Cross Validation

Microsoft

Intuition Cross Validation

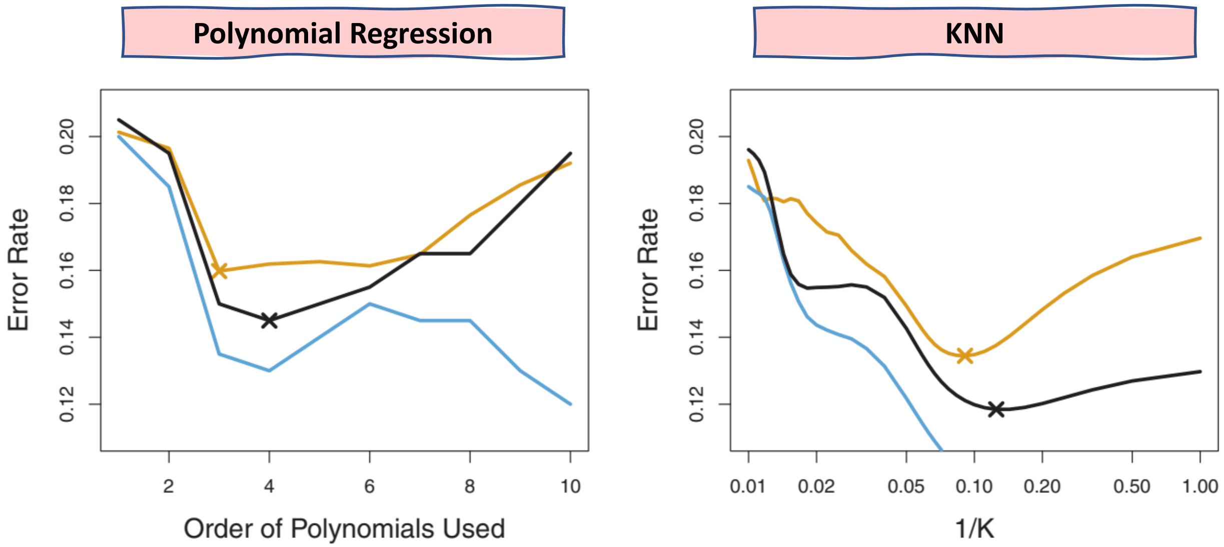


FIGURE 5.8. *Test error (brown), training error (blue), and 10-fold CV error (black) on the two-dimensional classification data displayed in Figure 5.7.* Left: *Logistic regression using polynomial functions of the predictors. The order of the polynomials used is displayed on the x-axis.* Right: *The KNN classifier with different values of K, the number of neighbors used in the KNN classifier.*

http://faculty.marshall.usc.edu/gareth-james/ISL/ | Chapter 5

Cross Validation techniques

- **Definition:**
  - ***Cross Validation*** can be understood as the ***process/es*** of testing your model against data that is ***not on the training set***.

- **Cross Validation Techniques.** The three main techniques are:
  - Validation/holdout Set.
  - LOOCV (Leave-One-Out Cross-Validation).
  - K-Fold Cross-Validation.

In practice we usually use a combination of validation set with either LOOCV or K-Fold.

http://faculty.marshall.usc.edu/gareth-james/ISL/ | Chapter 5

Validation (Holdout) set

Validation Set

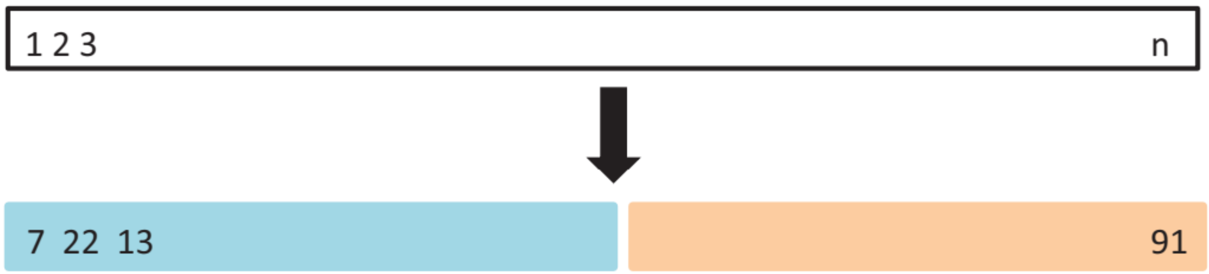


FIGURE 5.1. *A schematic display of the validation set approach. A set of n observations are randomly split into a training set (shown in blue, containing observations 7, 22, and 13, among others) and a validation set (shown in beige, and containing observation 91, among others). The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.*

The percentage of the training set can vary. However, in practice we typically find it to be from 70-80% of the original dataset. On the other hand, we often see the validation being around 20-30%.

http://faculty.marshall.usc.edu/gareth-james/ISL/ | Chapter 5

LOOCV (Leave-One-Out Cross-Validation)

