Machine Learning Report

ST 3189 Machine Learning 2021-22

Student Number: 200615086

Table of Contents

1. Introduction

1.1 The Background

Data is constantly generated and accumulated. Machine learning has made it possible to use this data to predict the future. Machine learning allows computers to learn on their own, keeping the model's performance constant. This is largely divided into three categories: supervised learning, unsupervised learning, and reinforcement learning. Among them, this project deals with supervised learning and unsupervised learning. Supervised learning means that computers are taught and supervised, which is usually called labeling. First, when a label indicating the correct answer is provided, it predicts the label not given through training. In other words, supervised learning helps prediction by providing the computer with the final result value of the data. However, unsupervised learning does not provide the correct answer, that is, Label. The unsupervised learning model is created through learning to analyze characteristics between data to distinguish differences within it and find patterns. The ultimate goal of this learning model is to predict which group or cluster the observations will belong to.

1.2 The Purpose

Through this report, various machine learning models will be introduced, and which models have high predictive performance when applied to actual data.

1.3 The Scope

Next, I will introduce the types and models used in each part.

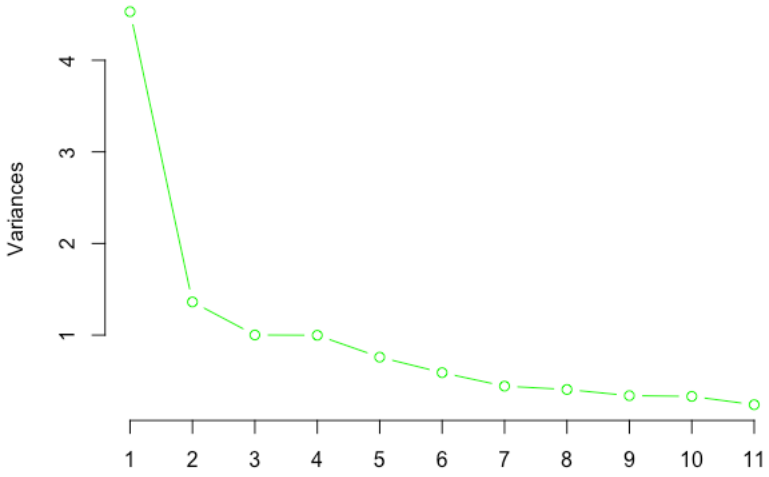
1) Part 1: Unsupervised Learning
   A. Principal Component Analysis (PCA)
   B. K-Means Clustering
2) Part 2: Supervised Learning
   A. Linear Regression
   B. Random Forest
   C. Support Vector Regression (SVR)
3) Part 3: Supervised Learning
   A. Rogistic Regression
   B. Random Forest
   C. Support Vector Classifier (SVC)

2. Data Analysis

2.1 Part 1

Part 1 uses Unsupervised Learning techniques among the three machine learning types mentioned above. First, the given dataset contains quite a lot of variables. Even if gender and age are excluded as categorical variable, the total number of variables is 9, which means that it is high-dimensional data including many variables. High-dimensional data results in inefficient modeling due to complex calculations, and visual representation is also significantly difficult. Therefore, machine learning modeling reduces the dimension by selecting or extracting only important variables. In this part, the dimension will be reduced and grouped through K-means clustering using Principal Component Analysis (PCA), a representative analysis technique of unsupervised feature extraction.

In short, PCA summarizes high-dimensional data into data consisting of k-variables that are not correlated. The summarized variable, scores, is generated as a linear combination of existing variables. In other words, this technique finds a new axis that can preserve the variance of the original variable as much as possible, and then projects the data on that axis. There are two ways to choose how many principal components to use. Select the number of main components corresponding to the elbow point in which the change rates of the eigenvalues are slowed or select the minimum main component that stores the cumulative report of variance at a certain level (usually 70-80% or higher).

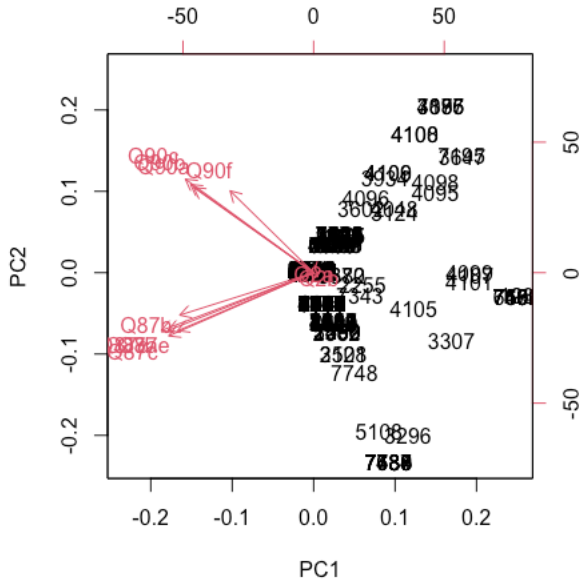


<Figure 1> Scree plot

The above scree plot shows the rate of change in eigenvalue according to the number of main components. This value goes through Principal Component 6 to PC7, and the slope is gentle enough to be almost parallel. In addition, the cumulative proportion in PC6 is 0.84, which is more than 80%. Therefore, in this task, the number

