Linear Discriminant Analysis on Pima Diabetes Data

Joshua Jose

2023-03-23

Aim

To develop a model that helps us identify if a particular person has diabetes or not(diabetes=1,no diabetes=0)based on the medical readings collected(glucose levels, Insulin levels,Age. . .).

Data Description

The Diabetes data set contains the variables

- Pregnancies - No of times the person has been pregnant.
- Glucose - Plasma glucose concentration at 2 hours in an oral glucose tolerance test.
- BloodPressure - Diastolic blood pressure (mm Hg).
- SkinThickness - Triceps skin fold thickness (mm).
- Insulin - 2-Hour serum insulin (mu U/ml).
- BMI - Body mass index (weight in kg/(height in metres squared)).
- DiabetesPedigreeFunction - Diabetes pedigree function.
- Age - Age (years)
- Outcome - test whether the patient shows signs of diabetes (coded 0 if negative, 1 if positive)

The data set contains a lot of observations with value 0 in the Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction and Age columns. These all seem to be observations where the reading are not noted down. In that case, we can either omit these entire rows or we can impute using the 'missForest' package , in order to get a good model.

Theory

Linear discriminant analysis(LDA) is used as a tool for classification. Decision theory for classification tells us that we need to know the class posteriors $Pr(G \mid X)$ for optimal classification. Suppose $f_k(x)$ is the class-conditional density of X in class $G = k$, and let π_k be the prior probability of class k, with $\sum_{k=1}^{K} \pi_k = 1$. A simple application of Bayes theorem gives us

$$Pr(G = k \mid X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

In this case we have more than one predictors(Pregnancies,Glucose, BloodPressure, SkinThickness. . .) we will assume that $X = (X_1, X_2, ..., X_p)$ is drawn from a multivariate Gaussian (or multivariate normal) distribution, with a class-specific multivariate mean vector and a common covariance matrix. This is important as Linear discriminant analysis (LDA) works on the basis of a few assumptions

- Sample measurements are independent from each other.
- We model each class density as multivariate Gaussian,another way of saying it is that the predictors are multivariate normal conditioned on the classes.

$$X \mid Y = k \sim N_p(\mu_k, \Sigma)$$

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}$$

- the classes have a common covariance matrix $\Sigma_k = \Sigma \; \forall k$.

In comparing two classes k and l, it is sufficient to look at the log-ratio, and we see that

$$log\frac{Pr(G=k \mid X=x)}{Pr(G=l \mid X=x)} = log\frac{f_k(x)}{f_l(x)} + log\frac{\pi_k}{\pi_l}$$

where the decision boundary(for class k and class l) is at 0. So, Decision Boundary:

$$log\frac{f_k(x)}{f_l(x)} + log\frac{\pi_k}{\pi_l} = 0$$

$$=> log\frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k^T\Sigma^{-1}\mu_k - \mu_l^T\Sigma^{-1}\mu_l) + (\mu_k - \mu_l)^T\Sigma^{-1}x = 0$$

$$=> (\mu_k - \mu_l)^T\Sigma^{-1}x = \frac{1}{2}(\mu_k^T\Sigma^{-1}\mu_k - \mu_l^T\Sigma^{-1}\mu_l) - log\frac{\pi_k}{\pi_l}$$

In practice we do not know the parameters of the Gaussian distributions, and will need to estimate them using our training data:

- $\hat{\pi}_k = N_k/N$ where N_k is the number of class-k observations;
- $\hat{\mu}_k = \sum_{g_i=k} x_i/N_k$
- $\hat{\Sigma} = \sum_{k=1}^{K}\sum_{g_i=k}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T/(N-K)$

With two classes there is a simple correspondence between linear discriminant analysis and classification by linear least squares, The LDA rule classifies to class 2 if

$$x^T\hat{\Sigma}^{-1}(\hat{\mu}_l - \hat{\mu}_k) > \frac{1}{2}\hat{\mu}_l^T\hat{\Sigma}^{-1}\hat{\mu}_l - \frac{1}{2}\hat{\mu}_k^T\hat{\Sigma}^{-1}\hat{\mu}_l + log\frac{N_k}{N} - log\frac{N_l}{N}$$

and class 1 otherwise.

