MLNN3 — Assignment 3

19th November 2024

The following two questions are based on the Chapters 3 and 4 of the lecture notes. For this assignment, you are required to write your answers in R Markdown so you can practice some of the R coding needed for the practical exam. The template for the R Markdown file is provided on the course website (Assignment_3_template.Rmd).

Please submit to Gradescope via the link provided on the course website. The deadline is 12:00 noon on the 2nd December.

Question 1

We are using a k-nearest neighbours model to predict the response variable y based on the predictor variables x_1 and x_2. We want to assign a value to the point with $x_1 = 1$ and $x_2 = 2$. Here are the nearest neighbours from the training set:

x_1	x_2	y
1	1	0.65
2	2	0.53
2	1	0.42
3	2	0.31
0	3	0.75

1(a)

What is the predicted value of y for the new point using the k-nearest neighbours model with $k = 3$ using the Euclidean distance?

1(b)

What is the predicted value of y for the new point using the k-nearest neighbours model with $k = 3$ using the Manhattan distance?

1(c)

Find the predicted value of y for the new point using the weighted variant of the k-nearest neighbours model with $k = 3$, Manhattan distance and the inverse distance weighting.

Question 2

In this question, we will utilise the `wine` dataset in R. The dataset contains information about 1,599 Portuguese red wines.

```r
wine <- read.csv(paste0("https://www.maths.dur.ac.uk/users/john.p.gosling/",
                        "MATH3431_presentations/winequality-red.csv"))
```

2(a)

The variable `quality` is listed as a numeric variable. Convert this variable to a binary factor with levels "bad" and "good". A wine is considered "good" if the quality is greater than or equal to 7.

2(b)

Perform a 60/40 split of the dataset into a training set and a test set.

2(c)

Using `rpart`, fit a decision tree model to the training set to predict the binary quality of the wine based on the 11 explanatory variables. What is the accuracy of the model on the test set? Provide a visualisation of the decision tree.

2(d)

Fit a k-nearest neighbours model to the training set using the `caret` and `class` packages. Use 15-fold cross-validation to determine the optimal value of k from the range 2 to 10. What is the optimal value of k and the associated accuracy of the model on the test set? Would you prefer to use the decision tree or the k-nearest neighbours model for this dataset?