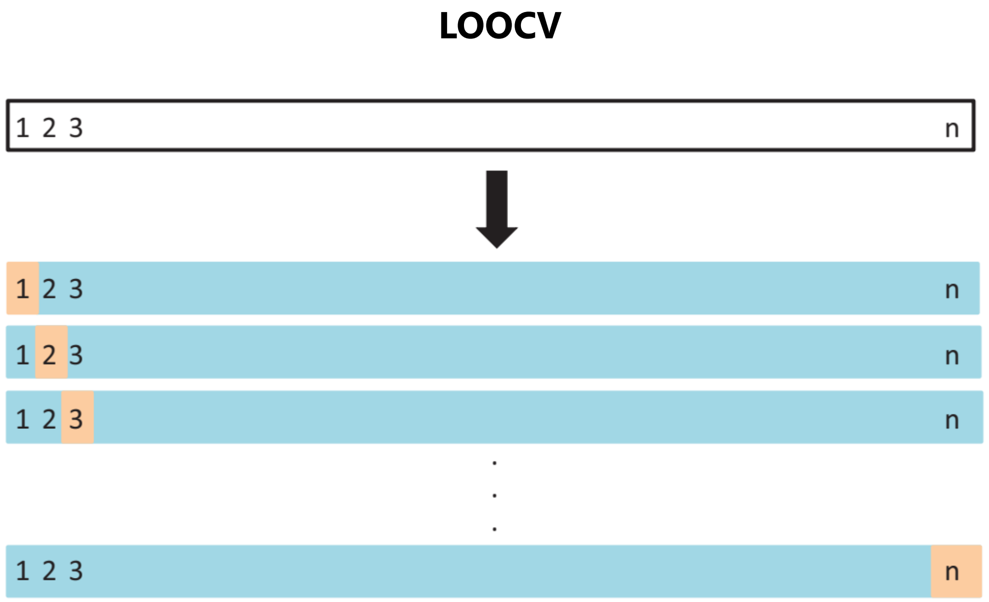


FIGURE 5.3. *A schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The test error is then estimated by averaging the n resulting MSE's. The first training set contains all but observation 1, the second training set contains all but observation 2, and so forth.*

One of the risks of the LOOCV is that we create so many models of the same data that we end up overfitting the model.

http://faculty.marshall.usc.edu/gareth-james/ISL/ | Chapter 5

K-Fold Validation

K-Fold

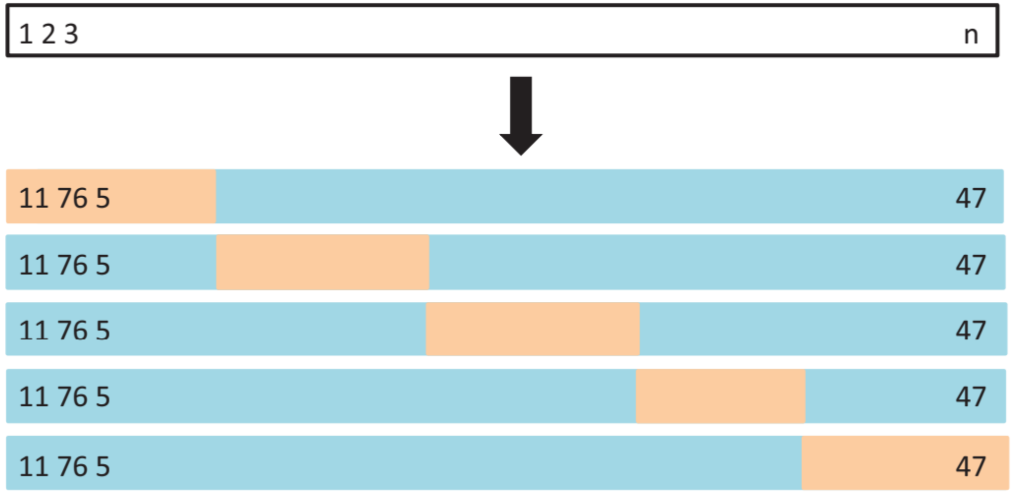


FIGURE 5.5. *A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.*

One of the advantages of the k-fold cross validation is that it measures how well the model is going to perform with new data without overfitting to the training data.

http://faculty.marshall.usc.edu/gareth-james/ISL/ | Chapter 5

Difference between LOOCV and K-Fold

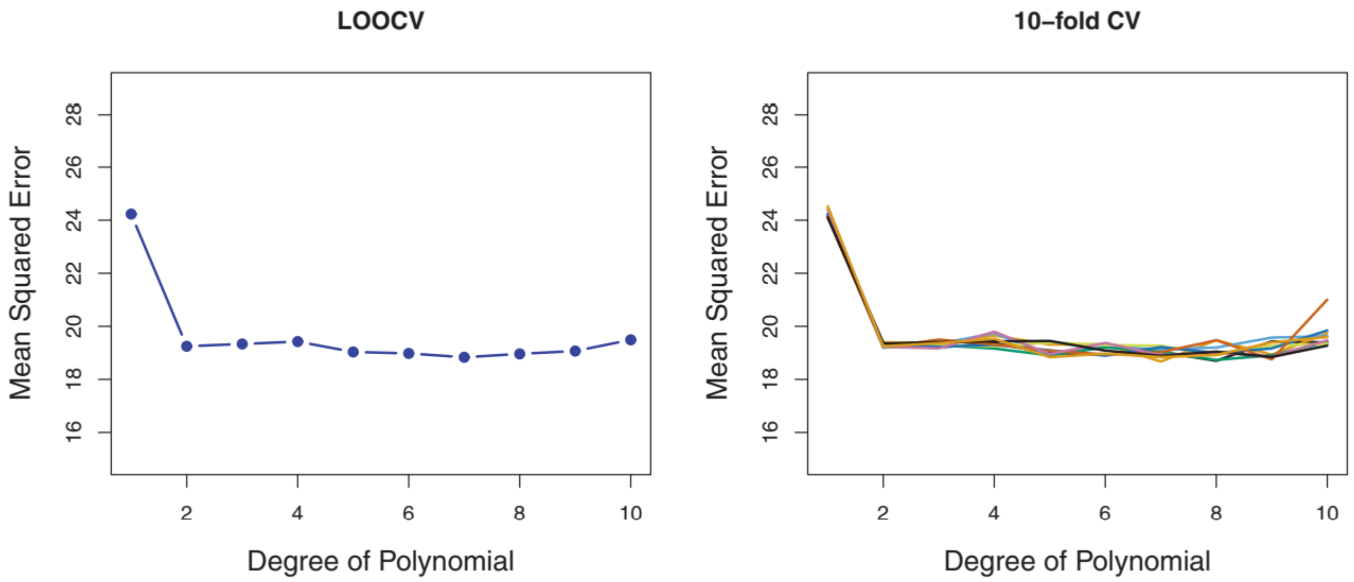
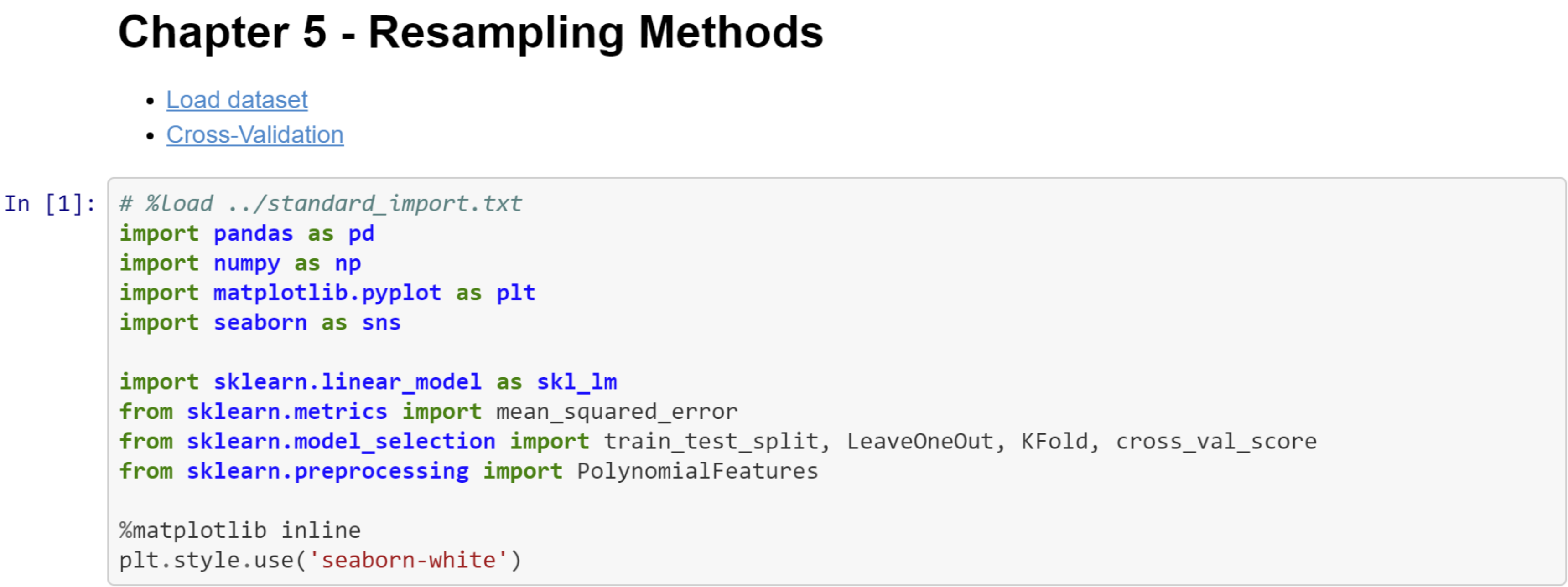