Data Imputation

```
library("MASS")
library("missForest")
library("twinning")
library("ROCR")
library("mice")
```

```
##
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

Loading the data set and checking out how it looks

```
diabetes = read.csv("diabetes.csv")
head(diabetes)
```

```
##   Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1           6     148            72            35       0 33.6
## 2           1      85            66            29       0 26.6
## 3           8     183            64             0       0 23.3
## 4           1      89            66            23      94 28.1
## 5           0     137            40            35     168 43.1
## 6           5     116            74             0       0 25.6
##   DiabetesPedigreeFunction Age Outcome
## 1                    0.627  50       1
## 2                    0.351  31       0
## 3                    0.672  32       1
## 4                    0.167  21       0
## 5                    2.288  33       1
## 6                    0.201  30       0
```

```
dim(diabetes)
```

```
## [1] 768   9
```

```
str(diabetes)
```

```
## 'data.frame':    768 obs. of  9 variables:
##  $ Pregnancies             : int  6 1 8 1 0 5 3 10 2 8 ...
##  $ Glucose                 : int  148 85 183 89 137 116 78 115 197 125 ...
##  $ BloodPressure           : int  72 66 64 66 40 74 50 0 70 96 ...
##  $ SkinThickness           : int  35 29 0 23 35 0 32 0 45 0 ...
##  $ Insulin                 : int  0 0 0 94 168 0 88 0 543 0 ...
##  $ BMI                     : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
##  $ DiabetesPedigreeFunction: num  0.627 0.351 0.672 0.167 2.288 ...
##  $ Age                     : int  50 31 32 21 33 30 26 29 53 54 ...
##  $ Outcome                 : int  1 0 1 0 1 0 1 0 1 1 ...
```

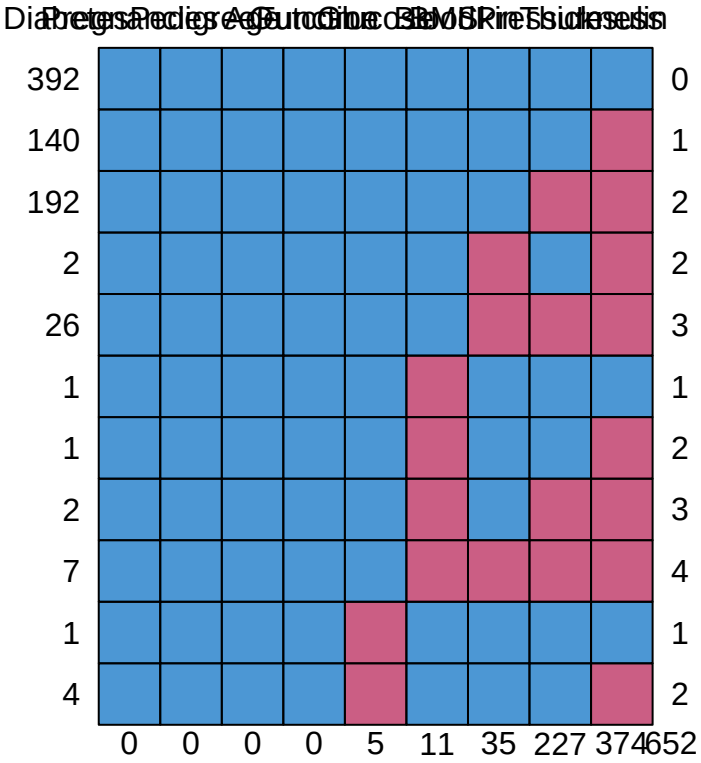
The unnoted observations that are currently filled with 0 have to be filed with NA instead of 0. We don't apply it for the 8th column, since that's a factor output.

