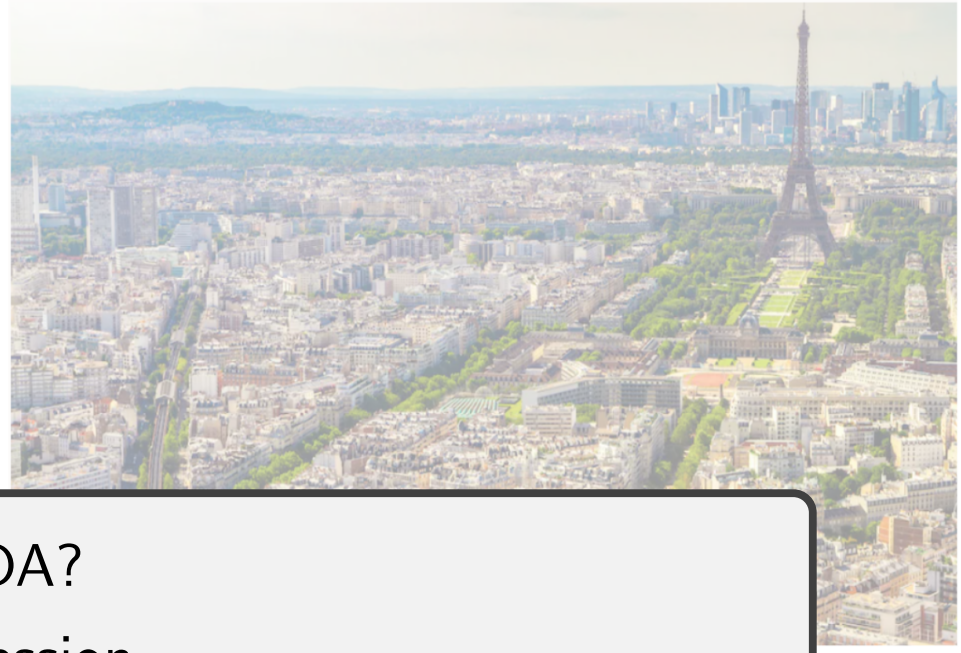
Linear discriminant analysis, quadratic discriminant analysis and my next vacation in Europe