FIGURE 5.4. *Cross-validation was used on the* `Auto` *data set in order to estimate the test error that results from predicting* `mpg` *using polynomial functions of* `horsepower`. Left: *The LOOCV error curve.* Right: 10-*fold CV was run nine separate times, each with a different random split of the data into ten parts. The figure shows the nine slightly different CV error curves.*

LOOCV and 10-fold often have similar results in accuracy.

http://faculty.marshall.usc.edu/gareth-james/ISL/ | Chapter 5

Walk-through of concepts



Chapter 5 - Resampling Methods

- Load dataset
- Cross-Validation

In [1]:

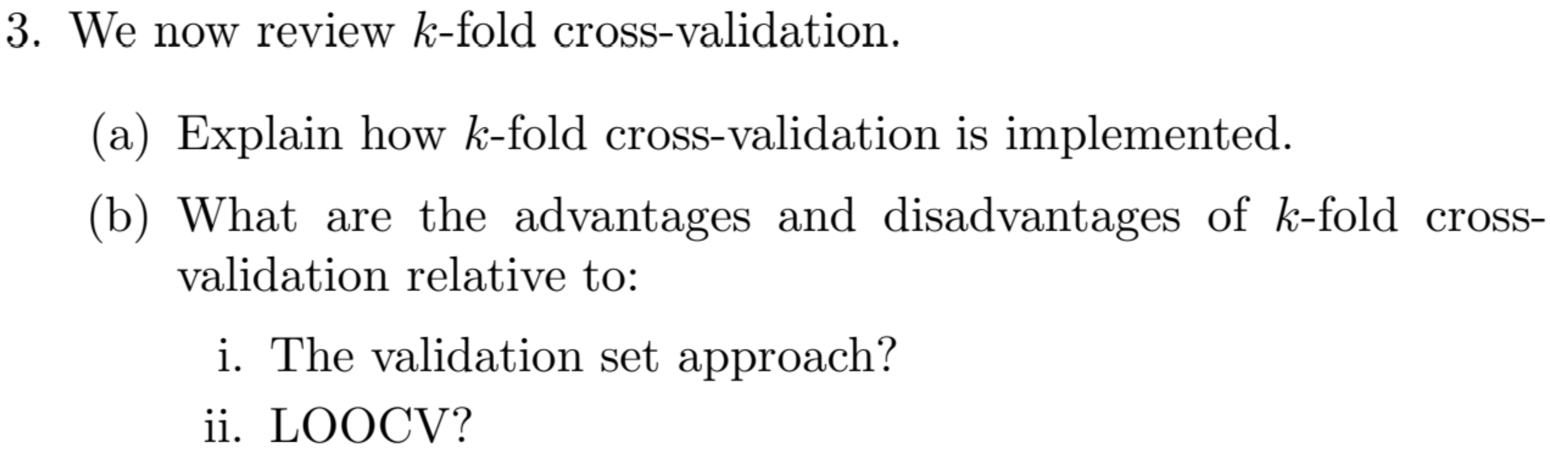
```
# %load ../standard_import.txt
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

import sklearn.linear_model as skl_lm
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split, LeaveOneOut, KFold, cross_val_score
from sklearn.preprocessing import PolynomialFeatures

%matplotlib inline
plt.style.use('seaborn-white')
```

https://github.com/JWarmenhoven/ISLR-python

https://github.com/JWarmenhoven/ISLR-python/blob/master/Notebooks/Chapter%205.ipynb

http://faculty.marshall.usc.edu/gareth-james/ISL/

Conceptual Question



3. We now review k-fold cross-validation.

   (a) Explain how k-fold cross-validation is implemented.