of main components (PCs) was determined to be 6. Next, check what each PC mea ns. The reason why variables have negative values is that the average of all variabl es is set to zero through data normalization.

- PC 1: Negative mean of variables excluding Q2a and Q2b
- PC 2: Average of Q90a, Q90b, Q90c, and Q90f
- PC 3: Negative mean of Q2a and Q2b
- PC 4: Q2a - Q2b
- PC 5: Negative number of Q90f
- PC 6: Average of Q87b and Q90b

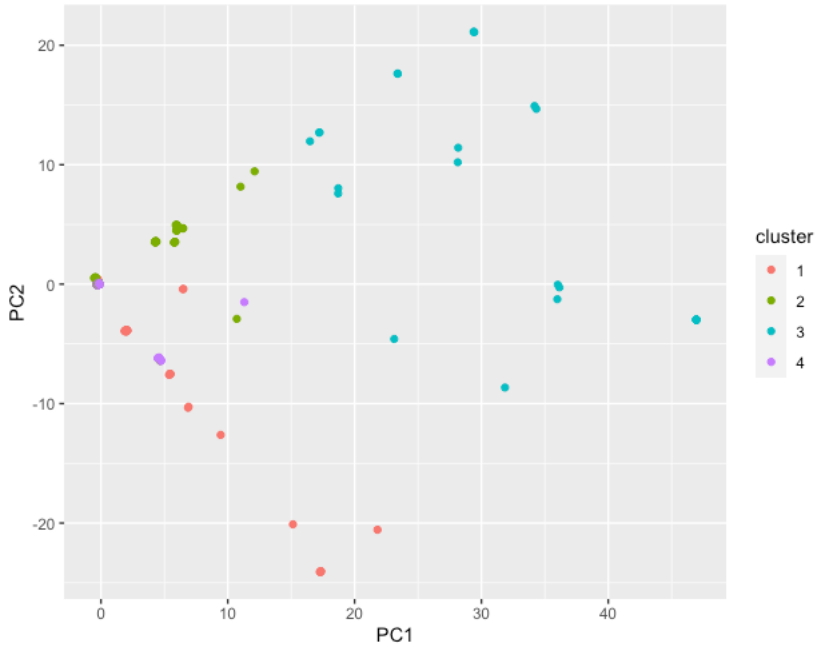


<Figure 2> Biplot

The above Biplot visualizes the distribution of data for two PCs (generally PC1 and PC2). Most observations are biased so that PC1 has a positive value, and PC2 has both positive and negative values and is widely distributed. Red texts means an ori ginal variable, and the variables are clustered because the correlation coefficients of the variables belonging to Q87 and Q90 are high.

After reducing the number of variables through principal component analysis (PCA), t he clustering process proceeds by grouping observations with similar attributes into s everal groups. First of all, clustering is a different concept from classification. Classifi cation has a predefined label as supervised learning, whereas clustering is unlabeled and aims to find the optimal group in the data. Among various clustering methods, this task will use K-Means Clustering, a representative algorithm. At first, K-Means C lustering sets the number of clusters K. Then, the cluster is reallocated to all observ ations based on the centroid of the generated cluster. Subsequently, the center of e ach changed cluster is recalculated. If the above process is repeated until the center

does not change, clusters with high similarity finally appear. The number of clusters to solve this problem was set to 4. The scatter plot below shows the quantity of o bservations and the characteristics by cluster.



