STATS 7022 - Data Science PG
Assignment 2

Trimester 2, 2024

Question 1: Recreating an Analysis

Data scientists are often asked the following question by their colleagues (especially domain experts):

"Can you check this analysis for me?"

Generally, the best way to "check" the analysis performed by someone else is to repeat their entire analysis. Doing so gives you the best possible understanding of what they did and why they did it. In this question, we will recreate a short data analysis report prepared by another data scientist.

Question

You have been provided with a data analysis report produced by another data scientist. The data scientist who prepared the report has forgotten to include their code. You task is to recreate the report, making sure to include all code and all relevant output.

A copy of the `diamonds` data used in the analysis is available on MyUni, or can be downloaded here: `diamonds.rds`

A copy of the original analysis prepared by the other data scientist is available on MyUni, or can be downloaded here: `DS_A2_Q1.html`

For further information, see the assessment rubric on MyUni.

[Question total: 12 marks]

Submission

A single html file containing your recreation of the original data analysis report, including all code and relevant output.

Checklist

Please ensure that:

- Your submission includes all `R` output and plots to support your answers where necessary.
- Your submission includes all `R` code.
- Your submission includes a caption for every plot and table.
- Your submission is a single html file - correctly oriented, clear, and legible.

- You have submitted your solution by online submission to the correct portal on MyUni. Submissions will not be marked and will receive a grade of 0 if they are not made through MyUni or if they are not made to the correct portal.

The course's regular late policy applies to this assignment. If you need to request an extension, or request a variation to the assessment on medical or compassionate grounds, please contact the course coordinator.

Question 2: Leave-One-Out Cross-Validation and Linear Regression

An important part of being a data scientist is understanding the mathematics that underpins the methods we use. One advantage of this is being able to recognise when methods can be simplified and when shortcuts exist. In this question, we will consider one such shortcut when using leave-one-out cross-validation with linear regression.

Question

Cross-validation is a model validation method used to determine how well a given statistical model will generalise to independent (new) data. In a week 4 topic video, we saw a particular result for leave-one-out cross-validation and linear regression. In this question, we will derive this result.

(a) Consider an invertible $p \times p$ matrix A and a vector $\boldsymbol{v} \in \mathbb{R}^p$ (that is, $\boldsymbol{v}$ is $p \times 1$). Show that

$$\left(A - \boldsymbol{v}\boldsymbol{v}^T\right)^{-1} = A^{-1} + \frac{1}{1 - \boldsymbol{v}^T A^{-1} \boldsymbol{v}} A^{-1}\boldsymbol{v}\boldsymbol{v}^T A^{-1}.$$

Hint: Recall that for matrices J and K, $K = J^{-1}$ if $JK = KJ = I$.

[5 marks]

(b) Consider an $n \times p$ matrix X. Let $\boldsymbol{x}_i^T$ be the i^{th} row of X and let $X_{(i)}$ be the matrix X with row i removed. Using part (a), show that

$$\left(X_{(i)}^T X_{(i)}\right)^{-1} = \left(X^T X\right)^{-1} + \frac{1}{1 - h_{i,i}} \left(X^T X\right)^{-1} \boldsymbol{x}_i \boldsymbol{x}_i^T \left(X^T X\right)^{-1}$$

where

$$h_{i,i} = \left[X \left(X^T X\right)^{-1} X^T\right]_{i,i} = \boldsymbol{x}_i^T \left(X^T X\right)^{-1} \boldsymbol{x}_i.$$

Hint: You may assume that $X^T X = \sum_{j=1}^{n} \boldsymbol{x}_j \boldsymbol{x}_j^T$.

[3 marks]

(c) Consider the regression model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$. The i^{th} deletion residual is defined by $\tilde{\epsilon}_i = y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)}$ where $\hat{\boldsymbol{\beta}}_{(i)}$ is the least squares estimate of $\boldsymbol{\beta}$ calculated from the dataset with the observation $(y_i, \boldsymbol{x}_i)$ removed. Using part (b), show that

$$\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)} = \frac{1}{1 - h_{i,i}} \left(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} - y_i h_{i,i}\right).$$

Hint: Recall that $\hat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T \boldsymbol{y}$ and so $\hat{\boldsymbol{\beta}}_{(i)} = \left(X_{(i)}^T X_{(i)}\right)^{-1} X_{(i)}^T \boldsymbol{y}_{(i)}$. You may assume that $X_{(i)}^T \boldsymbol{y}_{(i)} = X^T \boldsymbol{y} - y_i \boldsymbol{x}_i$.