```
diabetes[, 2:7][diabetes[, 2:7] == 0] <- NA
head(diabetes)
```

```
##   Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1           6     148            72            35      NA 33.6
## 2           1      85            66            29      NA 26.6
## 3           8     183            64            NA      NA 23.3
## 4           1      89            66            23      94 28.1
## 5           0     137            40            35     168 43.1
## 6           5     116            74            NA      NA 25.6
##   DiabetesPedigreeFunction Age Outcome
## 1                    0.627  50       1
## 2                    0.351  31       0
## 3                    0.672  32       1
## 4                    0.167  21       0
## 5                    2.288  33       1
## 6                    0.201  30       0
```

Checking the pattern of the missing data

```
md.pattern(diabetes)
```




```
##     Pregnancies DiabetesPedigreeFunction Age Outcome Glucose BMI BloodPressure
## 392           1                        1   1       1       1   1             1
## 140           1                        1   1       1       1   1             1
## 192           1                        1   1       1       1   1             1
## 2             1                        1   1       1       1   1             0
## 26            1                        1   1       1       1   1             0
## 1             1                        1   1       1       1   0             1
## 1             1                        1   1       1       1   0             1
## 2             1                        1   1       1       1   0             1
## 7             1                        1   1       1       1   0             0
## 1             1                        1   1       1       0   1             1
## 4             1                        1   1       1       0   1             1
##               0                        0   0       0       5  11            35
##     SkinThickness Insulin
## 392             1       1   0
## 140             1       0   1
## 192             0       0   2
## 2               1       0   2
## 26              0       0   3
## 1               1       1   1
## 1               1       0   2
## 2               0       0   3
## 7               0       0   4
## 1               1       1   1
## 4               1       0   2
##               227     374 652
```

There are no NA values in the columns: Pregnancies, DiabetesPedigreeFunction, Age, Outcome.There are 5 NA values for Glucose, 11 NA values for BMI, 35 NA values for BloodPressure,227 NA values for SkinThickness and 374 NA values for Insulin, which leads to a total of 652 NA's.

Missing data pattern by variable pairs

```
p <- md.pairs(diabetes); p
```

```
## $rr
##                           Pregnancies Glucose BloodPressure SkinThickness
## Pregnancies                       768     763           733           541
## Glucose                           763     763           728           536
## BloodPressure                     733     728           733           539
## SkinThickness                     541     536           539           541
## Insulin                           394     393           394           394
## BMI                               757     752           729           539
## DiabetesPedigreeFunction          768     763           733           541
## Age                               768     763           733           541
## Outcome                           768     763           733           541
##                           Insulin BMI DiabetesPedigreeFunction Age Outcome
## Pregnancies                   394 757                      768 768     768
## Glucose                       393 752                      763 763     763
## BloodPressure                 394 729                      733 733     733
## SkinThickness                 394 539                      541 541     541
## Insulin                       394 393                      394 394     394
## BMI                           393 757                      757 757     757
## DiabetesPedigreeFunction      394 757                      768 768     768
## Age                           394 757                      768 768     768
## Outcome                       394 757                      768 768     768
##
## $rm
##                           Pregnancies Glucose BloodPressure SkinThickness
## Pregnancies                         0       5            35           227
## Glucose                             0       0            35           227
## BloodPressure                       0       5             0           194
## SkinThickness                       0       5             2             0
## Insulin                             0       1             0             0
## BMI                                 0       5            28           218
## DiabetesPedigreeFunction            0       5            35           227
## Age                                 0       5            35           227
## Outcome                             0       5            35           227
##                           Insulin BMI DiabetesPedigreeFunction Age Outcome
## Pregnancies                   374  11                        0   0       0
## Glucose                       370  11                        0   0       0
## BloodPressure                 339   4                        0   0       0
## SkinThickness                 147   2                        0   0       0
## Insulin                         0   1                        0   0       0
## BMI                           364   0                        0   0       0
## DiabetesPedigreeFunction      374  11                        0   0       0
## Age                           374  11                        0   0       0
## Outcome                       374  11                        0   0       0
##
## $mr
##                           Pregnancies Glucose BloodPressure SkinThickness
## Pregnancies                         0       0             0             0
## Glucose                             5       0             5             5
## BloodPressure                      35      35             0             2
## SkinThickness                     227     227           194             0
## Insulin                           374     370           339           147
```

