ACTL3142

Week 1: Introduction to Statistical Learning

Tetian Madfouni

t.madfouni@unsw.edu.au

Icebreaker

Let’s go around the room, each person tell me

- Your name
- Degree and year
- How much experience do you have in programming?
- What’s one thing you want to get out of this course?

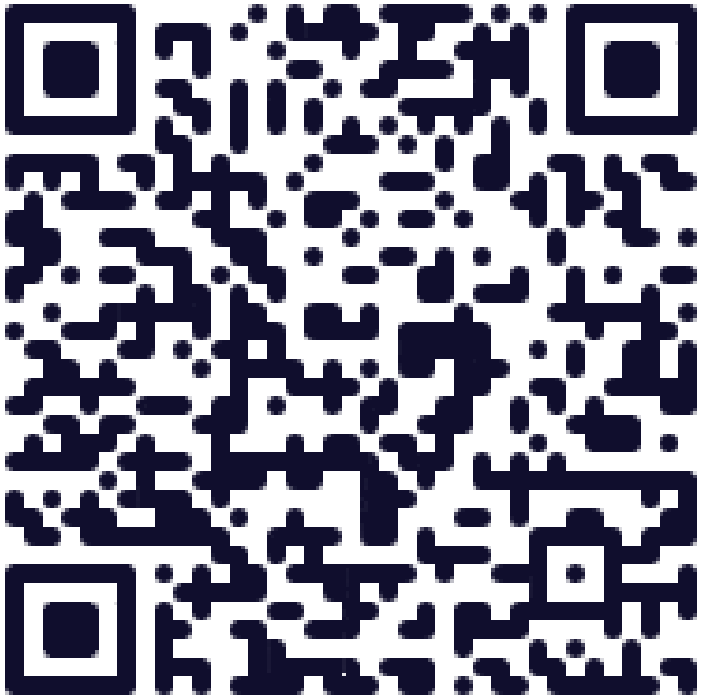
My Tips

Knowing how to code and being able to learn new concepts related to programming are slowly becoming more and more useful skills in an increasing number of jobs. This course gives you another opportunity to learn how to code in R in more depth, take advantage of it.

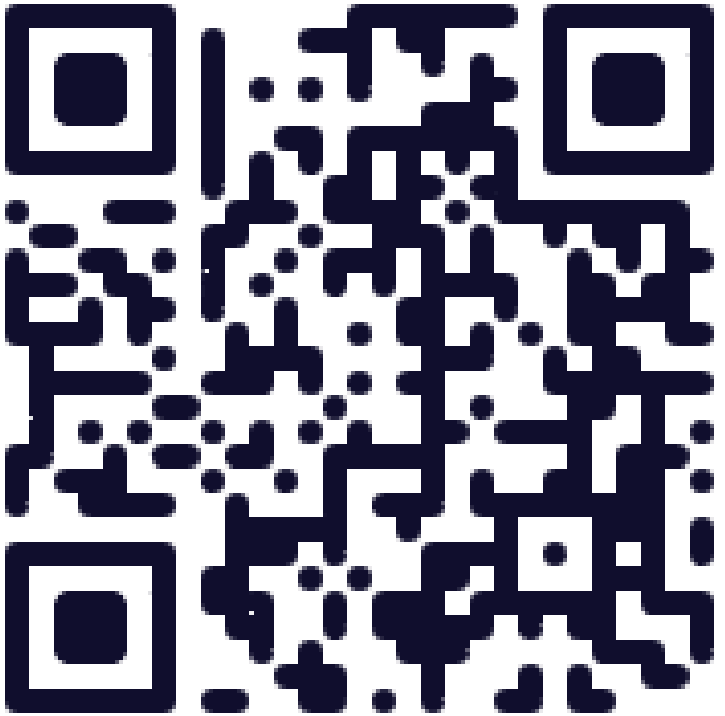
Read the ISLR textbook. Normally recommended textbooks for your courses can be convoluted or don’t add much to the course. In my opinion, this one will help you a lot if you actually do the readings.

ChatGPT…

Useful Links



Github Repo with Lab Code



Actuarial Students Discord Server

Conceptual Question 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. $\sigma^2 = \mathbb{V}(\epsilon)$, is extremely high.

Things to consider before answering the question

1. What is an inflexible method vs. flexible?
2. Can you give an example of each?

Related Concepts

1. What is overfitting vs. underfitting and how does this relate to flexibility?

Conceptual Question 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).

Things to consider before answering the question

1. What exactly are each of the things we're asked to plot?
2. What happens to each of them as we increase/decrease the flexibility of a model?

Conceptual Question 4

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?

Things to consider before answering the question

- What is the difference between modelling for inference vs. prediction? Would you prefer more accuracy or interpretability for each of those use cases?

Conceptual Question 5

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	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when $X1 = X2 = X3 = 0$ using K-nearest neighbours.

a. Compute the Euclidean distance between each observation and the test point, $X1 = X2 = X3 = 0$.

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?