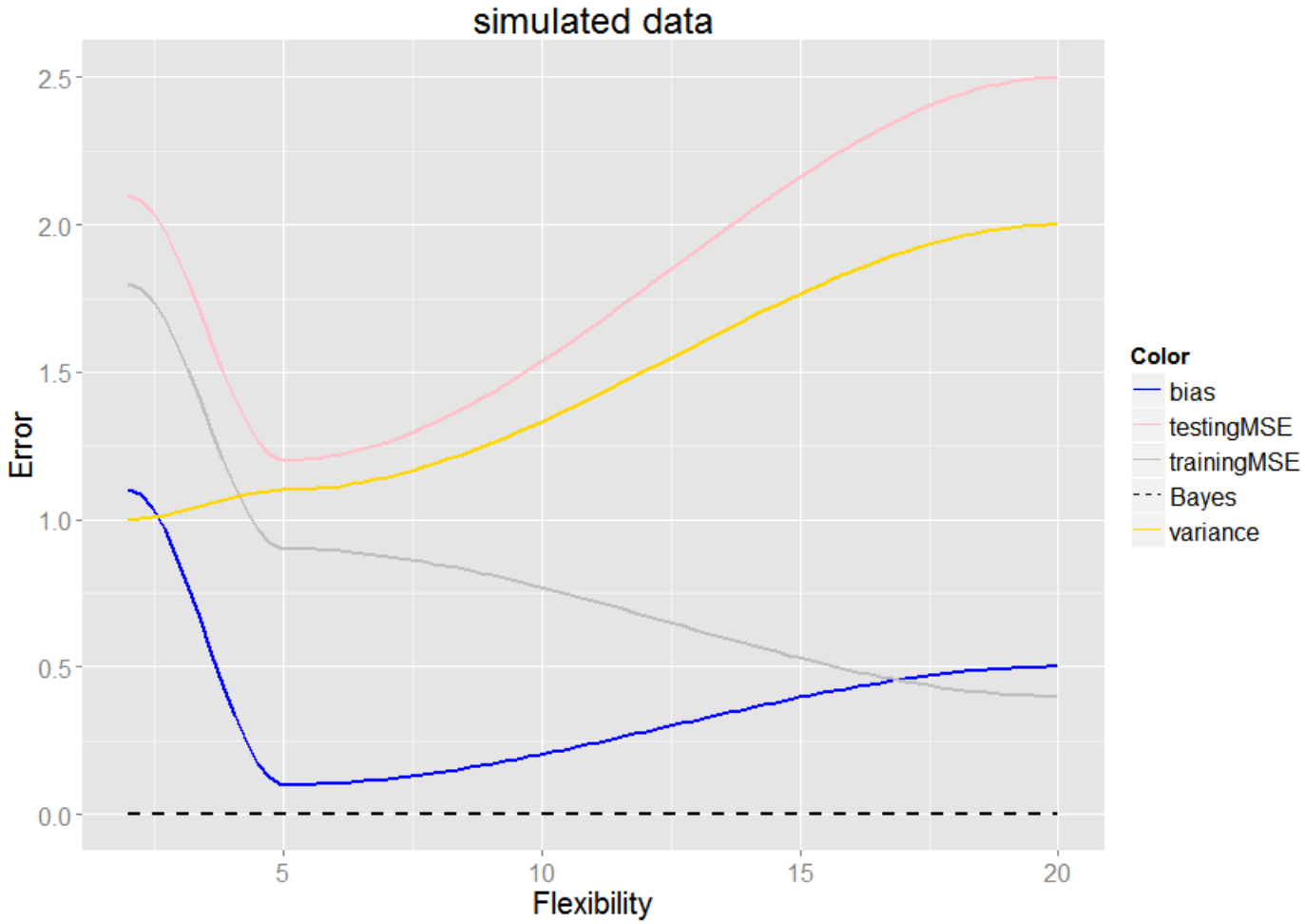
Tony Jiang

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. $\sigma^2 = \text{Var}(\epsilon)$, is extremely high.
Answer:
(a) better. when the sample size is large and the number of predictor is small, a more flexible provides a better fit
(b) worse. flexible method would suffer from over-fitting.
(c) better. more flexible method can better capture the non-linearity relationship.
(d) worse. when the data is noisy, flexible method tends to over fit the data which are mostly noise.

2. 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 US market, the % change in the British market, and the % change in the German market.
Answer:
(a) regression(inference). n=500,p=3 (number of employees, industry and the CEO salary)
(b) classification(prediction). n=20, p=13 (price charged for the product, marketing budget, competition price, and ten other variables)
(c) regression(prediction).n=52, p=3(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 flexible 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)
Answer:

(a)



(b)

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 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.
Answer:
(a)

- stock market price direction, prediction, response: up, down, input: yesterday's price movement change, etc.
- illness classification, inference, response: ill, healthy, input: resting heart rate, resting breath rate, mile run time
- car part replacement, prediction, response: needs to be replace, good, input: age of part, mileage used for, current amperage

(b)

- CEO salary. inference. predictors: age, industry experience, industry, years of education. response: salary.
- car part replacement. inference. response: life of car part. predictors: age of part, mileage used for, current amperage.
- illness classification, prediction, response: age of death, input: current age, gender, resting heart rate, resting breath rate, mile run time.

(c)

- cancer type clustering. diagnose cancer types more accurately.
- Netflix movie recommendations. recommend movies based on users who have watched and rated similar movies.
- marketing survey. clustering of demographics for a product(s) to see which clusters of consumers buy which products.

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?

Answer:

The advantages for a very flexible approach for regression or classification are obtaining a better fit for non-linear models, decreasing bias.

The disadvantages for a very flexible approach for regression or classification are requires estimating a greater number of parameters, follow the noise too closely (overfit), increasing variance. 6. Describe the differences between a parametric and a non-parametric statistical approach. What are the advantages of a parametric approach to regression or classification(as opposed to a non-parametric approach)? What are its disadvantages?

Answer:

A parametric approach reduces the problem of estimating f down to one of estimating a set of parameters because it assumes a form for f.

A non-parametric approach does not assume a functional form for f and so requires a very large number of observations to accurately estimate f.
The advantages of a parametric approach to regression or classification are the simplifying of modeling f to a few parameters and not as many observations are required compared to a non-parametric approach.
The disadvantages of a parametric approach to regression or classification are a potential to inaccurately estimate f if the form of f assumed is wrong or to overfit the observations if more flexible models are used.
7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X_1	X_2	X_3	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	0	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.
(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$
(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 non-linear, then would we expect the *best* value for K to be large or small? Why?
Answer:
(a)

Obs.	X_1	X_2	X_3	Y	Distance to (0,0,0)
1	0	3	0	Red	3
2	2	0	0	Red	2
3	0	1	3	Red	$\sqrt{10}$
4	0	1	2	Green	$\sqrt{5}$
5	-1	0	1	Green	$\sqrt{2}$
6	1	1	1	Red	$\sqrt{3}$

(b) Green. The nearest observation is 5.
(c) Red. The 3 nearest neighbors are 2,5 and 6 in which Red dominates over Green (2:1)
(d) small. needs less neighbors to make the boundary more wiggly.