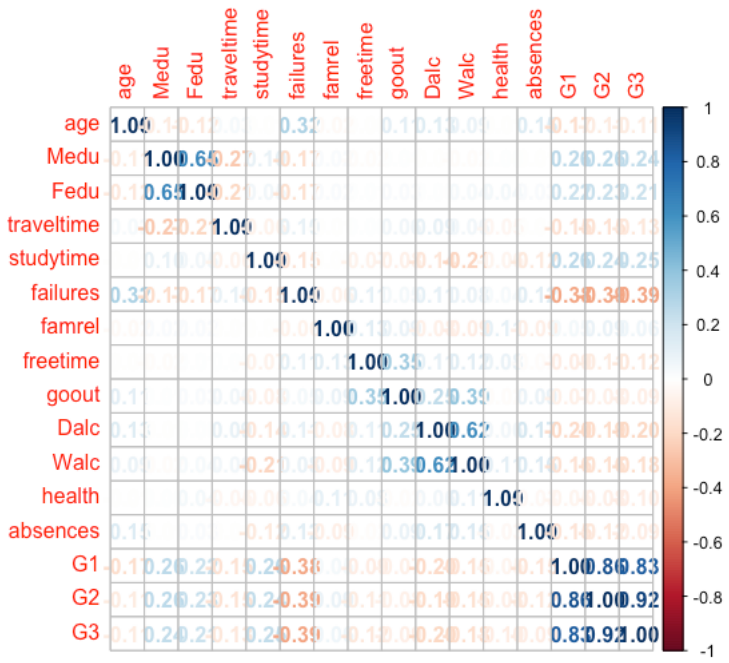
<Figure 3> Scatterplot (K-Means Clustering)

Looking at the graph, in clusters 1 and 4, both PC1 and PC2 have small values. On the other hand, in the case of cluster 3, both PC1 and PC2 have higher values than other clusters.

2.2 Part 2

Unlike Part 1, Part 2 provides the label final grade (y value) and uses a regression model belonging to supervised learning techniques. This model analyzes the weight of the effects of variables in the data on y. The goal of the regression model is to predict the y value with new variables by grasping the relationship between each va riable and the result y.

First, in the statistical summary of each variable, there are few variables that are bi ased to one side.



<Figure 4> Correlation Plot

Next, an analysis of the correlation between each variable is presented. The correlation interval between the defined first period grade (variable G1), second period grade (variable G2), and final grade (variable G3) is particularly large. Therefore, the approach of removing G1 and G2, which are direct causes, and predicting G3, is judged to be a better prediction. The first step for model construction is to divide the data into a train sample (70% of the total data) to understand the causal relationship and a test sample (30% of the total data) to predict the y-value.

This part finds the model that shows the best performance out of a total of three regression models.

First, linear regression is a model that predicts y by creating a line representing the relationship between y (dependent variable) and one or more x (independent variable). As a result of linear regression training, there are a total of 9 statistically significant variables, and the list is as follows. School, sex, Fedu, studytime, failures, school sup, higher, farmel, health. Next, the accuracy is confirmed by applying a regression line to the set test sample. The model's performance value, R squared, usually had a value of about 0.25.

The Random Forest model was used as the second regression model. It uses the base in the Ensemble model as the Decision Tree model. Random forests divide existing training data into several through Bagging and form their own decision tree. The variable used in this process is randomly selected and is called random subspace. When the y values from the decision tree are synthesized, the final y value is predicted. This part sets the number of trees to 500. As a result of measuring the final y value after random forest to the test sample, R squared steadily had a value of 0.35 or more.

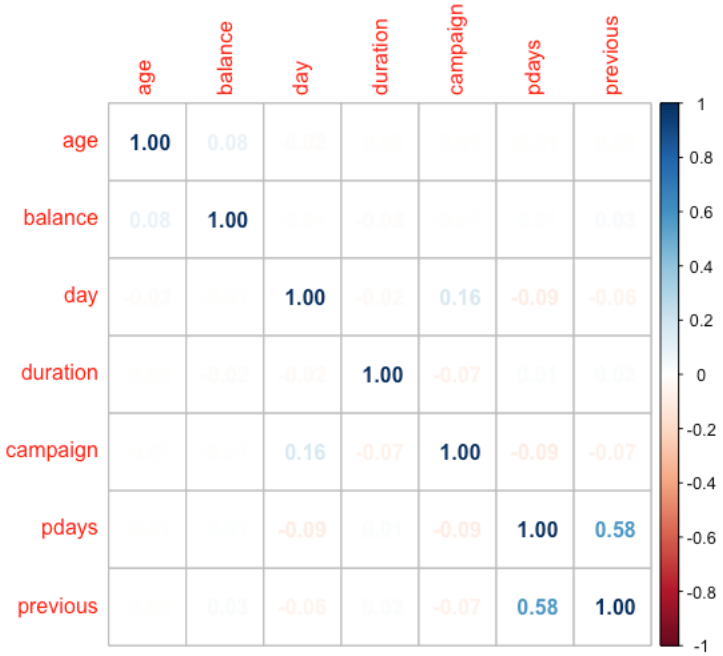
The last model is Support Vector Regression (SVR), which applies Support Vector M achines (SVMs) to regression. The SVM looks for a hyperplane that maximizes marg in, the distance between the regression line and the data. Through this process, it is possible to improve prediction performance by minimizing generalization errors. In th is case, a nonlinear structure may be created after expanding the variable space in a specific manner using kernals. As a result of the test, R squared usually had a v alue between linear regression and the R squared value of the random forest.