[5 marks]

(d) Hence, show that

$$\tilde{\epsilon}_i = \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}}{1 - h_{i,i}}$$

and give an expression for the leave-one-out cross validated estimate of the MSE ($CV_{(n)}$) for the linear regression model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

[3 marks]

[Question total: 16 marks]

Submission

A single pdf file containing your solutions. You may handwrite your answers and scan them, or typeset your answers.

Checklist

Please ensure that:

- Your submission is a single pdf file - correctly oriented, clear, and legible.
- You have shown all of your working, including probability notation where necessary.
- You have submitted your solution by online submission to the correct portal on MyUni. Submissions will not be marked and will receive a grade of 0 if they are not made through MyUni or if they are not made to the correct portal.

The course's regular late policy applies to this assignment. If you need to request an extension, or request a variation to the assessment on medical or compassionate grounds, please contact the course coordinator.

Question 3: QDA Function

One of the best ways to understand an algorithm is to implement it yourself. In this question, we will write a function to perform Quadratic Discriminant Analysis (QDA).

Question

In `R`, write a function called `get_QDA()`. Your function should take three vector inputs:

- `x`: a vector of predictors.
- `y`: a vector of classes.
- `x0`: a vector of `new_data` predictors.

Your function should return a vector of predicted classes for `x0`.

Here is an example to illustrate its use:

```
x <- 8:0
y <- rep(LETTERS[26:24], each = 3)
x0 <- c(7,5)
get_qda(x = x, y = y, x0 = x0)
```

```
## [1] "Z" "Y"
```

Here is a template you can use:

```
# Function to perform QDA
#
# INPUT
# Three vectors:
#   x: predictors
#   y: classes
#   x0: new predictors
#
# OUTPUT
# A vector of predicted classes for x0.

get_QDA <- function(x, y, x0) {
  # YOUR CODE HERE
}
```

Your function should also ensure that the correct type of input is passed to the function and that edge cases are handled appropriately. For further information, see the assessment rubric on MyUni.

[Question total: 8 marks]

Submission

A single `R` file containing your function. Your code will be automatically unit-tested, so any changes to names etc. may result in a loss of marks. Your function must not use QDA functions from other packages - the code must be your own work.

Checklist

Please ensure that:

- You have submitted your solution by online submission to the correct portal on MyUni. Submissions will not be marked and will receive a grade of 0 if they are not made through MyUni or if they are not made to the correct portal.

The course's regular late policy applies to this assignment. If you need to request an extension, or request a variation to the assessment on medical or compassionate grounds, please contact the course coordinator.

Question 4: Technical Communication

There are four key skills in a data scientist's arsenal: analysis, mathematics, coding, and communication. In this question, we'll practice your communication skills.

Whenever you are communicating, it's important to consider your audience - different audiences will need to know different information, and be able to understand different levels of technical detail.

Question

Suppose you are a professional data scientist who works predominantly in `R` using `tidymodels`. Your manager is considering getting your whole team to use `tidymodels`, and has asked you to prepare a single-page overview of `tidymodels` to help them make a decision.

Your overview should:

- describe the different components of a `tidymodels` workflow;
- explain the types of data and models that can be used with `tidymodels`;
- explain the advantages of adopting `tidymodels`; and
- discuss any disadvantages of using `tidymodels`.

You may assume that your manager has about the same level of knowledge of data science and `R` as a student in STATS 7022 Data Science PG (except that they are unfamiliar with `tidymodels`!). Marks will be allocated for the quality and correctness of your writing, as well as how suitable your writing is for the intended audience. Your report must use Harvard-style referencing as detailed here:
https://www.adelaide.edu.au/library/referencing-support

Your overview should not exceed one page in length (excluding references). For further information, see the assessment rubric on MyUni.

[Question total: 12 marks]

Submission

A single doc or pdf file containing your report. Your report should be one A4 page (excluding references), with 10-12pt font, single line spacing, and standard page margins.

Checklist

Please ensure that:

- Your submission is a single doc or pdf file - correctly oriented, clear, and legible.

- Your submission is not longer than 1 page (including tables and figures, but excluding references).
- Your submission is correctly referenced. For more information, see the link above or visit the Writing Centre.
- You have submitted your solution by online submission to the correct portal on MyUni. Submissions will not be marked and will receive a grade of 0 if they are not made through MyUni or if they are not made to the correct portal.

The course's regular late policy applies to this assignment. If you need to request an extension, or request a variation to the assessment on medical or compassionate grounds, please contact the course coordinator.