Discriminant Analysis on Mobile Phone Data

Yehya Abdelmohsen

4/4/2022

Contents

Introduction

Mobile phones have become widely available in the past couple years. This availability has allowed a surge in the number of users worldwide. To elaborate on this point, in 2022 there were 6.6 billion registered smartphones (2022a). To put this number into perspective, the world population in 2022 is around 7.9 billion (2022b). That means that around 83% of the world population owns a smartphone. Of course, we are not accounting for people with multiple smartphones and other factors. To further exaggerate the growth in the number of smartphones, only 6 years ago in 2016 the number of active smartphones were around 3.7 billion. These facts are all driven by the fact that mobile phones are becoming an essential tool for the majority of people, whether they use it as a means of communication, entertainment, work, or other uses. In this project, a data set containing multiple variables including the variable price range is presented. We will be using discriminant analysis to classify observations into their respective classes.

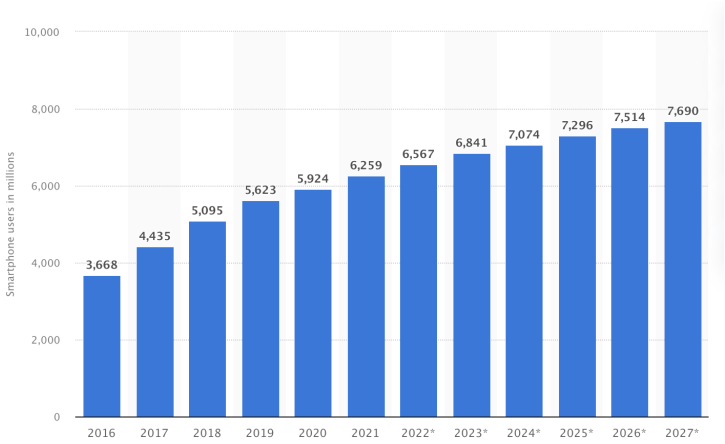

Figure 1: Number of Smartphones, Statista (2022)

About the Data

The data presented contains information related to mobile phones. It contains 21 variables, 14 of these are quantitative, 6 are binary, and 1 is categorical. The data is already split into training and testing. The training data contains 2000 observations. The testing data contains 1000 observations, however, it does not contain the class variable. Therefore, we will not be able to use it to calculate external classification error, and we will have to split our training data into training and testing again.

The data is obtained from the kaggle website, linked below.

Source: https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification

Variables

- Battery power - Total energy a battery can store in one time in mAh. This is a quantitative variable.
- Blue - Whether the phone has bluetooth or not. This is a qualitative binary variable i.e. 0 or 1.
- Clock speed - Speed at which the processor executes instructions in GHz. This is a quantitative variable.

- Dual sim - Whether the phone has dual sim support or not. This is a qualitative binary variable.
- Fc - Front camera megapixels. One megapixel is a million pixels. This is a quantitative variable.
- Four g - Whether the phone has 4g or not. This is a qualitative binary variable.
- Int memory - Internal memory in GB. This is a quantitative variable.
- M dep - Mobile depth in cm. This is a quantitative variable.
- Mobile wt - Weight of mobile phone in grams. This is a quantitative variable.
- N cores - Number of cores of the processor. This is a quantitative variable.
- Pc - Primary camera megapixels. This is a quantitative variable.
- Px height - Pixel resolution height. This is a quantitative variable.
- Px width - Pixel resolution width. This is a quantitative variable.
- Ram - Random access memory in MB. This is a quantitative variable.
- Sc h - Screen height of mobile in cm. This is a quantitative variable.
- Sc w - Screen width of mobile in cm. This is a quantitative variable.
- Talk time - Longest time that a single battery charge will last when you are using the calling someone on the phone. This is a quantitative variable.
- Three g - Has 3g or not. This is a qualitative binary variable.
- Touch screen - Has a touch screen or not. This is a qualitative binary variable.
- Wifi - Has wifi or not. This is a qualitative binary variable.
- Price range - Low cost (0), medium cost (1), high cost (2), and very high cost (3). This is a qualitative categorical variable.