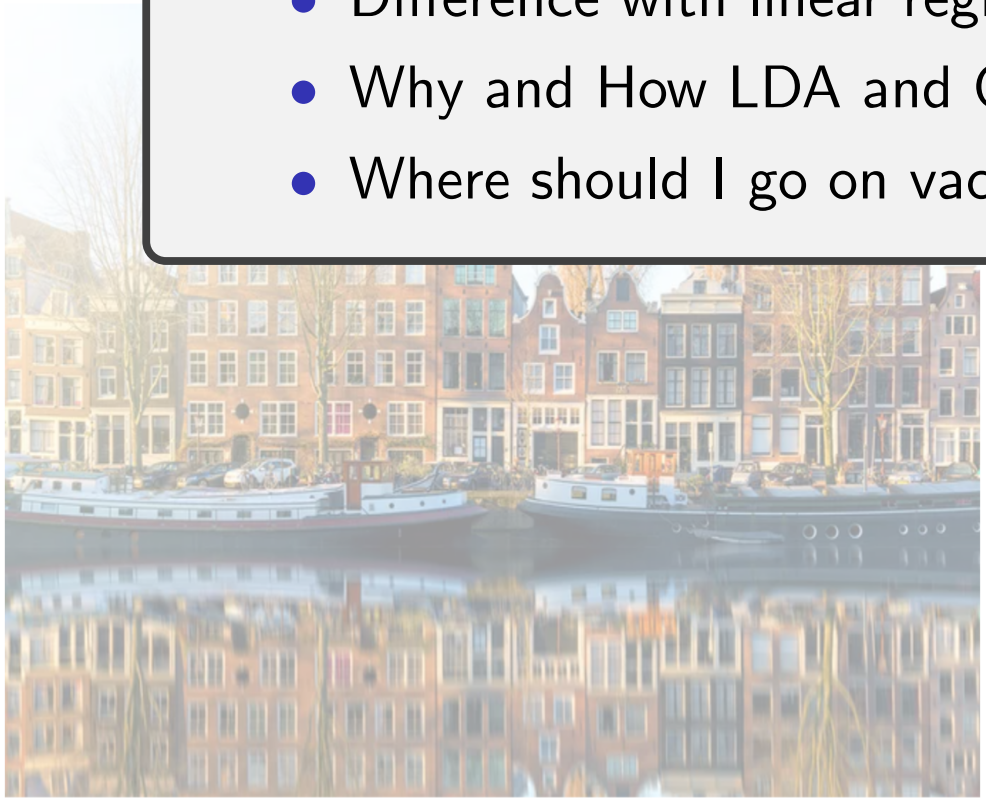
Isis A. Gallardo

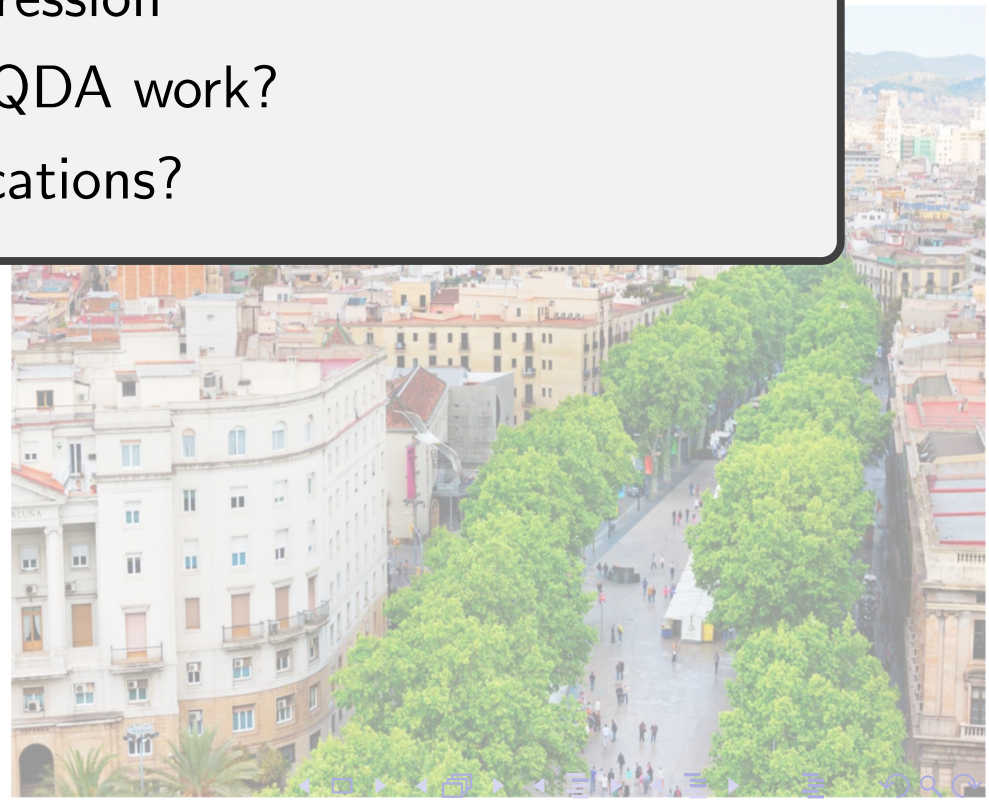
March 2023

Outline



- What is LDA? What is QDA?
- Difference with linear regression
- Why and How LDA and QDA work?
- Where should I go on vacations?

