

## DATS 7860 Statistical and Machine Learning for Big Data

### Homework 3

**Textbook (required):** *An Introduction to Statistical Learning with Applications in R*, 2nd Edition (ISBN: 978-1071614174), by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2021), Springer (**Referred as ISL in this course**).

The PDF version of the book, R codes for labs and datasets are publicly available from the book's website: <https://www.statlearning.com/>

A PDF version of ISL was also uploaded to D2L "textbook" folder.

1. Please study and run the two examples (1). "Cross-Validation – The Right and Wrong Way", and (2). "Monte Carlo simulation and the bootstrap" (uploaded in the folder "Examples" in Week 5 on D2L). Write a few sentences summarizing the results and what you have learned from these examples.

**Problems below are all from Exercises 5.4 of Chapter 5 of ISL:**

#### 2. Problem 2 (on Page 219).

Some specific notes for this problem:

- For part (a), (b) and (c), just briefly explaining your answer (say, in one sentence) is sufficient.
- On a separate note: we will use the results from this problem in Chapter 8 - tree-based methods.

#### 3. Problem 5 (on Page 220).

Some specific notes for this problem:

- Please submit your complete R codes for this problem.
- The Default dataset is available from the R package ISLR2. You can use the following R codes to load the dataset:

```
require(ISLR2)
```

```
data(Default)
```

- For part (d), splitting the dataset one time to estimate the test error is sufficient, although you can try to repeat the process multiple times like in part (c).

#### 4. Problem 7 (on Page 222).

Some specific notes for this problem:

- Please submit your complete R codes for this problem.
- The Weekly dataset is available from the R package ISLR2. You can use the following R codes to load the dataset:

```
require(ISLR2)
```

```
data(Weekly)
```

- You can answer part (d) and (e) together, but please do remember to include your R codes for part (d).

### **5. Problem 9 (on Page 223)**

Some specific notes for this problem:

- Please submit your complete R codes for this problem.
- The Boston dataset is available from the R package MASS. You can use the following R codes to load the dataset:

```
require(MASS)
```

```
data(Boston)
```

- You may want to use the R package boot to perform the bootstrap, or you can also write your own functions.
- The number of bootstrap samples used should not be too small. The suggested number should be at least 100.