Some Remarks Regarding the Data

Below is an initial look at the data. The data does not contain any missing values. We will be using all 21 variables and the class variable price range to perform discriminant analysis.

battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	pc	px_height
842	0	2.2	0	1	0	7	0.6	188	2	2	20
1021	1	0.5	1	0	1	53	0.7	136	3	6	905

px_height	px_width	ram	sc_h	sc_w	talk_time	three_g	touch_screen	wifi	price_range
20	756	2549	9	7	19	0	0	1	1
905	1988	2631	17	3	7	1	1	0	2

Problem Statement

The aim of this analysis is to compare the results of Fishers Linear Discriminant Analysis (FLDA) and Multinomial Random Variable (MRV). In addition, we would like to classify new observations. We will try both FLDA and MRV and decide which is most suitable for classifying the data at hand.

Assumptions

Normality

To satisfy the normality assumption in Fisher Linear Discriminant Analysis, we must transform the variables so that they follow a normal distribution. In Figure 2, 3, and 4 we can see that none of the variables follow a normal distribution.

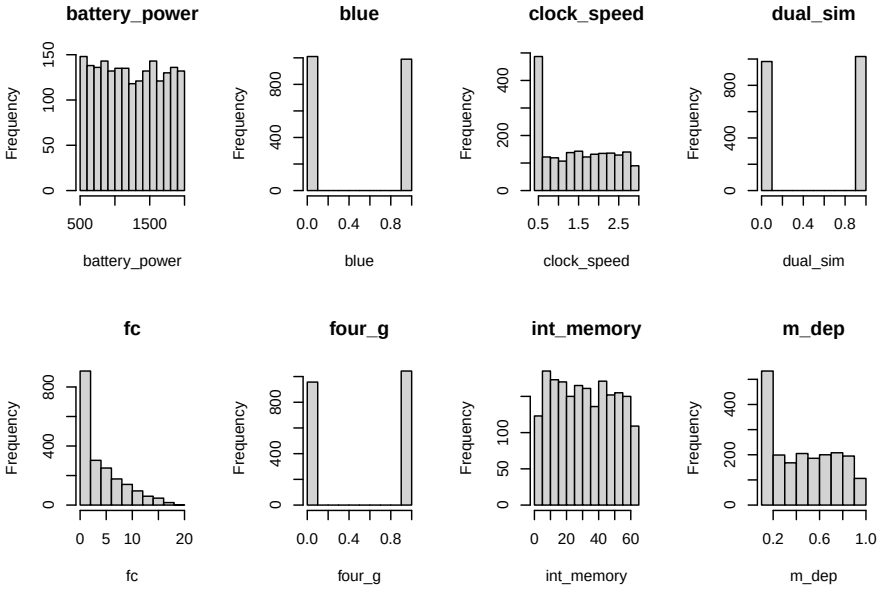

Figure 2: Histograms of All Variables

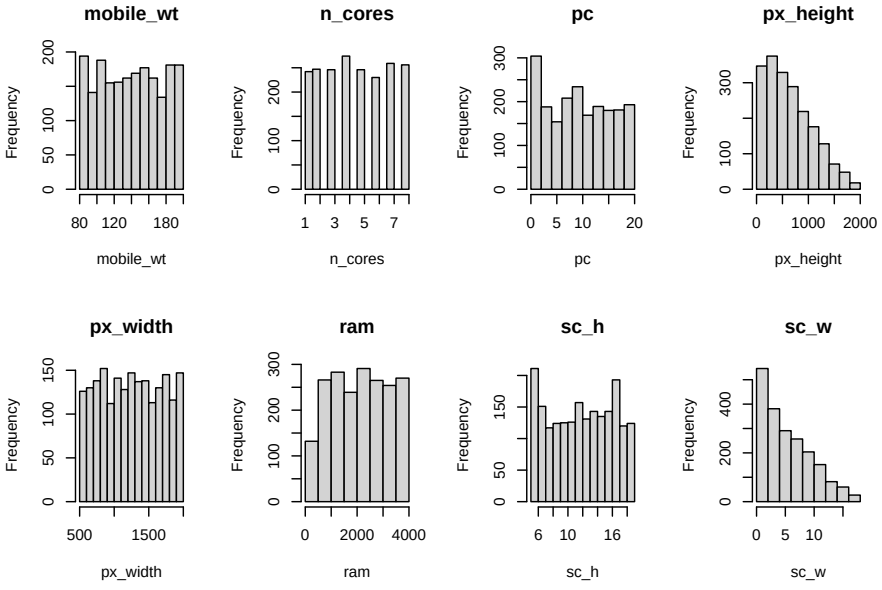

