MLNN3 — Assignment 2

5th November 2024

The following two questions are based on the Chapters 2 and 3 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_2_template.Rmd).

Please submit to Gradescope via the link provided on the course website. The deadline is 12:00 noon on the 18th November.

Question 1

1(a)

Consider a simple linear regression model with Gaussian errors. We simulate data from the model

$$Y = X\boldsymbol{\beta} + \epsilon,$$

where $\epsilon \sim \text{Normal}(0, \sigma^2)$, using the code in the Markdown template.

Fit 11 models using the simulated data and the `lm` function in R. The first model should include only the intercept, the second model should include the intercept and the first predictor, the third model should include the intercept and the first two predictors, and so on.

1(b)

Create a plot with the adjusted R^2 on the x-axis and the AIC on the y-axis. What do you observe?

(Hint: if our model is called `lm1`, we can extract the adjusted R^2 using `summary(lm1)$adj.r.squared` and the AIC using `AIC(lm1)`)

1(c)

We posit the following model for the relationship between the response variable Y and the predictor variable X:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where $\epsilon \sim \text{Lapace}(0, \sigma)$. Let's say that we have n observations, then the maximum likelihood estimate for σ is given by:

$$l(\hat{y}, \hat{\sigma}; x) = \frac{1}{2\hat{\sigma}} \exp\left(-\frac{|y - \hat{y}|}{\hat{\sigma}}\right),$$

where $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Let's say that we have n observations, then the maximum likelihood estimate for σ is given by:

$$\hat{\sigma} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|.$$

Write out an expression for the AIC in terms of the mean absolute error.

Question 2

In this question, we will utilise the `mtcars` dataset in R. The dataset contains information about 32 cars from the 1974 model year.

2(a)

Five of the variables are more factors than numeric variables. Identify these variables, and convert them to factors.

2(b)

We are going to try to predict the number of cylinders (`cyl`) based on the remaining variables. Split the dataset into a training set and a test set. The training set should contain 70% of the data and the test set should contain the remaining 30%.

2(c)

Fit a k-nearest neighbours model to the training set using leave-one-out cross validation to determine k. What is the optimal value of k and the associate accuracy of the model (for both the training and test set)?