   (b) What are the advantages and disadvantages of k-fold cross-validation relative to:

      i. The validation set approach?

      ii. LOOCV?

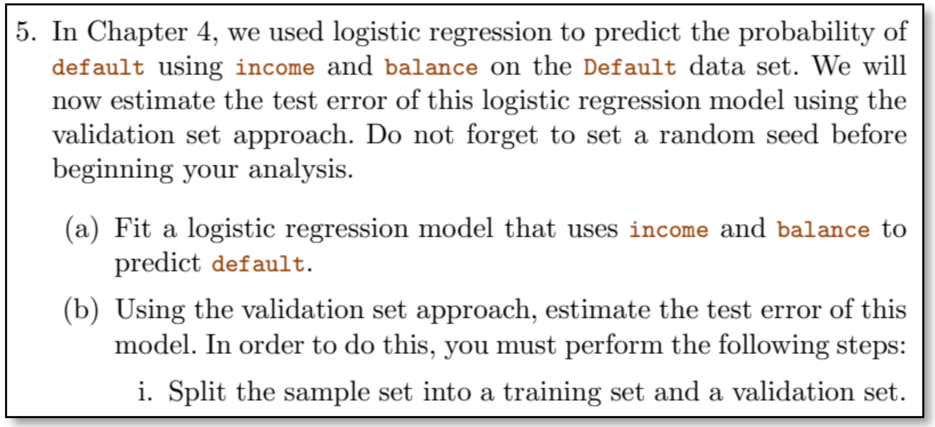
https://botlnec.github.io/islp/sols/chapter5/exercise3/

http://faculty.marshall.usc.edu/gareth-james/ISL/

Cross Validation

Microsoft

Exercise 5.5

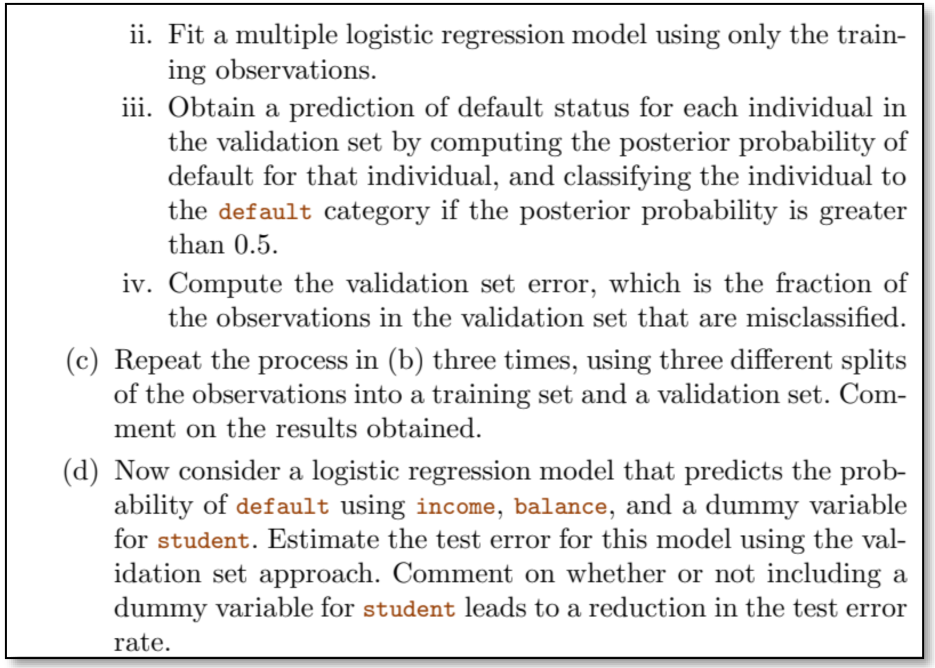


5. In Chapter 4, we used logistic regression to predict the probability of `default` using `income` and `balance` on the `Default` data set. We will now estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis.

   (a) Fit a logistic regression model that uses `income` and `balance` to predict `default`.

   (b) Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:

      i. Split the sample set into a training set and a validation set.



      ii. Fit a multiple logistic regression model using only the training observations.

      iii. Obtain a prediction of default status for each individual in the validation set by computing the posterior probability of default for that individual, and classifying the individual to the `default` category if the posterior probability is greater than 0.5.

      iv. Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.

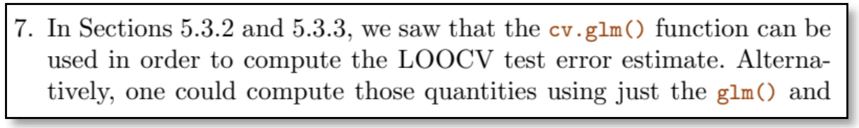
   (c) Repeat the process in (b) three times, using three different splits of the observations into a training set and a validation set. Comment on the results obtained.

   (d) Now consider a logistic regression model that predicts the probability of `default` using `income`, `balance`, and a dummy variable for `student`. Estimate the test error for this model using the validation set approach. Comment on whether or not including a dummy variable for `student` leads to a reduction in the test error rate.