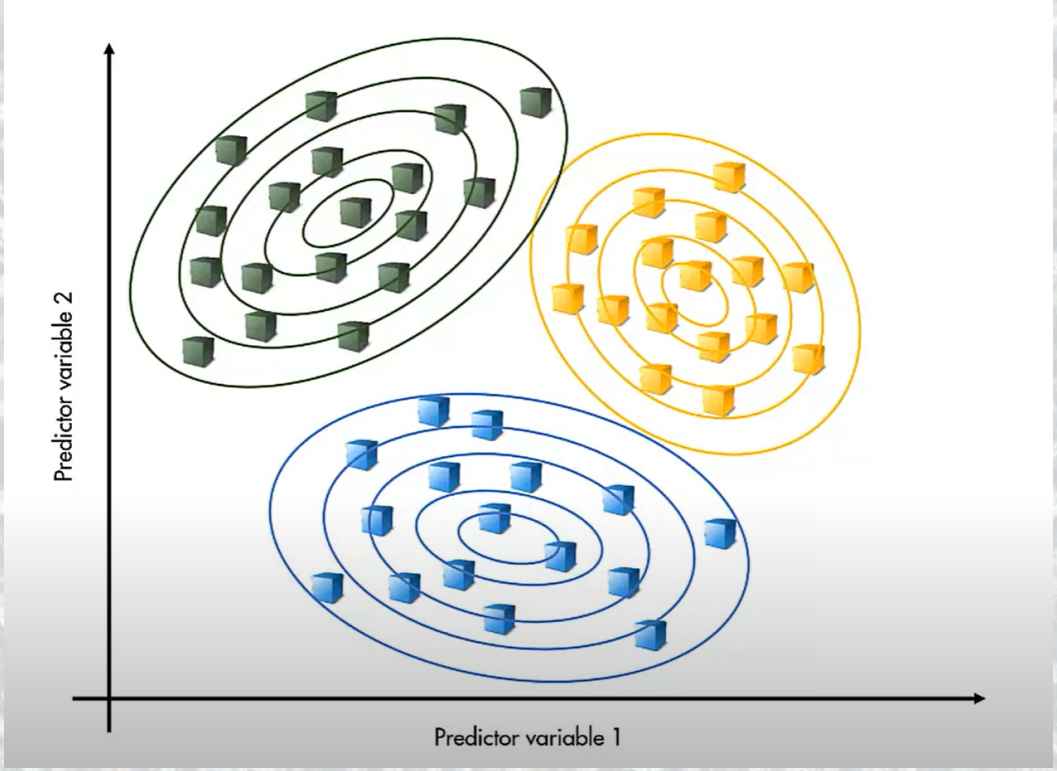




What are LDA and QDA

LDA and QDA are methods that allow us to model our data in order to recognize classifications.

Example



LDA and QDA allow us to generate models for the population that help us decide if a point belongs to one or another class.

Difference between LDA/QDA and logistic regression

In logistic regression, we calculate $P(C_i|\mathbf{x})$ while LDA/QDA uses $P(\mathbf{x}|C_i)$ and then use Bayes rule:

$$P(C_i|\mathbf{x}) = \frac{P(x|C_i)P(C_i)}{\sum_i P(x|C_i)P(C_i)}$$

Assumptions

Consider an n-dimensional data set, such that our data belong to k different classes. Both LDA and QDA methods can be used when we can assume that our classes have n-dimensional normal distributions, i.e. for all $i = 1, \ldots, k$.

$$f_i(x) = \frac{1}{\sqrt{(2\pi)^n det \sum_i}} exp[-\frac{1}{2}(x-\mu_i)^T \sum_i^{-1} (x-\mu_i)]$$

where

μ_i : is the mean

$\sum_i$: covariance matrix

Quadratic discriminant Analysis (QDA) and Linear discriminant Analysis (LDA)

Suppose we have two classes C_1 and C_2 with $\pi_1 = \pi_2$.
We would like to find the decision boundary. That is, a set of pints $\mathbf{x}$ such that

$$P(\mathbf{x}|C_1) = P(\mathbf{x}|C_2)$$

Meaning that

$$\frac{1}{\sqrt{(2\pi)^n det \sum_1}} exp[-\frac{1}{2}(x-\mu_1)^T \sum_1{}^{-1}(x-\mu_1)] = \frac{1}{\sqrt{(2\pi)^n det \sum_2}} exp[-\frac{1}{2}(x-\mu_2)^T \sum_2{}^{-1}(x-\mu_2)]$$

Applying ln( ) both sides

$$-\frac{1}{2} ln(\Sigma_1) - \frac{1}{2}(x-\mu_1)^T \sum_1{}^{-1}(x-\mu_1) = -\frac{1}{2} ln(\Sigma_2) - \frac{1}{2}(x-\mu_2)^T \sum_2{}^{-1}(x-\mu_2)$$

If we assume $\Sigma_1 = \Sigma_2$

$$(x-\mu_1)^T \sum_1{}^{-1}(x-\mu_1) = (x-\mu_2)^T \sum_2{}^{-1}(x-\mu_2)$$

How does it work?