Figure 3: Histograms of All Variables

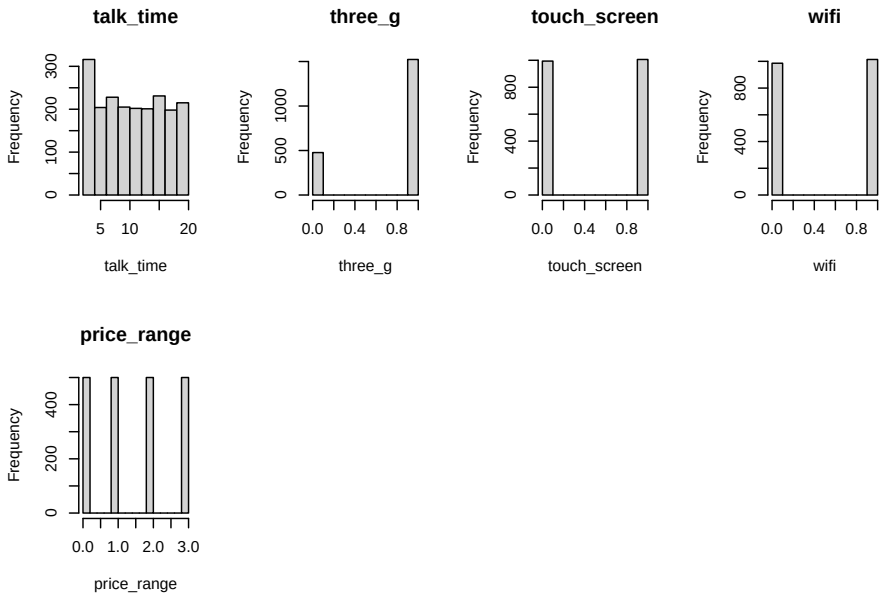

Figure 4: Histograms of All Variables

Transformations

Adding ten to the variables with 0 values in order to allow the log transformation.

```
add_one = c('blue','dual_sim','fc','four_g','pc','px_height','sc_w','three_g','touch_screen','wifi','pr
for(i in add_one){
  train[,i] = train[,i] + 10
}
```

We then perform the necessary transformations. They are shown below. We can also see in Figure 5 the distribution of these variables after transformation.

$$
\begin{aligned}
FrontCamera &= log(FrontCamera) \\
MobileDepth &= log(MobileDepth) \\
NumberofCores &= log(NumberofCores \\
ScreenWidth &= log(ScreenWidth) \\
ScreenHeight &= log(ScreenHeight) \\
ClockSpeed &= log(ClockSpeed) \\
PixelHeight &= \sqrt{PixelHeight}
\end{aligned}
$$

The transformations in Figure 5 are not perfectly normally distributed, however, we are trying to approach a normal distribution as much as possible. Therefore, we perform the transformations regardless of whether we reach a perfect normal distribution or not. To conclude, the assumption of normality is not valid for all variables.