```
## BMI                              11      11              4              2
## DiabetesPedigreeFunction          0       0              0              0
## Age                               0       0              0              0
## Outcome                           0       0              0              0
##                          Insulin BMI DiabetesPedigreeFunction Age Outcome
## Pregnancies                    0   0                        0   0       0
## Glucose                        1   5                        5   5       5
## BloodPressure                  0  28                       35  35      35
## SkinThickness                  0 218                      227 227     227
## Insulin                        0 364                      374 374     374
## BMI                            1   0                       11  11      11
## DiabetesPedigreeFunction       0   0                        0   0       0
## Age                            0   0                        0   0       0
## Outcome                        0   0                        0   0       0
##
## $mm
##                          Pregnancies Glucose BloodPressure SkinThickness
## Pregnancies                        0       0             0             0
## Glucose                            0       5             0             0
## BloodPressure                      0       0            35            33
## SkinThickness                      0       0            33           227
## Insulin                            0       4            35           227
## BMI                                0       0             7             9
## DiabetesPedigreeFunction           0       0             0             0
## Age                                0       0             0             0
## Outcome                            0       0             0             0
##                          Insulin BMI DiabetesPedigreeFunction Age Outcome
## Pregnancies                    0   0                        0   0       0
## Glucose                        4   0                        0   0       0
## BloodPressure                 35   7                        0   0       0
## SkinThickness                227   9                        0   0       0
## Insulin                      374  10                        0   0       0
## BMI                           10  11                        0   0       0
## DiabetesPedigreeFunction       0   0                        0   0       0
## Age                            0   0                        0   0       0
## Outcome                        0   0                        0   0       0
```

We are using the 'missForest' package for imputing the NA values in our data set.

```
set.seed(287)
imp_diabetes <- missForest(diabetes)
```

We don't use the 'mice' package to impute the values since, 'mice' function creates multiply imputed data sets(mids), the lda() function doesn't accept all the mids and later allow us to make a pooled model(like what can be done with lm() function)

The normalized root mean squared error (NRMSE) is defined as:

$$\sqrt{\frac{mean((X_{true} - X_{imp})^2)}{var(X_{True})}}$$

The NRMSE in this case can be calculated by

```
imp_diabetes$OOBerror
```

```
##     NRMSE
## 0.5921172
```

Data Partition and Modeling

```
set.seed(673)
twin_indices = twin(imp_diabetes$ximp, r=5)
diabetes_test = imp_diabetes$ximp[twin_indices, ]
diabetes_train = imp_diabetes$ximp[-twin_indices, ]
```

Training the model

```
lda_model <- MASS::lda(Outcome ~., data = diabetes_train)
preds_train <- predict(lda_model)
head(preds_train$posterior)
```

