K-Nereast Neighbours

Lab 3

Jay Paul Morgan

In this lab, we're going to create a K-Nearest Neighbour (KNN) classifier. This lab is optional, and will not be marked.

Download the Data

In creating the KNN, we'll use another toy dataset, the wine quality dataset https://archive.ics.uci.edu/dataset/186/wine+quality. Specifically, the red wine.

To get access to the dataset, click on the download button. Inside the now downloaded folder, there will be two CSV files, one for red wine, and another for white wine. For this lab, we'll be using the red wine dataset.

Load the Data

The next task is to load the data in Python. For this section, we'll want to write some code to load the red wine CSV file as a pandas dataframe.

Split the data

Now that we've loaded the data as a pandas dataframe, we'll partition the data off into a train and test subset.

Randomly sample 70% of the data for training, and the other 30% will be used for testing.

Normalise the data

To ensure all of the columns are within the same range, we'll normalise them (except from the score column of course!).

We can normalise a column x by the following equation:

$$\text{normalise}(x) = \frac{x - \min_x}{\max_x - \min_x}$$

Therefore, we'll need to calculate the min and max values of each column in the training subset, then apply the normalising function, using these values.

Looking at the correlations

Using the training subset, we'll want to find out which columns to use in our classifier.

We'll visually inspect the correlations between the input columns and the target score column.

Create a series of 2D scatter plots with an input column across the x-axis, and the target score column across the y-axis.

Which of these columns show a strong relationship between the input and target columns?

Create a K-Nearest Neighbour

Create a KNN classifier and $k = 3$, using some of the input colums (the best ones as decided by the correlation plots).

Create Classification performance metrics

Create a series of classification metrics for the new KNN classifier:

1. F_1 score (averaged over all classes)
2. Plot a confusion matrix.

Plot performance metrics for different k

Using the KNN classifier, modify the k value from 1 to N and plot the F_1 score for each value of k. What is the optimal value for k?