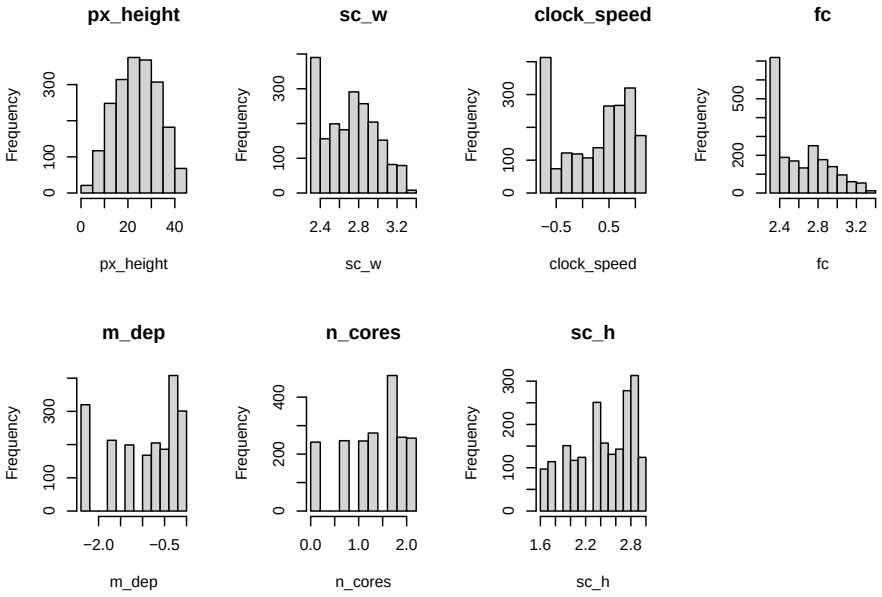

Figure 5: Transformation of Variables

Equal Variance

The assumption of equal variance-covariance matrices for all classes will be relaxed, therefore, we assume the following.

$$\Sigma_0 = \Sigma_1 = \Sigma_2 = \Sigma_3$$

Where the subscript represents the price range: low cost (0), medium cost (1), high cost (2), and very high cost (3).

Classification

Fisher Linear Discriminant Analysis

Internal Classification

We perform linear discriminant analysis. As we can see in the confusion matrix Table 3, there are 98 internally misclassified observations. Therefore, the internal misclassification rate is 4.9%.

```
## Fisher Linear Discriminant:
```

Table 3: Confusion Matrix (Internal)

	10	11	12	13
10	485	15	0	0
11	8	476	16	0
12	0	31	465	4

	10	11	12	13
13	0	0	24	476

```
## Internal Classification Error Rate = 4.9 %
```

Training and Testing

We split the data into training and testing in order to compute external misclassification. The training set contains 1500 observations while the testing set contains 500 observations.

```
# Shuffling data
set.seed(1652001)
s = sample(1:2000,1500,replace=F)
new_train = train[s,]
new_test = train[-s,]
```

We use the model to compute the confusion matrix Table 4 for the testing set. We find that the external misclassification error is 5.4%. Out of 500 observations in the testing set 27 were misclassified. The external misclassification is higher than internal misclassification, which is logical. However, the internal misclassification is computed on the whole data $n = 2000$ while the external is computed on $n = 500$. The fact that the two error rates are not too far off from each other is a good sign. In addition, an error rate less than 10% is sufficient for classification in my opinion.

Table 4: Confusion Matrix (External)

	10	11	12	13
10	124	3	0	0
11	5	121	12	0
12	0	2	114	5
13	0	0	0	114

```
## External Misclassification Error: 5.4 %
```

Leave One Out

We also attempt the leave one out method for cross validation. In this case, we will be using the whole data. The number of runs is $n = 2000$.

```
loo=function(x,class){
  n=length(class)
  rslt={}
  for(i in 1:n){
     a = lda(x[-i,], class[-i])
     b = predict(a,x[i,])
     rslt[i]=b$class #[i]==class[i]
  }
return(rslt)
}
```

We can see in Table 5 that the off diagonal elements are 112 observations. Therefore, we have 112 misclassifications out of 2000 observations.