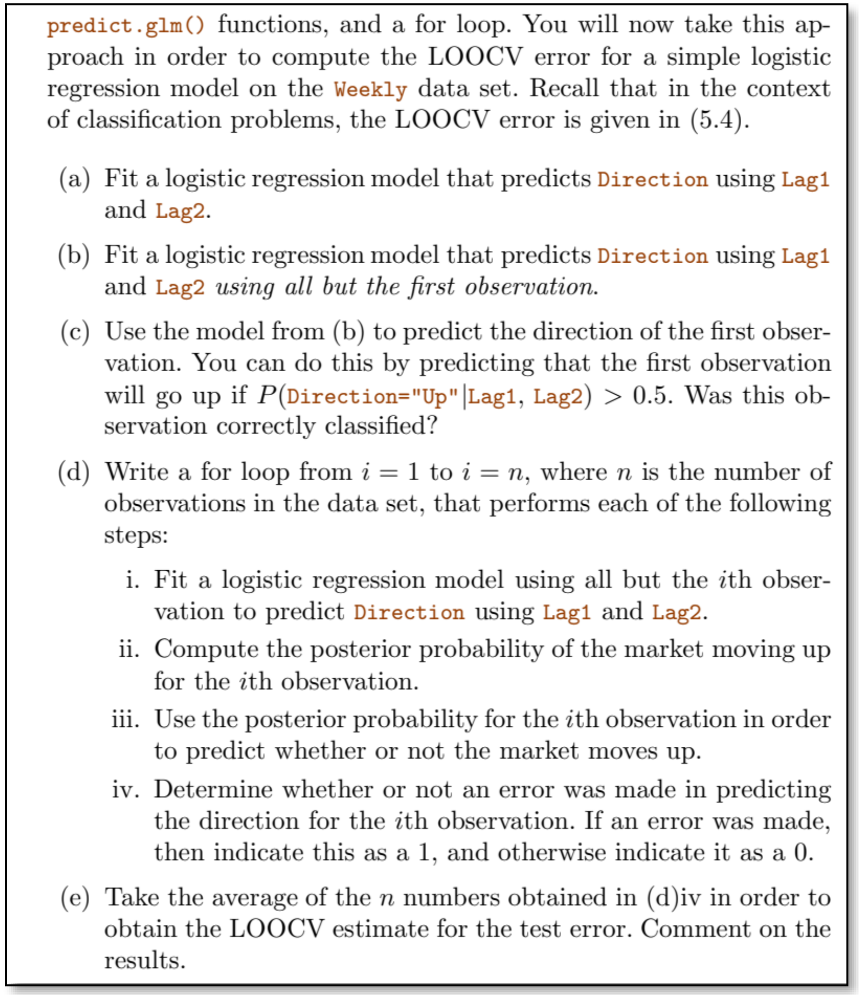
https://botlnec.github.io/islp/sols/chapter5/exercise5/

http://faculty.marshall.usc.edu/gareth-james/ISL/ | Chapter 5

Exercise 5.7



7. In Sections 5.3.2 and 5.3.3, we saw that the `cv.glm()` function can be used in order to compute the LOOCV test error estimate. Alternatively, one could compute those quantities using just the `glm()` and



`predict.glm()` functions, and a for loop. You will now take this approach in order to compute the LOOCV error for a simple logistic regression model on the `Weekly` data set. Recall that in the context of classification problems, the LOOCV error is given in (5.4).

(a) Fit a logistic regression model that predicts `Direction` using `Lag1` and `Lag2`.

(b) Fit a logistic regression model that predicts `Direction` using `Lag1` and `Lag2` *using all but the first observation.*

(c) Use the model from (b) to predict the direction of the first observation. You can do this by predicting that the first observation will go up if $P($`Direction="Up"`$|$`Lag1`, `Lag2`$) > 0.5$. Was this observation correctly classified?

(d) Write a for loop from $i = 1$ to $i = n$, where n is the number of observations in the data set, that performs each of the following steps:

   i. Fit a logistic regression model using all but the ith observation to predict `Direction` using `Lag1` and `Lag2`.

   ii. Compute the posterior probability of the market moving up for the ith observation.

   iii. Use the posterior probability for the ith observation in order to predict whether or not the market moves up.

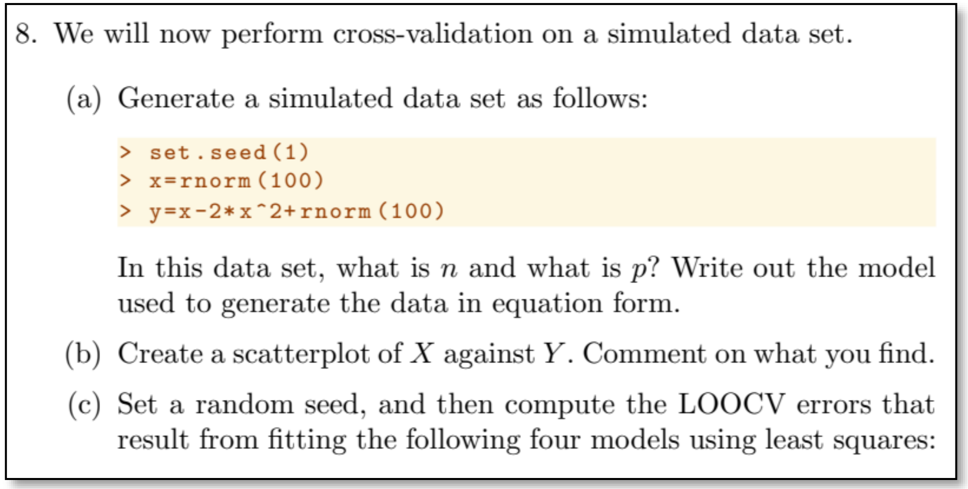
   iv. Determine whether or not an error was made in predicting the direction for the ith observation. If an error was made, then indicate this as a 1, and otherwise indicate it as a 0.

(e) Take the average of the n numbers obtained in (d)iv in order to obtain the LOOCV estimate for the test error. Comment on the results.

