Department of Mathematics, College of Engineering,
Design and Physical Sciences, Brunel University

Fundamentals of Machine Learning

Academic Year 2022 - 2023

Assignment Report

By David Blair
Lecturer: Dr Simon Shaw

1 Task 1

The k-NN method is a machine learning method designed for problems involving classification and regression although it is more commonly used for classification. When it is given a new input vector, it locates the k nearest neighbours and uses the majority output class as the resulting prediction.

It is non-parametric meaning it makes no assumptions about the underlying distribution of the data. It is a lazy learning method meaning that it does not fit against a particular model but stores the entire training data. This makes training fast but may mean predictions are slow.

There are three common hyperparameters: the number of neighbours, the distance metric and the weighting scheme. The number of neighbours (the k in k-NN) indicates how many neighbours to consider when classifying the new input vector. The distance metric is how the classifier determines proximity to the input vector. Finally, the weighting scheme allows us to potentially weight each neighbour by a given number. A common weighting scheme is to use the inverse of the distance metric in order to allow closer neighbours to be weighted higher.

It was decided to make the train test split as 30% for the testing and 70% for the training. This is because it gave a sufficient number of testing samples to make the accuracy score meaningful. We also decided to stratify by the response column to make sure we had enough of each category to improve the quality of the results.

In order to calculate an estimate for $\mathbf{P}(P|-)$ we took the number of times the classifier predicted Benign when the true label was Malignant and divided it by the sum of both times where the prediction was Benign.

2 Task 2

~~Principle~~ Principal Component Analysis is a method of dimensionality reduction that attempts to maximise the explained variance along new dimensions or "principle components". Given an n-dimensional dataset, we first need to calculate the covariance matrix:

$$\mathbf{\Sigma} = \begin{bmatrix} \operatorname{cov}(x_1, x_1) & \operatorname{cov}(x_1, x_2) & \cdots & \operatorname{cov}(x_1, x_p) \\ \operatorname{cov}(x_2, x_1) & \operatorname{cov}(x_2, x_2) & \cdots & \operatorname{cov}(x_2, x_p) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(x_p, x_1) & \operatorname{cov}(x_p, x_2) & \cdots & \operatorname{cov}(x_p, x_p) \end{bmatrix} \tag{1}$$

You can think of this as a matrix that encapsulates the correlations between each dimension. If $\operatorname{cov}(x_p, x_q) > 0$, the variables increase or decrease together. If $\operatorname{cov}(x_p, x_q) < 0$ one increases as the other decreases. The $\operatorname{cov}(x_p, x_p)$ is given by:

$$\operatorname{cov}(x_p, x_q) = E[(x_p - \mu_{x_p})(x_q - \mu_{x_q})] \tag{2}$$

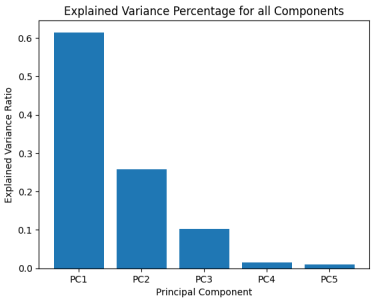

Figure 1: Scree Plot Example.

The covariance between a variable and itself is the variance of that variable. The next step is to calculate the eigenvalues and eigenvectors of the covariance matrix using the following formula:

$$\mathbf{\Sigma v} = \lambda \mathbf{v} \tag{3}$$

The eigenvectors form a new basis (or eigenbasis) for a coordinate system that maximises the variance along each dimension. If we order the eigenvectors in decreasing order of size according to their eigenvalues, we can determine the amount of variance explained along that dimension. This allows us to produce a scree plot (see figure 1) that exhibits the amount of variance explained by each dimension, determined by the eigenvalue. Lastly, we need to project our data $\boldsymbol{X}$ onto the new coordinate system of the eigenbasis:

$$\boldsymbol{Y} = \boldsymbol{XV} \tag{4}$$

Where $\boldsymbol{V}$ is the matrix of k eigenvectors. To reduce the dimensions of the data, we can then remove the number of dimensions we want starting with those with the lowest eigenvalue / explained variance.