```
##           0          1
## 1 0.2980953 0.70190471
## 2 0.9641042 0.03589583
## 3 0.1930411 0.80695892
## 4 0.9654624 0.03453762
## 5 0.1238325 0.87616750
## 6 0.8725070 0.12749298
```

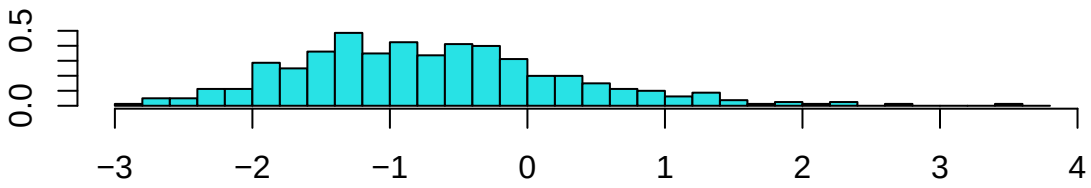
The output above shows the probabilities of being classified into the 'Diabetes'(1) or 'No Diabetes'(0) group. For example, observation one has not been tested positive for diabetes with a probability of 98%. Observation two has been diagnosed with diabetes with a probability of 87%. The model uses a 50% threshold for the posterior probabilities.

```
lda_model
```

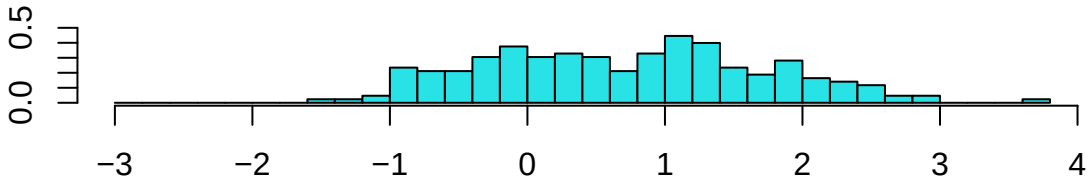
```
## Call:
## lda(Outcome ~ ., data = diabetes_train)
##
## Prior probabilities of groups:
##         0         1
## 0.6530945 0.3469055
##
## Group means:
##   Pregnancies  Glucose BloodPressure SkinThickness  Insulin      BMI
## 0    3.309227 110.7227      70.76519      27.10931 130.8583 30.85762
## 1    4.901408 141.5348      74.84681      32.61962 207.9294 35.35402
##   DiabetesPedigreeFunction     Age
## 0                0.4278653 31.1596
## 1                0.5609061 37.1784
##
## Coefficients of linear discriminants:
##                                 LD1
## Pregnancies               0.0922984500
## Glucose                   0.0265552580
## BloodPressure            -0.0074817417
## SkinThickness             0.0065036779
## Insulin                   0.0005801987
## BMI                       0.0575026441
## DiabetesPedigreeFunction  0.6055522244
## Age                       0.0090370577
```

From our output we can read off the prior probabilities $\pi_1 = 0.347$ and $\pi_2 = 0.653$. This means that around 34.7% of our data set includes people who have been diagnosed with diabetes and 65.3% who have not been diagnosed with diabetes.

```
plot(lda_model)
```



group 0



group 1

More specifically, the scores, or coefficients of the output of the linear discriminant, are a linear combination that forms the LDA decision rule. When the linear combination of these coefficients is negative, then the probability increases that observation has diabetes (see plot), whereas when the linear combination is positive, observation is more likely to belong to the "No Diabetes" group.

Using the posterior for the test set we try predicting whether they will have diabetes or not

```
preds_test <- predict(lda_model,diabetes_test)
head(preds_test$posterior)
```

```
##             0         1
## 686 0.6994258 0.3005742
## 298 0.8266944 0.1733056
## 352 0.6912457 0.3087543
## 569 0.4706025 0.5293975
## 629 0.6317891 0.3682109
## 726 0.6971475 0.3028525
```

Validating the model using the test set, we use a confusion matrix to tabulate our finding

```
diabetes_test <- data.frame(diabetes_test, predicted = preds_test$class)
xtabs(~ predicted + Outcome, data = diabetes_test)
```

```
##           Outcome
## predicted  0  1
##         0 88 25
##         1 11 30
```

