STAT 154 Lab 9: Classification with Discriminant Analysis

Yuansi Chen and Raaz Dwivedi

Apr 15, 2019

1 White-board discussion with LDA/QDA

2 LDA and QDA by hand

We have training data for a two class classification problem as laid out in Figure 1. The black dots are examples of the positive class ($y = +1$) and the white dots examples of the negative class ($y = -1$)[1].

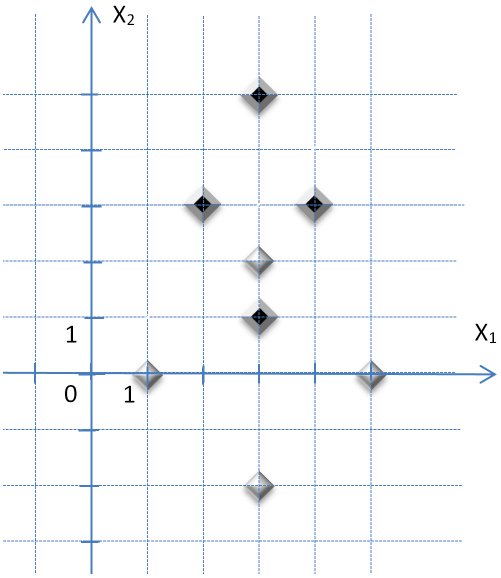

Figure 1: Draw your answers to the QDA problem.

(a) Draw on Figure 1 the position of the class centroids $\mu_{(+)}$ and $\mu_{(-)}$ for the positive and negative class respectively, and indicate them as circled $(+)$ and $(-)$. Give their coordinates:

$$\mu_{(+)} = \begin{bmatrix} \quad \end{bmatrix} \qquad \mu_{(-)} = \begin{bmatrix} \quad \end{bmatrix}$$

(b) Compute the covariance matrices for each class:

$$\Sigma_{(+)} = \begin{bmatrix} \quad \end{bmatrix} \qquad \Sigma_{(-)} = \begin{bmatrix} \quad \end{bmatrix}$$

[1] This problem borrows ideas for the content of CS 189, Spring 2018.

(c) Assume each class has data distributed according to a bi-variate Gaussian, centered on the class centroids computed in question (a). Draw on Figure 1 the contour of equal likelihood $p(X = x|Y = y)$ going through the data samples, for each class. Indicate with light lines the principal axes of the data distribution for each class.

(d) Compute the determinant and the inverse of $\Sigma_{(+)}$ and $\Sigma_{(-)}$:

$$|\Sigma_{(+)}| = \qquad\qquad |\Sigma_{(-)}| =$$

$$\Sigma_{(+)}^{-1} = \begin{bmatrix} & \\ & \end{bmatrix} \qquad\qquad \Sigma_{(-)}^{-1} = \begin{bmatrix} & \\ & \end{bmatrix}$$

(e) The likelihood of examples of the positive class is given by:

$$p(X = x|Y = +1) = \frac{1}{2\pi|\Sigma_{(+)}|^{1/2}} \exp\big(-\frac{1}{2}(x - \mu_{(+)})^T \Sigma_{(+)}^{-1}(x - \mu_{(+)})\big)$$

and there is a similar formula for $p(X = x|Y = -1)$. Compute $f_{(+)}(x) = \log\big(p(X = x|Y = +1)\big)$ and $f_{(-)}(x) = \log\big(p(X = x|Y = -1)\big)$. Then compute the discriminant function $f(x) = f_{(+)}(x) - f_{(-)}(x)$:

$f_{(+)}(x) =$

$f_{(-)}(x) =$

$f(x) =$

(f) Draw on Figure 1 for each class contours increasing equal likelihood. Geometrically construct the Bayes optimal decision boundary. Compare to the formula obtained with $f(x) = 0$ after expressing x_2 as a function of x_1:

$x_2 =$

What type of function is it?

(g) Now assume $p(Y = -1) \neq p(Y = +1)$, how does it change the decision boundary?

(h) Repeat your answers if the method of choice was LDA. Which method do you expect would do better if LDA or QDA?

3 The *Default* Data Set

Consider the **Default** data set that comes in the R package **ISLR**. We are interested in predicting whether an individual will default on his or her credit card payment, on the basis of annual income and monthly credit card balance.

```
library(ISLR)
library(ggplot2)
colnames(Default)
dim(Default)
```

```r
summary(Default)
```

- Begin with some exploratory analysis of the data. You can start with a scatterplot of balance and income, distinguishing observations based on default, like in the images below.
- Make density plots of balance and income, for instance.
- Try fitting OLS model by regressing **default** on **balance**. What is wrong with the OLS fit?

```r
# transform default as numeric
default_numeric <- rep(0, nrow(Default))
default_numeric[Default$default == 'Yes'] <- 1
Default$default_num <- default_numeric

ols_reg <- lm(default_num ~ balance, data = Default)
summary(ols_reg)
```

- Think about how LDA will perform on this dataset. Try LDA in R.

4 LDA / QDA in R

We now compare and contrast an LDA/QDA classifier on iris dataset to predict the Species based on other features. You may find the following libraries useful: **caret** library (*confusionMatrix* function) to compute the errors and **klaR** library (*partimat* function) to visualize the boundary. Some useful links:

- Tutorial on LDA `https://rpubs.com/ifn1411/LDA`.
- Using ggplot2 for visualizing the data `https://stackoverflow.com/questions/20197106/linear-discriminant-analysis-plot-using-ggplot2`

1. Do a $50\% - 50\%$ split for the training and test set. Compute the train and test-error (for LDA/QDA).
2. For LDA, plot the projection of the data on the linear-discriminant vectors (how many of them are there?) and color the points using the real Species label. Can you guess the boundary?