To calculate the total variance explained for 4 principle components, we calculate the total sum of the variance explained by all 4 components as 98.97%.

The dimensionality reduction increased the accuracy from 88.89% to 91.23% and reduced the false negative rate from 15.38% to 12.31%. This is most likely because the PCA captured the most important parts of the dataset, eliminating the less important. This means that I would recommend using 4-principle components.

3 Task 3

The singular value decomposition (SVD) allows you to decompose a matrix of any rank into the sum of rank-1 matrices. It allows you to do many things such as matrix approximation, matrix compression and dimensionality reduction. The theorem is as follows:

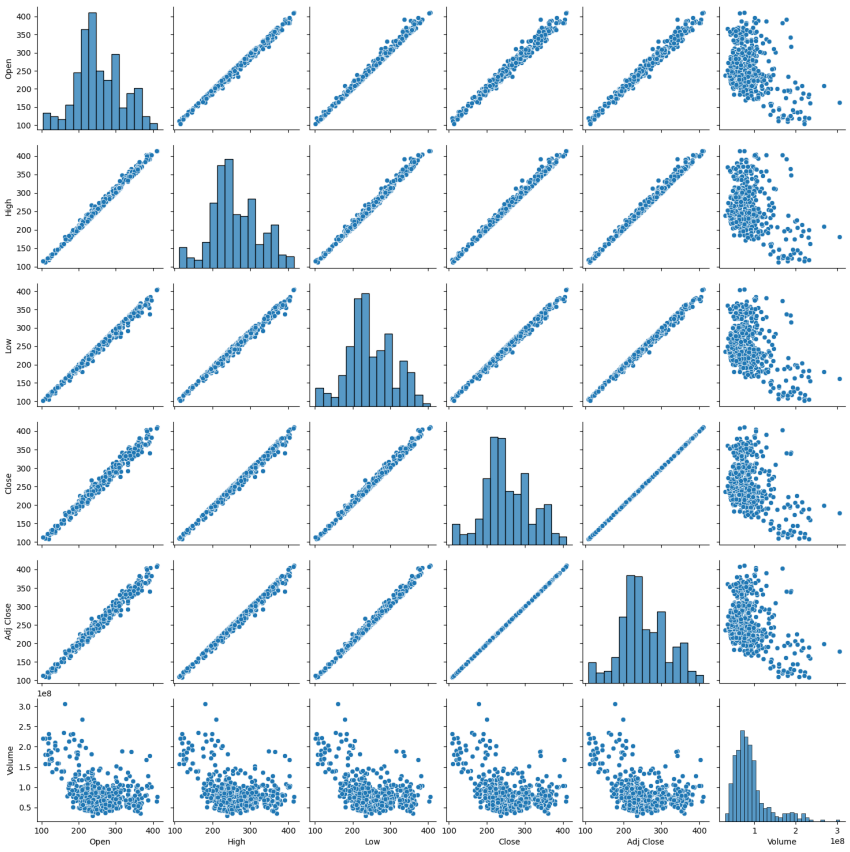

Figure 2: Pair Plot of Tesla Stock Data (Historic)

Theorem 3.1 *(SVD Theorem)* Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ be a rectangular matrix of rank $r \in [0, min(m, n)]$. The SVD of $\boldsymbol{A}$ is a decomposition of the form:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}$$

with an orthogonal matrix $U \in \mathbb{R}^{m \times m}$ with column vectors $\mathbf{u}_i$, $i = 1, \ldots, m$, and an orthogonal matrix $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ with column vectors $\mathbf{v}_j$, $j = 1, \ldots, n$. Moreover, $\boldsymbol{\Sigma}$ is an $m \times n$ matrix with $\Sigma_{ii} = \sigma_i \leq 0$ and $\Sigma_{ij} = 0$, $i \neq j$.

Figure 2 is the pair plot of the Tesla stock data for the 5 variables under consideration. It looks as though all variables are correlated with on another excluding the Volume. This would indicate that two dominant independent components lie in this data. The dataset TSLHistory.csv has a rank of 5 after dropping all non-numeric columns. The variable Kc in the notebook is a rank-c approximation of the original dataset.