SID: 480386401

Executive Summary

Data3888 Disciplinary Assignment 2

Overarching Question

*"Is SVM **always** a more appropriate method of predicting acute rejection in kidney transplant patients given the following gene set, than KNN?"***.**

After intial analysis of the gene sequence data set GSE120396 in lab 6, it was obvious that the modelling SVM was more accurate in predicting failures from the current gene sequence than SVM. It is not clear whether this is always the case, or whether KNN given optimal input values would provide a more suitable method of modelling.

Key Techniue

Although there are 2 different models mentioned in the question, the main technique that will be focussed on in this study is the KNN model.

The KNN approach is short for "K nearest neighbours". This technique uses an algorithm on labelled data, in our case we have various genes and their expression values as data, with the label we're interested in being whether the the patient's kidney transplant was successful. The algorithm is able to make decisions for future data that is unlabelled as it compares the unknown data to the labelled data it is familiar with by taking it's nearest neighbours. For every feature in the known dataset (the number of genes), an n- dimentional graph is created. The algorithm assumes that labels will group together on the graph, meaning when an unknown data point is asked to be labelled based on its closest k data points, the most popular neighbout label will be assigned.

Approach:

A data set of 88 patients, detailing 11721 genes and their numeric expression values, describes a group of patients that have undergone a kidney transplant, 66 labelled "No", i.e. there was no kidney failure after the procedure, and 22 "Yes" denoting that there was. In order to determine whether SVM or KNN is a more appropriate model to predict future data, the first thing we must be able to do is visualise the accuracy of each model.

As we do not have additional data in the same format, we are made to both test and create the KNN and SVM models with the same data. This means that using r studio we will be using a k fold technique, meaning the data is split randomly into a set number of folds, and the models will be tested on a split section of the initial data used to create the models. This will be done using the programming language r.

Once the results of these models have been obtained, we will be able to assess the accuracy of each. The accuracy is defined as the total number of labels the model got correct out of the total number of labels it had to predict. In order to both get a more reliable accuracy figure and a way to visualise the data, we will repeat the process of getting accuracy a set number of times. When we have fully completed the accuracy readings, we will use the various accuracy values to create a boxplot, and get the overall average accuracy.

As the value of nearest neighbours can affect the overall accuracy of the knn model, we want to observe the whether the difference in accuracy in the knn and svm models, changes as we select different values of nearest neighbours. In order to observe this we will need to repeat the accuracy assessment process multiple times, and compare the values and boxplots over a variety of different nearest neighbour values.

Finally we will observe whether the observed trend is the same for different values given for the number of folds, and for the number of overall repeats. The number, although changing, will stay the same for both the knn and svm models to keep the test fair. If the same trend remains, then the conclusion of the experiment will be that regardless of the input values given, provided they are the same, SVM will be a more suitable method of modelling kidney failure given the following genes than KNN.

Potential Shortcomings

The way that the process decribes retrieving modelling can be very computationally expensive. SVM and KNN over a large data set alone will take a lot of processing time. There becomes a point at which modelling can be unfeasible. In order to combat this short coming, input values for creating the models will have to be capped at a certain point. The nearest neighbours value in the KNN algorithm is known to increase the accuracy the higher it is, before eventually dropping off, whilst also being computationally expensive. A potential negative is that if the gap in accuracy closes as k increases, it will not be known whether a certain value of nearest neighbouts trumps SVM, if that value is over the threshold. In order to combat this, effort will be made to make computation as efficient as possible, and trends will be noted to determine whether this is a possibility of not.

Shiny App

Link: https://github.com/ddur4870/data3888_Discipline_project_2

Clone Link: https://github.com/ddur4870/data3888_Discipline_project_2.git

The question being explored requires a lot of graphical visualisations, rather than a normal report having to observe many graphs, in a confusing format, the shiny application is able to graph these boxplots and display different measures of accuracy in a comprehensible way.

The app allows the user to choose the desired model, and will display the boxplot of the average accuracy given other inputs that are on the side panel. The inputs are given default values of nearest

neighbours = 5, number of repeats = 10, number of folds = 5. These can be changed as the user desires.

If the checkbox in the side panel is ticked, the accuracy trend for each model as certain inputs change are displayed, giving the user a glimpse at the overall trend in accuracy as user inputted values change.

The user should be able to easily observe the difference in accuracy given different user input, allowing them to make the assessment of which modelling method is the best, thus answering the question outlined in this study.