Introduction to Machine Learning
Week 5 Activities

1 Knowledge Transfer Activities

The main text for Week 5 is "An Introduction to Statistical Learning with Applications in R" available here. There are also online copies available in our HWU library system. An R package `ISRL` is available on CRAN which provides the collection of data sets used in the book.

To access a data set from the `ISRL` pacakge, e.g. `Auto`, first install the `ISRL` package if you have not done so using the command `install.packages("ISRL")`. Then, type `library(ISRL)` in the RStudio command prompt followed by `attach(Auto)` to make the data set available to you in R. Also, look at page 14 of the main text to see all the data sets that are available from the `ISRL` package.

Complete the following activities for Week 5 and post any questions that you have on the Forum/Discussion Board.

1.1 Unit 1:

(a) Read the following sections of the main text:

- Chapter 1 Introduction pages 1-9 just before Who Should Read This Book?
- Chapter 2 Sections 2.1, 2.1.1, 2.1.2, 2.1.3; Sections 2.2, 2.2.1, 2.2.2.
- Chapter 3 Sections 3.1, 3.1.1; Sections 3.2, 3.2.1; This is revision reading.
- Chapter 5 Section 5.1.1 (partly)

(b) Go through Unit 1 Lecture Slides and Video Presentation(s) on the Week 5 Vision page.

(c) Complete the following exercises given in Section 2: 2.1 and 2.2

1.2 Unit 2:

(a) Read the following sections of the main text:

- Chapter 5 Section 5.1.1 (completely)
- Chapter 2 Sections 2.1.5, 2.2.3 just before K-Nearest neighbor.
- Chapter 4 Sections 4.1, 4.2 and 4.3 just before subsection 4.3.5

(b) Go through Unit 2 Lecture Slides and Video Presentation(s) on the Week 5 Vision page.

(c) Complete the following exercises given in Section 2: 2.3 and 2.4

1.3 Unit 3:

(a) Read the following sections of the main text:

- Chapter 2 Section 2.1.4.
- Chapter 10 Sections 10.1 and 10.3.1

(b) Go through Unit 3 Lecture Slides and Video Presentation(s) on the Week 5 Vision page.

(c) Complete the following exercises given in Section 2: 2.5 and 2.6

2 Exercises

2.1 Consider a pair of quantitative variables (X, Y) with a joint PDF given by

$$\begin{aligned} \pi(x, y) &= 2 \quad \text{if } x \geq 0, y \geq 0 \text{ and } x + y \leq 1, \text{ and} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Suppose we observe $X = x = 0.2$ and would like to predict the corresponding Y.

(a) Under the MSE criteria, what is the expression for the best regressor of Y given $X = x$?

(b) Now suppose that $\pi(x, y)$ is unknown to you but we have a data set consisting of iid samples $\{(x_i, y_i),\ i = 1, 2, \cdots, N\}$ from $\pi(x, y)$. Based on the expression derived in (a), determine a class of functions $\mathcal{C}$ that will be optimal in estimating the best regressor. Find the estimate of the best regressor based on $\mathcal{C}$ and the data set $\{(x_i, y_i),\ i = 1, 2, \cdots, N\}$.

(c) Determine a class of functions that is more restrictive compared to $\mathcal{C}$, and another class of functions that is more flexible compared to $\mathcal{C}$.

2.2 The `Advertising` dataset consists of the `sales` of a product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: `TV`, `radio`, and `newspaper`.

(a) Obtain scatterplots of `sales` versus each of the three different media. What do you observe regarding the general trend in these scatterplots?

(b) Find the least squares regression line for each scatterplot and obtain summaries of the fit.

(c) Is simple linear regression an adequate class of models for explaining the data in the scatterplots? Provide relevant diagnostics.

(d) Provide advice with suitable quantitative evidence on what could be the best media to allocate funds in order to increase sales.

2.3 On the flexibility of models and the bias-variance decomposition:

(a) Sketch of typical (squared) bias, variance, training error and test error on a single plot, as we increase the flexibility of the class of models used to fit the data. The x-axis should represent increasing degree of flexibility of the model class. There should be four curves. Make sure to label each one.

(b) Explain why each curve has the shape that you have drawn.

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

2.4 Consider the `Auto` data set from the `ISRL` package.

(a) How many rows are in this data set? How many columns? What do the rows and columns represent?

(b) Obtain a scatter plot of `horsepower` versus `mpg` and comment on the trend.

(c) Obtain the least squares regression line of `horsepower` on `mpg` and plot it on the scatterplot. Comment on the fit visually and based on residual diagnostics.

(d) Use a cross validation procedure to find the best regressor of `horsepower` on `mpg` based on a class of models $\mathcal{C}_p$ where

$$\mathcal{C}_p = \left\{ f(x) \,:\, f(x) = \beta_0 + \beta_1\, x + \beta_2\, x^2 + \cdots \beta_p\, x^p \right\}$$

with range of p in $p_1 \leq p \leq p_2$ where p_1 and p_2 are chosen appropriately by you.

2.5 This question related to the `Weekly` data set, which is part of the ISLR package. The `Weekly` data set contains $1,089$ weekly returns for the S&P 500 stock index for 21 years, from the beginning of 1990 to the end of 2010.

(a) Use the full data set to perform a logistic regression with `Direction` as the response and the five lag variables plus `Volume` as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

(b) Compute the misclassification error rate on the entire data set.

(c) Now fit the logistic regression model using a training data period from 1990 to 2008, with `Lag2` as the only predictor. (HINT: Use the function `filter` from the `dplyr` package to form the training and test sets).

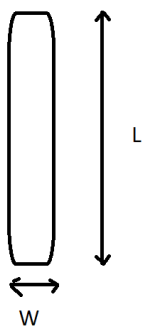
(d) Compute the misclassification error rate on the test data set.

2.6 This exercise illustrates the use of the `kmeans` clustering algorithm on a dataset in the `cluster.datasets` package. This dataset consists of milk composition of various mammals and the aim is to group the mammals according to similarities in the composition of their milk.

(a) Install the `cluster.datasets` package from CRAN and make the datasets available to you in R. Consider the `all.mammals.milk.1956` data set for `kmeans` clustering.

(b) Run the `kmeans` clustering for different number of clusters, K. Choose the optimal K, K^*, based on the elbow criteria.

(c) `kmeans` clustering is sensitive to scaling of the variables.



Consider this example: Let L and W represent the length and width of a skiboard, respectively, both measured in meters. It is clear that $L >> W$. If we want to cluster a collection of skiboards with data on (L, W) using `kmeans`, the clustering will be dominated by L. To avoid this, we scale both variables so that they are comparable. This is done using the R function `scale`. Typing `help(scale)` will give you the details on how this function scales all *appropriate* columns in a data frame so that they are comparable to each other.

(d) Investigate the variables in the `all.mammals.milk.1956` data set. Should the variables be scaled prior to running `kmeans`? Why?

(e) Run the `kmeans` algorithm on the scaled data set. Find the optimal K^* as before. Has your findings changed?

(f) Repeat (b) using GMMs. Use the option `modelNames = "EII"` in `mclustBIC`.