INFO304 Assignment 3 2022

Ethan Smith - 5652106

QUESTION ONE (15 MARKS)

Assume that you do not know the K-means clustering algorithm, but want to build a clustering tool that finds K cluster centres given a table of numeric data with F features (i.e. F explanatories, but no response). You decide to use a genetic algorithm (GA) to evolve the solution for the cluster centres.

1. Given some K value (number of clusters to evolve) and number of features F of the data, describe an appropriate representation for an individual.
2. What is the fitness function for this problem? That is, what makes a good solution to a clustering algorithm, and how is this evaluated? Assume that you have a data table of N rows and F numeric feature variables.

QUESTION TWO (25 MARKS)

You are tasked with using a k-nearest neighbour (kNN) model to do prediction, but the dataset you are using has a large number of explanatories. Assume you have N explanatory variables and a response Y.

Rather than doing more traditional feature reduction/selection methods, you decide to use a Genetic Algorithm (GA) to search for the best M features (M <= N) to optimise the predictive behaviour of the kNN model for a FIXED value of k.

Describe how you would set up this combined GA and kNN model to search for the "best" kNN model for a specific dataset and set value of k. Ensure you include:

1. A representation (and how you would use the representation) for the GA and kNN;
2. The fitness function and example search operators (crossover, mutation);

Now assume that you also want to optimise the number of neighbours k.

3. Define a representation that would allow you to optimise both the value of "k" and the M features during the evolution. Ensure you explain how the representation is interpreted and used within the model.

QUESTION THREE (60 MARKS) - Investment Portfolio Management

This section deals with modelling the selection of stocks, bonds and cash to make up an investment portfolio. Load the data as follows: invest <- read.table("invest.tab")

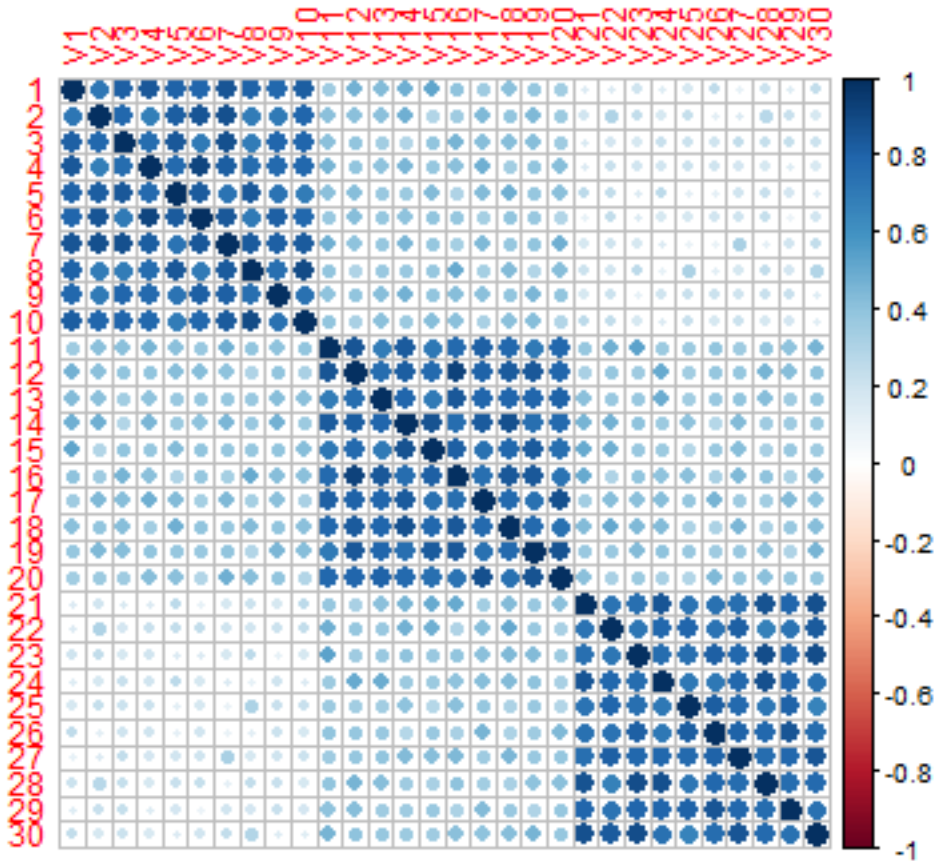
The ROI column is the percentage predicted return on investment, the Risk column is a measure of the risk associated with this particular investment, and the Type indicates the type of investment. Note that each row is labelled with the type of investment and a number, so that you can (if needed) refer to individual investments.

1. Visualise and discuss the different ROI and Risk associated with Stocks, Bonds and Cash. In addition, load, visualise and give a simple interpretation for the correlation table (HINT: See part 5 for details of "corr.tab").

