Universidad de Navarra

TESTING PREDICTION ALGORITHMS' ACCURACIES

MONTERO MUÑOZ, GONZALO
RICCI ZARAGOZA, DANIEL A.
VILLANUEVA JOVE, JAVIER E.

MACHINE LEARNING

INTRODUCTION

For the purpose of this project we will be analysing a data set containing statics of 126 students from Mercyhurst University in the United States. Using this information, we will implement eight different algorithms to test their prediction's accuracies.

THE DATA SET

The data set was taken from kaggle and is called “Food Choices” https://www.kaggle.com/borapajo/food-choices/data, This dataset includes information on food choices, nutrition, preferences, childhood favourites, and other information from college students but we decide that for our project the most important ones are the GPA, Daily calories consumption, income and weight. There are 126 responses from students of the Mercyhurst University in the United States.

From our variables, the college GPA is measure in the American way with a range between 0 – 4, in the data as you can see in the chart below the GPA have a range between 2,2 – 4 with an average of 3,42. The daily calories that are express in thousands have a range between 2 - 4 and an average of 3,02 calories per day. The income is also express in thousands and have a range between 1 – 6 with an average of 4,52$ and the final variable that is weight have a range between 100 – 265 pounds with an average of 159,05 pounds, the conversion rate from pounds to kilograms is 0,453592.

Range	GPA	Daily Calories	Income	Weight
Max	4	4	6	265
Average	3.42	3.02	4.52	159.05
Min	2.2	2	1	100

OBJECTIVE

Our purpose is to identify the gender of a student based on the 4 variables previously explained, we believe that there are differences in student behaviours that will let us know, depending on the interaction between the variables, if the student is male or female. To predict the results, we used eight different algorithms (Linear discriminant analysis, Quadratic discriminant analysis, Logistic regression, Rpart, Support vector machines, Naïve Bayes, Cross Validation 10 and Kmeans).

After we analyse the data with each algorithm we will determine the accuracy of the process and explain the differences between the eight of them, at the end, we will compare and determine which ones are the best algorithms to predict an analyse these type of data.

ANALYSIS

# CLEANING THE DATA SET

After uploading the dataset, the first step needed was to clean the data by creating a copy of the dataset that would not include any empty spaces:

```
➢ df <- na.omit(FoodChoices)
```

# DIVIDING THE DATA SET IN TRAIN AND TEST SETS

We installed the library needed and proceeded to split de data with a 0.6 ratio, having now a train set with a random 60% of the data and a test set with 40%. We also transformed the variable Gender having in mind that some algorithms we will use later will need it. (Cross 10 Validation & Kmeans will not use this split)

```
➢ library(caTools)
➢ df$Gender=factor(df$Gender)
➢ split <- sample.split(df$Gender, SplitRatio = 0.6)
➢ train <- subset(df, split==TRUE)
➢ test <- subset(df, split==FALSE)
```

# TO MEASURE ACCURACY

Before we began applying any of the predictive algorithms, we created a function that will allow us to evaluate and display each accuracy:

```
➢ accuracy <- function(x){sum(diag(x))/sum(x)}
```

Then, we began working with the selected algorithms. After successfully implementing all of them, we'll proceed to compare their results.

ANALYSIS

# ALGORITHM N.1: LINEAR DISCRIMINANT ANALYSIS

"***Linear Discriminant Analysis*** (LDA) was proposed by R. Fischer in 1936. It consists in finding the projection hyperplane that minimizes the interclass variance and maximizes the distance between the projected means of the classes.

This hyperplane can be used for classification, dimensionality reduction and for interpretation of the importance of the given features." *1.*

```
library(MASS)
m.lda <- lda(Gender~., data = train)
p.lda <- predict(m.lda,test)
(t.lda <- ftable(Predicted=p.lda$class, Correct=test$Gender))
accuracy(t.lda)
```

	CORRECT	FEMALE	MALE
PREDICTED			
FEMALE		20	6
MALE		5	9

Accuracy: 0.725

ANALYSIS

# ALGORITHM N.2: QUADRATIC DISCRIMINANT ANALYSIS

"***Quadratic discriminant analysis*** is a modification of LDA that does not assume equal covariance matrices amongst the groups.

In quadratic discriminant analysis, the group's respective covariance matrix is employed in predicting the group membership of an observation, rather than the pooled covariance matrix in linear discriminant analysis." *2.*

```
➢ m.qda <- qda(Gender~., data = train)
➢ p.qda <- predict(m.qda,test)
➢ (t.qda <- ftable(Predicted=p.qda$class, Correct=test$Gender))
➢ accuracy(t.qda)
```

	CORRECT	FEMALE	MALE
PREDICTED			
FEMALE		20	7
MALE		5	8

Accuracy: 0.7

ANALYSIS

# ALGORITHM N.3: MULTINOMIAL LOGISTIC REGRESSION

"***Multinomial Logistic Regression*** is the linear regression analysis to conduct when the dependent variable is nominal with more than two levels. Thus it is an extension of logistic regression, which analyzes dichotomous (binary) dependents.

