Machine learning CA project

Student name: Bui Kim Dung
ID: E0146998

1. **Build machine learning models for classification using R**

1.1. Prepare data

The sample data has 13 chemical attributes and 3 different types of wine, with medium size of data (m = 178). With 3 types of classification, I would choose LDA and QDA as the classification model.

sum(is.na(wines)) method show that there is no NA (not available) value in any column.

Since LDA and QDA are more reliable if the all the variables are normality distributed, I used Shapiro test to run on all variables to check if they are at 95% normality level. The test showed that only Ash_Alcalinity is not normal, but all other variables are fine, so LDA and QDA approximation should be reliable.

For the purpose of testing, I split the 178 rows sample into a training set and test set with ration 80:20. Test set has type A, B, C as 46, 57, 39, respectively, total 142 rows. Train set has type A, B, C with 13, 14, 9 respectively, total 36 rows.

1.2. Build ML model using LDA

There are 3 types of wine with 13 variables, so we can find at most 2 useful discriminant functions to separate the wines. The percentage separation achieved by each discriminant function return on the train data as follow:

```
Proportion of trace:
   LD1    LD2
0.7058 0.2942
```

After that, run the lda.fit on the test set to predict classifications and confusion matrix table show the prediction accuracy as 100%.

```
> table(lda.pred$class,test$Type)
     A  B  C
  A 13  0  0
  B  0 14  0
  C  0  0  9
```

1.3. Build ML model using QDA

Build similar ML model with QDA on train data, and then run prediction on test data also showed the similar results as QDA with accuracy 100%.

2. Compare model performance using Leave-One-Out Cross-Validation

Run LOOCV validation on LDA and QDA on the whole data set showed results that LDA and QDA can achieve accuracy of 98.87% and 99.4%, respectively. Although when running with train and test data, both model gave 100% correct.

LOOCV also showed that QDA can achieve better results. Since LDA assumes that the probability density functions of the predictors are normal, and covariance matrices between classes are the same. QDA typically produces larger variances than LDA, but it is more flexible and has lower bias. Since the test and train data set are quite small, both models achieve 100% so it is not possible to observe if QDA is better than LDA with normal training and predicting process.

3. Further discussion

3.1. Use Logistic Regression for the same multi classification problem

Since Logistic Regression is typically for binary classification, to solve this problem using Logistic regression, I have used multinom() to fit the model. Run LOOCV on logistic regression gave accuracy of 94%.

3.2. Should all 13 variables be used

To build the model, I have used all 13 variables. But if run the best subset selection on sample data, then try to remove some columns according to best subset selection.

If removes Malic_Acid, Ash, Magnesium, NonFlavanoid_Phenols, Proanthocyanins, the result of LOOCV on LDA returns lower than with all 13 variables (95.5%)

If removes Magnesium, NonFlavanoid_Phenols, Proanthocyanins, LOOCV on LDA returns similar result (98.87%).