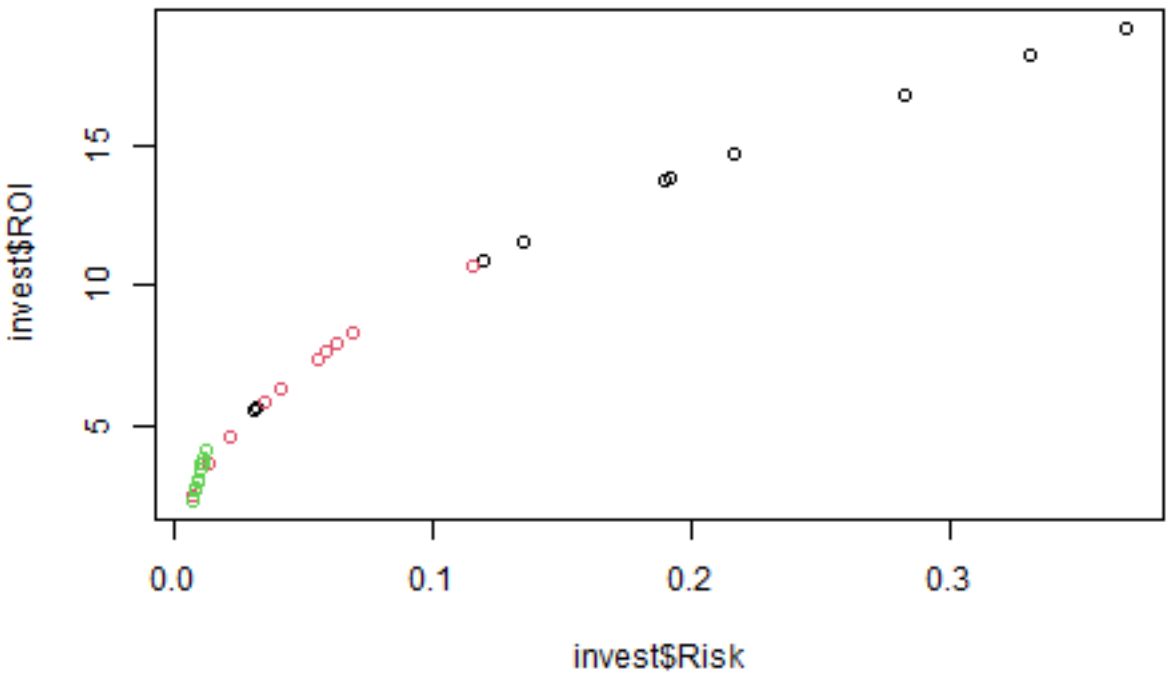
```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
invest <- read.table("invest.tab")
invest.corr <- read.table("corr.tab")
corrplot(as.matrix(invest.corr))
```




```
plot(invest$Risk, invest$ROI, col = invest$Type)
```



2. Run the "invest.R" script. This script does a multi-objective criteria analysis to determine the best mix of stocks, bonds and cash over a range of tradeoffs. Describe in words what the "invest.R" script is doing. Ensure that you address how a solution is represented and interpreted, and the relationship between the solution space, the objective space and the constraints.

3. Table 1 (below) shows the recommended blend of stocks, bonds and cash for a number of different brokerage houses (i.e. businesses that take your money and invest it to give you a return).

Using the result of the nsga2 model you have previously run, examine and present the blend of stocks, bonds and cash for a low risk, moderate risk and high risk investment blend (just pick one from each general category). Discuss, in relation to Table 1, the level of risk that seems to be taken by the brokerage houses and whether the one year return performance is related to the associated risk of the brokerage house.

4. Examine the plot shown in Figure 1. This shows how the percentage of bonds, stocks and cash vary as you move along the pareto front from the least to greatest percentage return. Write an R script to produce this figure given the output from nsga2. Submit your R code and the associated figure.

5. A matrix representing the estimated correlation between any two investments is given in the table corr.tab. For our example with 30 investments this is a 30x30 table. A good portfolio should aim for investments where correlation of behaviour is minimised, so that if some of the investments are decreasing, others in the portfolio may behave differently. This reduces the risk of the portfolio as a whole losing money and should result in more stable returns.

Extend the model in invest.R to have an additional objective which is to minimise the correlation for the investment blend.

a) Describe in words the approach you have chosen and explain the rational for the model.

b) Include your R code for the function that calculates this objective.

c) Produce pairwise plots for the 3 objectives after a successful optimisation.

d) Discuss and justify what type of tradeoff of risk/return/correlation you would choose for a portfolio based on the results from part (c).