In conclusion, although the results may be different by randomly dividing samples, R squared was mostly higher in the order of linear regression, SVR, and random fores t. Accordingly, in part 2, it may be said that the random forest model shows the bes t performance.

2.3 Part 3

The model to be built in Part 3 is supervised learning with labels, as in Part 2. Ho wever, the difference is that it is a classification model, not a regression. The label y in Part 2 was the figure specified as the final grade. The goal of Part 3 is to pre dict whether label y will be classified as yes or no in the subscription of a term dep osit of a client.

In the y data, which means whether to subscribe a term deposits, there are 4000 n o and 521 yes. Class imbalance occurs more in no, which is due to the difference i n prediction accuracy between the two classes, so it must be balanced.



<Figure 5> Correlation Plot

In the correlation between variables before solving the class imbalance, there are no variables that affect y except for the variable 'previous'. The 'previous' variable mea

ns number of contacts performed before this campaign and for this client, and if the frequency of contact is high, the probability of regular deposit subscription is high.

To solve the class imbalance, this part uses oversampling. Oversampling will be appl ied in the train sample. Therefore, in the test sample process, a total of 200 data ar e sampled with ‘No’ 100 and ‘Yes’ 100, and a relatively insufficient amount of ‘Yes’ data is restored and extracted according to 3900 of the remaining ‘No’. In conclusio n, the same number is created in ‘Yes’ and ‘No’. As such, oversampling means cop ying insufficient classes of data to match the amount of data, whereas undersamplin g means reducing the amount of data on many classes to fit the smaller. In conclus ion, ‘No’ 3900 and ‘Yes’ 3900 are configured in the train sample through oversampli ng.

In this part, three classification models are used as part 2.

The first model is Logistic Regression. This model has the same analysis process a s the Linear Regression in Part 2, but the output is different. If the prediction of y (dependent variable) in linear regression is continuous, the y prediction in logical regr ession is discrete. This is mainly applied when solving binary classification problems such as Part 3. Accuracy, Sensitivity, and Specificity values are used for model eval uation. Accuracy is the probability that the predicted result of the whole is the same as the actual result. Sensitivity is the probability that the prediction is correct during a test with a y value of ‘Yes’, and Specificity is the probability that the prediction du ring a test with a y value of ‘No’ is correct. The results of the Logistic Regression model show compliance performance with 0.8, Sensitivity with 0.85, and Specificity w ith 0.75.

The second model is a classification model using the random forest mentioned in P art 2. Similarly, the number of trees was set to 500, and the process is the same a s described above, but only the type of y as a result is different. The model's evalu ation is 0.67 for Accuracy, 0.95 for Sensitivity, and 0.4 for Specificity. Compared to l ogistic regression, all values excluding sensitivity are shown to be low, which is judg ed to be somewhat insufficient performance. It is analyzed that sensitivity was meas ured high because the prediction result was relatively often ‘Yes’.

The last model is also a Support Vector Classifier (SVC) using SVM, like the last m odel in Part 2. This makes the line dividing the ‘Yes’ data group and the ‘No’ data group non-linear, not straight. Through this, it predicts which group the new y-value will belong to. Evaluating the model's performance, Accuracy is 0.82, Sensitivity is 0. 82, and Specificity is 0.8, which is superior to the two previous models. The SVC m odel is observed as the most optimal model that conforms to both ‘No’ and ‘Yes’.

3. Conclusion

Part 1 created a model that reduces variables in data to analyze the correlation of newly created variables and clusters similar data. In addition, Parts 2 and 3 created a model that predicts the result value using existing data variables. The type of learning to be used is determined according to the desired result value. The predictor must choose which model to use even after the type of learning is determined. Like the performance results of the previously obtained models, even the same data shows different performance depending on which algorithm model is used. In addition, it was confirmed that performance was improved by generating nonlinearity through SVM in linear regression.

As parts deal with data in various fields, many fields currently use and require machine learning. The accuracy of prediction is ultimately what we want to obtain through machine learning. Higher accuracy requires the process of creating many models and evaluating performance. The more various models are used, the wider the performance evaluation range of the models is. This means that the probability of obtaining high accuracy increases. Therefore, in the use of machine learning, which models to use for data and how many models can be used are important points related to accuracy.