https://botlnec.github.io/islp/sols/chapter5/exercise7/

http://faculty.marshall.usc.edu/gareth-james/ISL/ | Chapter 5

Exercise 5.8



8. We will now perform cross-validation on a simulated data set.

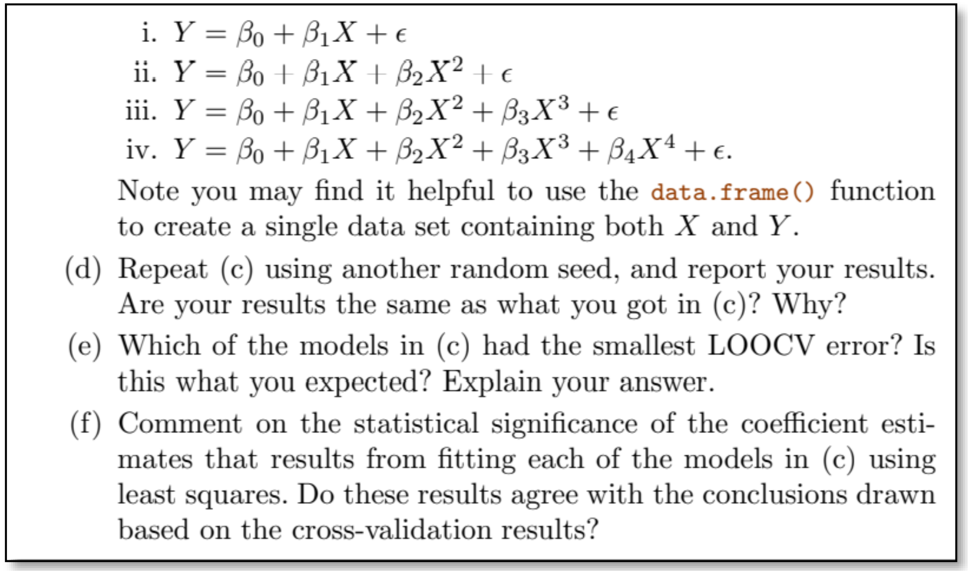
   (a) Generate a simulated data set as follows:

   ```
   > set.seed(1)
   > x=rnorm(100)
   > y=x-2*x^2+rnorm(100)
   ```

   In this data set, what is n and what is p? Write out the model used to generate the data in equation form.

   (b) Create a scatterplot of X against Y. Comment on what you find.

   (c) Set a random seed, and then compute the LOOCV errors that result from fitting the following four models using least squares:



   i. $Y = \beta_0 + \beta_1 X + \epsilon$
   ii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$
   iii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$
   iv. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$.

   Note you may find it helpful to use the `data.frame()` function to create a single data set containing both X and Y.

   (d) Repeat (c) using another random seed, and report your results. Are your results the same as what you got in (c)? Why?

   (e) Which of the models in (c) had the smallest LOOCV error? Is this what you expected? Explain your answer.

   (f) Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in (c) using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?

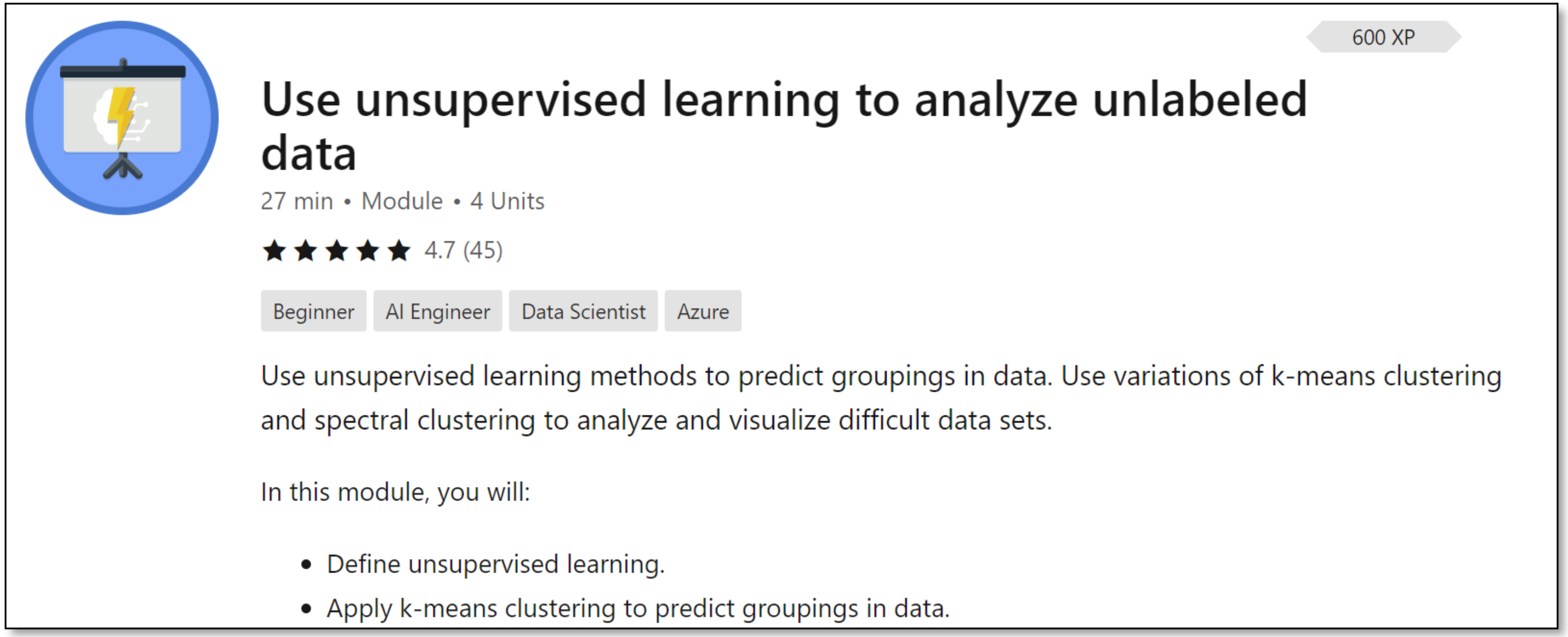
https://botlnec.github.io/islp/sols/chapter5/exercise8/

http://faculty.marshall.usc.edu/gareth-james/ISL/ | Chapter 5

Clustering

Microsoft

Clustering



600 XP

Use unsupervised learning to analyze unlabeled data

27 min • Module • 4 Units

★★★★★ 4.7 (45)

Beginner | AI Engineer | Data Scientist | Azure

Use unsupervised learning methods to predict groupings in data. Use variations of k-means clustering and spectral clustering to analyze and visualize difficult data sets.