Like all linear regressions, the multinomial regression is a predictive analysis. Multinomial regression is used to describe data and to explain the relationship between one dependent nominal variable and one or more continuous-level(interval or ratio scale) independent variables." *3.*

```
➢ library(nnet)
➢ m.mlr <- multinom(Gender~., data = train)
➢ (t.mlr<- ftable(Predicted=predict(m.mlr,test), Correct=test$Gender))
➢ accuracy(t.mlr)
```

	CORRECT	FEMALE	MALE
PREDICTED			
FEMALE		20	5
MALE		5	10

Accuracy: 0.75

ANALYSIS

# ALGORITHM N.4: TREES, RPART ALGORITHM

“Trees (also called decision trees, recursive partitioning) are a simple yet powerful tool in predictive statistics. ***The Rpart*** programs build classification or regression models of a very general structure using a two stage procedure. The idea is to split the covariable space into many partitions and to fit a constant model of the response variable in each partition. In case of regression, the mean of the response variable in one node would be assigned to this node.

The structure is similar to a real tree (from the bottom up): there is a root, where the first split happens. After each split, two new nodes are created (assuming we only make binary splits). Each node only contains a subset of the observations.
The partitions of the data, which are not split any more, are called terminal nodes or leafs. This simple mechanism makes the interpretation of the model pretty easy.” *4.*

```
library(rpart)
m.rp <- rpart(Gender~., data = train)
(t.rp <- ftable(Predicted=predict(m.rp, test, type = "class"),
Correct=test$Gender))
accuracy(t.rp)
```

	CORRECT	FEMALE	MALE
PREDICTED			
FEMALE		20	7
MALE		5	8

Accuracy: 0.7

ANALYSIS

# ALGORITHM N.5: SUPPORT VECTOR MACHINE

“***Support Vector Machine***” (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems.

In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot).

Support Vectors are simply the co-ordinates of individual observation.” *5.*

```
library(kernlab)
m.svm <- ksvm(Gender~., data=train, scale=FALSE)
(t.svm <- ftable(Predicted=predict(m.svm,test), Correct=test$Gender))
accuracy(t.svm)
```

	CORRECT	FEMALE	MALE
PREDICTED			
FEMALE		20	10
MALE		5	5

Accuracy: 0.625

ANALYSIS

# ALGORITHM N.6: NAIVE BAYES

“It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a ***Naive Bayes*** classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.” *6.*

```
➢ library(e1071)
➢ m.nb <- naiveBayes(Gender ~ ., data = train, subset)
➢ p.nb <- predict(m.nb, test, threshold = 0.05)
➢ (t.nb <- ftable(Predicted=p.nb, Correct=test$Gender))
➢ accuracy(t.nb)
```

	CORRECT	FEMALE	MALE
PREDICTED			
FEMALE		20	6
MALE		5	9

Accuracy: 0.725

ANALYSIS

# ALGORITHM N.7: CROSS VALIDATION 10

“***Cross-validation*** is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

For classification problems, one typically uses stratified k-fold cross-validation, in which the folds are selected so that each fold contains roughly the same proportions of class labels. In repeated cross-validation, the cross-validation procedure is repeated n times, yielding n random partitions of the original sample. The n results are again averaged (or otherwise combined) to produce a single estimation. ” *7.*

```
➢ library(caret)
➢ library(klaR)
➢ x = df[,-1]
➢ y = df$Gender
➢ model = train(x,y,'nb',trControl=trainControl(method='cv',number=10))
➢ predict(model$finalModel,x)
➢ table(predict(model$finalModel,x)$class,y)
➢ t.cv = table(predict(model$finalModel,x)$class,y)
➢ accuracy(t.cv)
```

	CORRECT	FEMALE	MALE
PREDICTED			
FEMALE		52	17
MALE		10	21

Accuracy: 0.73

ANALYSIS

# ALGORITHM N.8: KMEANS ALGORITHM

*"**K-means*** clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided. Data points are clustered based on feature similarity." *8.*

```
➢ Results <- kmeans(df[,2:5], 2, nstart=20)
➢ (t.km <- ftable(Predicted=Results$cluster, Correct=df$Gender))
➢ 1-accuracy(t.km)
```

The accuracy is displayed as "1-accuracy" because we observed that the algorithm successfully managed to identify men over women but simply organized them into the wrong order of clusters.