```
# prediction accuracy
round((89+31)/(154), 4)
```

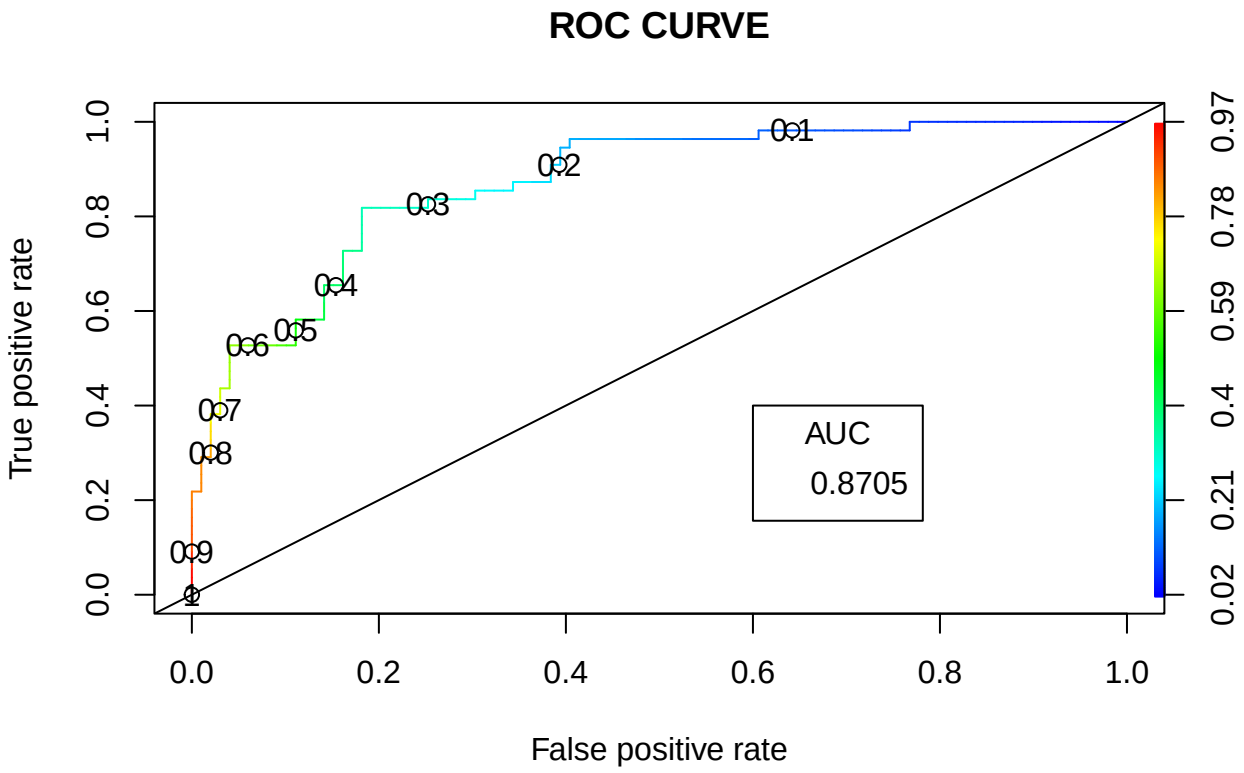
```
## [1] 0.7792
```

The prediction accuracy is 77.92% if the threshold for the posterior is 0.5.

```
ROCPred <- prediction(preds_test$posterior[,2], diabetes_test$Outcome)
ROCPer <- performance(ROCPred, measure = "tpr", x.measure = "fpr")
auc <- performance(ROCPred, measure = "auc")
auc <- auc@y.values[[1]]
auc
```

```
## [1] 0.8705234
```

```
plot(ROCPer, colorize = TRUE,
print.cutoffs.at = seq(0.1, by = 0.1),
main = "ROC CURVE")
abline(a = 0, b = 1)
auc <- round(auc, 4)
legend(.6, .4, auc, title = "AUC", cex = 1)
```



We have an AUC of about 87.05%, which considerably good for predicting whether a person has a possibility of having diabetes in the next few years.