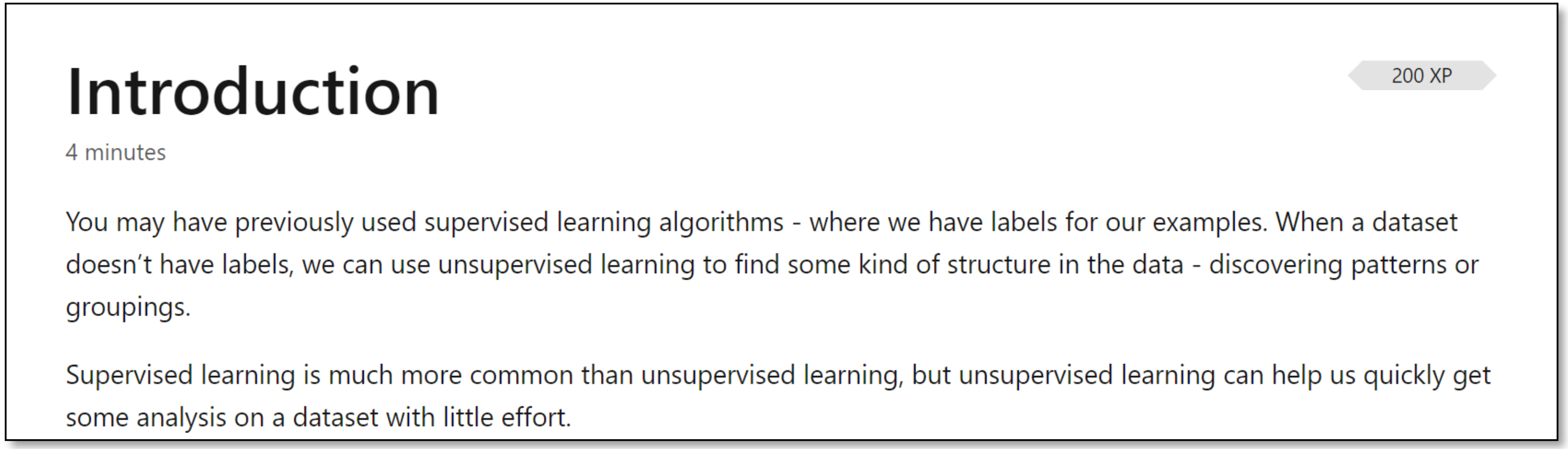
In this module, you will:

- Define unsupervised learning.
- Apply k-means clustering to predict groupings in data.

https://docs.microsoft.com/en-us/learn/modules/introduction-to-unsupervised-learning/

http://faculty.marshall.usc.edu/gareth-james/ISL/

Clustering



Introduction

4 minutes

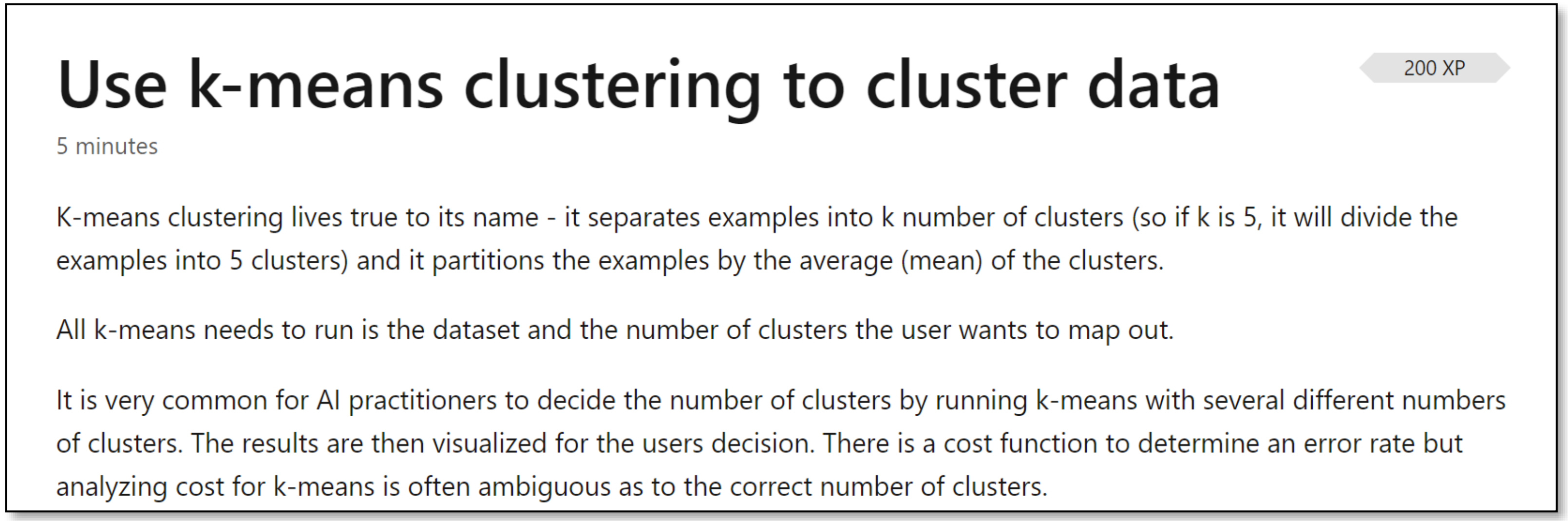
You may have previously used supervised learning algorithms - where we have labels for our examples. When a dataset doesn’t have labels, we can use unsupervised learning to find some kind of structure in the data - discovering patterns or groupings.

Supervised learning is much more common than unsupervised learning, but unsupervised learning can help us quickly get some analysis on a dataset with little effort.

https://docs.microsoft.com/en-us/learn/modules/introduction-to-unsupervised-learning/1-introduction

http://faculty.marshall.usc.edu/gareth-james/ISL/

Clustering



Use k-means clustering to cluster data

200 XP

5 minutes

K-means clustering lives true to its name - it separates examples into k number of clusters (so if k is 5, it will divide the examples into 5 clusters) and it partitions the examples by the average (mean) of the clusters.