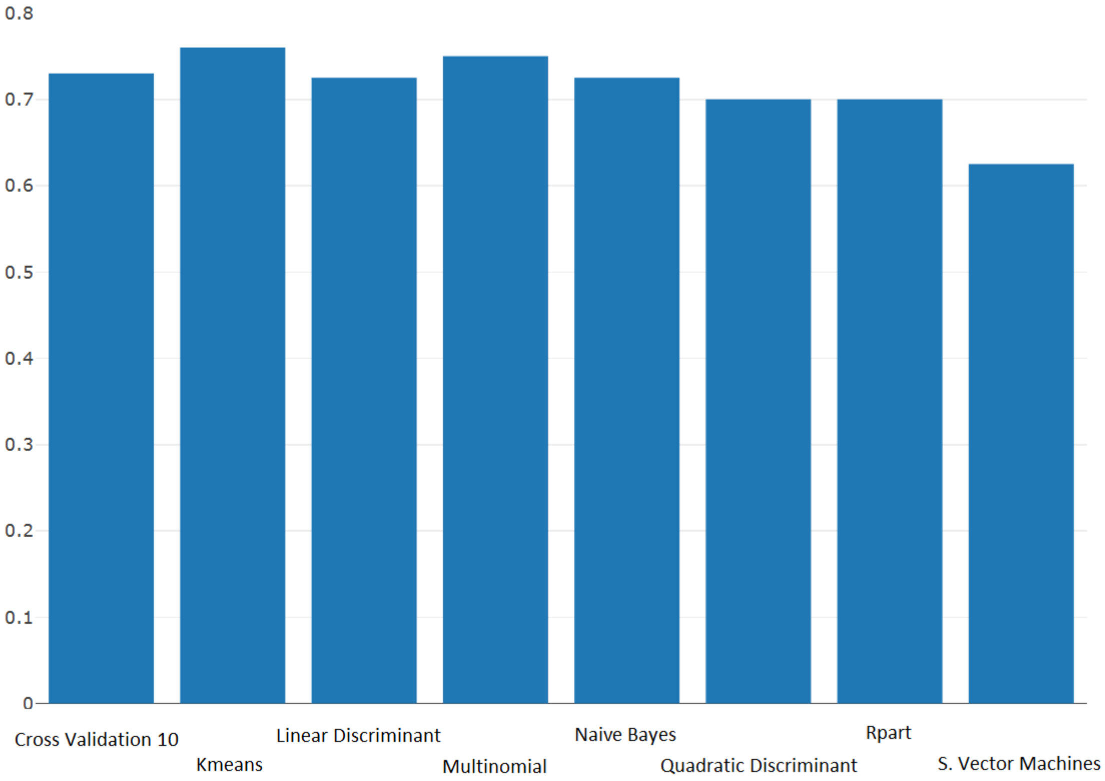
	CORRECT	FEMALE	MALE
PREDICTED			
1		15	29
2		47	9

Accuracy: 0.76

CONCLUSION

In conclusion, we have successfully implemented eight algorithms into our dataset capable of identifying student gender based on their behavioural characteristics. However, even though the algorithms were able to successfully analyse the data, the margin of error lies between 25% - 40%. So even though our algorithms are able to analyse the majority of the data, they fall short from giving us a perfect representation of reality.

We observed that the most accurate were the two algorithms that didn’t used the split, being Cross Validation 10 and Kmeans. Another interesting observation was that all the other algorithms that used the split correctly identified 20 women and incorrectly identified 5; having the accuracy percentage differenced only by its ability in identifying the males.



REFERENCES

1. *Xanthopoulos P., Pardalos P.M., Trafalis T.B. (2013) Linear Discriminant Analysis. In: Robust Data Mining. SpringerBriefs in Optimization. Springer, New York, NY*

2. *Rencher, A. (n.d.). Methods of Multivariate Analysis (2nd ed.). Brigham Young University: John Wiley & Sons, Inc.*

3. *Statistics Solutions. (2017). Retrieved from http://www.statisticssolutions.com/mlr/*

4. *Molnar, C. (2012, November 13). R - Bloggers. Retrieved from https://www.r-bloggers.com/trees-with-the-rpart-package/*

5. *Ray, S. (2017, September 17). Analytics Vidhya . Retrieved from https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/*

6. *Ray, S. (2017, September 11). Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/*

7. *Open Machine Learning . (n.d.). Retrieved from https://www.openml.org/a/estimation-procedures/1*

8. *Trevino, A. (2016, December 6). Data Science. Retrieved from https://www.datascience.com/blog/k-means-clustering*