Table 5: Confusion Matrix (External - Leave One Out)

10	11	12	13
483	10	0	0
17	471	35	0
0	19	461	27
0	0	4	473

The misclassification error is 5.6%, which is higher than when we split the data into training and testing sets. Therefore, we conclude that splitting the data into training and testing is superior to the leave one out method in this case and results in a lower misclassification rate 5.4% compared to 5.6%.

```
## Leave One Out External Misclassification Error: 5.6 %
```

Multinomial Random Variables

Now we will be using the multinomial distribution to classify observations.

```
library(nnet)
mn = multinom(price_range ~ ., data = train)
```

```
## # weights:  88 (63 variable)
## initial  value 2772.588722
## iter  10 value 2231.079111
## iter  20 value 1941.385751
## iter  30 value 1874.780749
## iter  40 value 1296.816032
## iter  50 value 912.761933
## iter  60 value 318.592381
## iter  70 value 134.119499
## iter  80 value 127.517858
## iter  90 value 117.246714
## iter 100 value 97.931969
## final  value 97.931969
## stopped after 100 iterations
```

```
results = predict(mn)
```

Internal Classification

We compute the confusion matrix. As can be seen from Table 6, there are 40 misclassified observations out of 2000 observations.

Table 6: Confusion Matrix (Internal)

	10	11	12	13
10	496	4	0	0
11	5	484	11	0
12	0	4	487	9
13	0	0	7	493

The internal misclassification error is 2%, which is lower than the one obtained from FLDA 4.9%.

```
## Internal Misclassification Error: 2 %
```

Training and Testing

Now splitting our data into training and testing again. We find the classification error is lower than FLDA when we split the data into training and testing.

The off diagonal elements sum is 25 observations. Out of 500 observations 25 were misclassified.

Table 7: Confusion Matrix (External)

	10	11	12	13
10	127	2	0	0
11	2	123	10	0
12	0	1	111	5
13	0	0	5	114

The external misclassification error is 5%, which is lower than FLDA's 5.4% in the same case. Also, the difference between the internal misclassification error 2% and external misclassification error 5% is large compared to the difference observed in FLDA.

```
## External Misclassification Error: 5 %
```

Leave One Out

Finally, we will be applying the leave one out method for multinomial random variables.

Using the leave one out method for the multinomial distribution we get 74 misclassifications out of 2000 observations.

Table 8: Confusion Matrix (External - Leave One Out)

10	11	12	13
487	12	0	0
13	473	11	0
0	15	475	9
0	0	14	491

The misclassification error is 3.7%, the lowest out of all the external classifications measured. In addition, it is lower than the external misclassification error for training and testing. This is the opposite of what we observed in FLDA.

```
## External Misclassification Error: 3.7 %
```

Conclusion

In the table below we can see that the Multinomial Random Variable (MRV) out performs FLDA. MRV has a lower error rate for all three tests. The last row with the variable n represents the number of observations in the training set. We conclude that given new data regarding mobile phone price ranges. We will be using MRV to classify observations into the relevant price range. It's important to note that MRV outperforms FLDA in this case and that in some other cases FLDA may be superior. Also, we did not satisfy the assumptions of FLDA. Finally, we should note that the difference between the external classification when splitting the data into training and testing is not that much. Splitting the data into training and testing is the most practical use case. For FLDA the misclassification rate is 5.4% and MRV it is 5%, so the difference is not too significant.

Table 9: Comparing Methods

	Internal Misclassification Error	Training and Testing (External Misclassification Error)	Leave One Out (External Misclassification Error)
FLDA	4.9%	5.4%	5.6%
MRV	2%	5%	3.7%
n	2000	1500	1999

References

2022b. https://worldpopulationreview.com.
2022a. https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/.