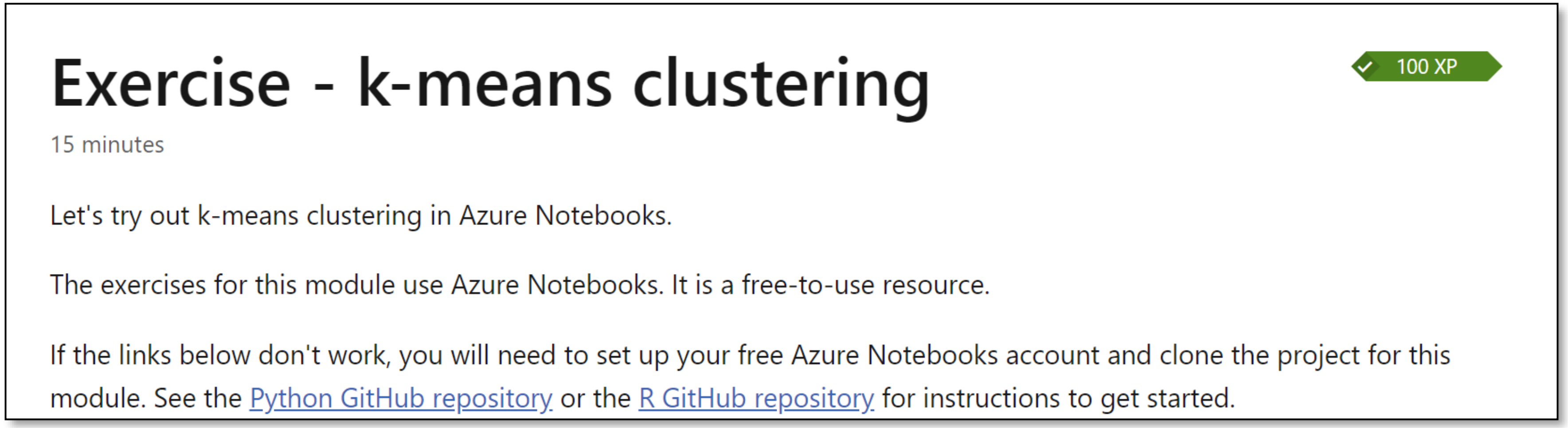
All k-means needs to run is the dataset and the number of clusters the user wants to map out.

It is very common for AI practitioners to decide the number of clusters by running k-means with several different numbers of clusters. The results are then visualized for the users decision. There is a cost function to determine an error rate but analyzing cost for k-means is often ambiguous as to the correct number of clusters.

https://docs.microsoft.com/en-us/learn/modules/introduction-to-unsupervised-learning/2-k-means-clustering

http://faculty.marshall.usc.edu/gareth-james/ISL/

Clustering



Exercise - k-means clustering

15 minutes

Let's try out k-means clustering in Azure Notebooks.

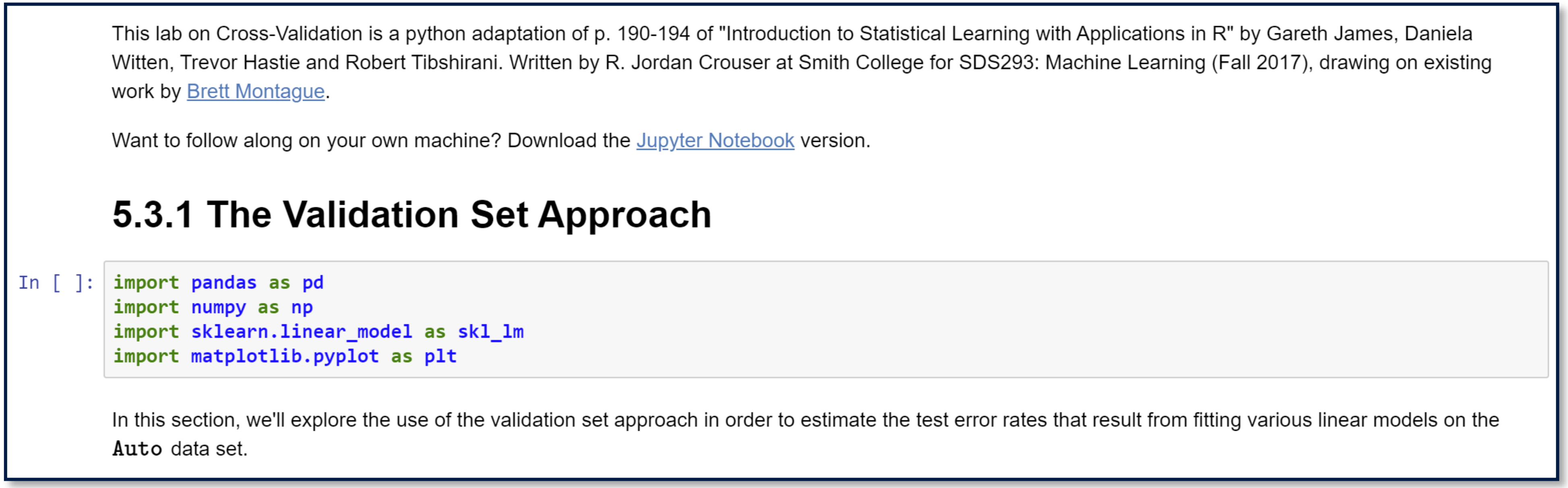
The exercises for this module use Azure Notebooks. It is a free-to-use resource.

If the links below don't work, you will need to set up your free Azure Notebooks account and clone the project for this module. See the Python GitHub repository or the R GitHub repository for instructions to get started.

https://docs.microsoft.com/en-us/learn/modules/introduction-to-unsupervised-learning/3-exercise-k-means

http://faculty.marshall.usc.edu/gareth-james/ISL/

Review + Additional Content



This lab on Cross-Validation is a python adaptation of p. 190-194 of "Introduction to Statistical Learning with Applications in R" by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Written by R. Jordan Crouser at Smith College for SDS293: Machine Learning (Fall 2017), drawing on existing work by Brett Montague.

Want to follow along on your own machine? Download the Jupyter Notebook version.

5.3.1 The Validation Set Approach

In [ ]:

```
import pandas as pd
import numpy as np
import sklearn.linear_model as skl_lm
import matplotlib.pyplot as plt
```

In this section, we'll explore the use of the validation set approach in order to estimate the test error rates that result from fitting various linear models on the `Auto` data set.

http://www.science.smith.edu/~jcrouser/SDS293/labs/lab7-py.html