Suppose we have a sample data set $C_1 \cup \ldots \cup C_k$ where C_i denotes the class i. We need to start by estimating a model for C_i for $i = 1, \ldots, k$. We assume the classes have normal distributions, so we can estimate μ_i, Σ_i and μ_i by:

$$\hat{\pi}_i = \frac{n_i}{n}$$

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}$$

$$\hat{\Sigma}_i = \frac{1}{n_i - 1} \sum_{\mathbf{x} \in C_i} (x - \mu_i)(x - \mu_i)^T$$

For LDA, all $\hat{\Sigma}_i$ must be the same then

$$\hat{\Sigma} = \frac{1}{n - k} \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} (x - \mu_i)(x - \mu_i)^T$$

The Important stuff: Where to vacation

Consider the Data set "Airbnb Prices in European Cities" from Kaggle. We narrow down our options to 4 cities (Amsterdam, Paris, Barcelona, and Athens) and we consider 4 important characteristics of the most popular Airbnb rentals in these cities:

- Prices of Airbnb rental.
- Distance from Metro/Town.
- Cleaning rating.
- Overall customer rating.

Data set

	A	B	C	D	E	F	G
1	City	Price	Distance	Clean	Rate		
2	Paris	356.7434	0.014268	10	85		
3	Barcelona	174.7853	0.017045	9	88		
4	Paris	261.441	0.020239	10	98		
5	Paris	297.325	0.020249	8	80		
6	Paris	303.3834	0.027891	8	85		
7	Amsterdan	310.9695	0.038355	9	94		
8	Paris	336.0052	0.038691	10	100		
9	Paris	324.3546	0.042113	10	99		
10	Paris	262.839	0.042178	10	90		
11	Barcelona	180.6037	0.043091	9	77		
12	Paris	1077.454	0.048533	4	20		
13	Barcelona	117.9975	0.050492	8	84		
14	Paris	284.9753	0.052276	10	93		
15	Barcelona	173.6216	0.053187	9	86		
16	Paris	393.7925	0.056391	7	86		
17	Amsterdan	420.4063	0.057698	9	94		
18	Paris	474.6481	0.058714	10	97		
19	Athens	405.6429	0.065383	10	90		
20	Paris	340.4325	0.065837	9	65		
21	Athens	76.39491	0.066279	9	94		
22	Paris	435.5019	0.069532	9	80		

DataCities

Implementing LDA

First, we divide our data into training and testing data. Then we train our model and test it.

```
# Select our data and divide it into "training" and "test" data

z=read.csv(file="C:/Users/isisa/Dropbox/DU/Winter 2023/Statistical Inference with R/Final/DataCities.csv")
train_indices <- createDataPartition(z$City,p=0.6,list = FALSE,times = 1)
a=z[train_indices , ]
b= z[-train_indices, ]

# Apply the LDA method
m=MASS::lda(City ~Price + Distance + Clean + Rate, data = a)

# Test the model
x=predict(m,b)
y=x$class
table(y,b[,1])

# Where shuld I go?
z=read.csv(file="C:/Users/isisa/Dropbox/DU/Winter 2023/Statistical Inference with R/Final/mytrip.csv")
where=predict(m,z)
w=where$class
```

The result.

After applying LDA

```
Call:
lda(City ~ Price + Distance + Clean + Rate, data = a)

Prior probabilities of groups:
Amsterdam    Athens Barcelona     Paris
     0.25      0.25      0.25      0.25

Group means:
              Price  Distance    Clean     Rate
Amsterdam 575.4349 1.0906719 9.540000 94.82000
Athens    147.6968 0.4591003 9.633333 95.32667
Barcelona 300.2475 0.3579288 9.340000 91.12667
Paris     489.4088 0.2036761 9.240000 90.55333

Coefficients of linear discriminants:
                  LD1         LD2          LD3
Price    -0.002031166  0.00256468  0.001065321
Distance -1.900019322 -0.73753223 -0.811198220
Clean     0.058469620 -0.12979110 -0.559481838
Rate     -0.017044865 -0.07088569  0.161797480

Proportion of trace:
   LD1    LD2    LD3
0.7282 0.2590 0.0128
```

The result

Test our model by creating a confusion matrix

```
> # Test the model
> x=predict(m,b)
> y=x$class
> table(y,b[,1])

y            Amsterdam Athens Barcelona Paris
  Amsterdam         55      5         2     0
  Athens            15     83        36    16
  Barcelona         13     10        42    24
  Paris             17      2        20    60
>
```

Now, assuming my budget is 150 Euros a day, I expect a distance of 0.5 to the metro town, a cleaning score of 8, and a general rate of 80, my ideal Airbnb is in the city...

```
> w
[1] Barcelona
Levels: Amsterdam Athens Barcelona Paris
```


Thank you!!