STATS 7022 - Data Science PG
Assignment 3

Trimester 2, 2024

Question 1: Modelling with Data

We have seen many predictive models in this course, as well as how to fit them. In this question, we will fit a predictive model to a dataset.

Question

You have been provided with the `diamonds2` dataset - a copy of the data is available on MyUni. Read this dataset in to `R` and fit a random forest model to predict log(`price`) in terms of all other variables in the dataset.

In addition to fitting the model, you code should perform the following:

- split the data into training and testing sets;
- tune the model; and
- perform a final test of the final model.

For further information, see the assessment rubric on MyUni.

[Question total: 13 marks]

Submission

A single pdf file detailing the fitting of your random forest model.

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 pdf 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: Piecewise Cubic Splines

An important part of being a data scientist is understanding the mathematics that underpins the methods we use. The deepest understanding includes recognising why certain methods have certain forms and recognising when different forms are equivalent. In this question, we will investigate some properties of piecewise cubic splines.

Question

(a) Consider a piecewise cubic spline with k knots that is continuous up to (and including) the second order derivative. Show that the number of parameters needed to describe this piecewise cubic is $k + 4$.

[5 marks]

(b) Consider a piecewise cubic spline with a single knot ξ. This can be represented in two ways:

Method 1
A linear combination of the basis functions:

$$\begin{aligned} h_1(x) &= 1 \\ h_2(x) &= x \\ h_3(x) &= x^2 \\ h_4(x) &= x^3 \\ h_5(x) &= (x - \xi)^3_+ \end{aligned}$$

Method 2
Two cubic functions

$$\begin{aligned} f(x) =& a_1 + b_1 x + c_1 x^2 + d_1 x^3, \quad x < \xi \\ g(x) =& a_2 + b_2 x + c_2 x^2 + d_2 x^3, \quad x \geq \xi, \end{aligned}$$

such that

$$\begin{aligned} f(\xi) &= g(\xi) \\ f'(\xi) &= g'(\xi) \\ f''(\xi) &= g''(\xi) \end{aligned}$$

Show that Method 1 implies Method 2.

[7 marks]
[Question total: 12 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: k-Means

One of the best ways to understand an algorithm is to implement it yourself. In this question, we will write a function to implement the k-means algorithm.

Question

In `R`, write a function called `get_kmeans()`. Your function should take two inputs:

- `x`: a numeric vector.
- `k`: the number of groups.

Your function should return a vector of labels indicating group membership for each value in `x`. You should format your labels as follows:

- the labels of the groups should be the numbers $1, \ldots, k$.
- the groups should be ordered such that group 1 has the largest mean, group 2 has the second largest mean, and so on, up to group k has the smallest mean.

Here is an example to illustrate its use:

```
x <- rep(2:4, each = 3)
get_kmeans(x = x, 3)
```

```
## [1] 3 3 3 2 2 2 1 1 1
```

Here is a template you can use:

```
# Function to implement k-means algorithm
#
# INPUT:
#   x: numeric vector
#   k: number of groups
#
# OUTPUT
# A vector of integer labels indicating group membership.

get_kmeans <- function(x, k) {
  # 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: 7 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 k-means 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.