Preamble to Discriminant Analysis (part 1)

Predictive Modeling & Statistical Learning

Gaston Sanchez

Introduction

Introduction

In these slides I'll talk about the concept of Variance decomposition taking into account a group structure (i.e. analysis of variance or *one-way anova*).

The idea is to layout a couple of foundational principles that should allow you to understand discriminant methods in a more comprehensive way.

BTW: this material is not in the textbooks *ISL* and *APM*.

Iris Data

Dataset `iris` in R

$n = 150$ Observations, i.e. iris flowers

$p = 4$ predictors

- `Sepal.Length`
- `Sepal.Width`
- `Petal.Length`
- `Petal.Width`

One response (categorical)

- `Species` (3 classes: setosa, versicolor, virginica)

Famous data set collected by Edgar Anderson (1935), and used by Ronald Fisher (1936) in his paper about Discriminant Analysis.

Dataset iris in R

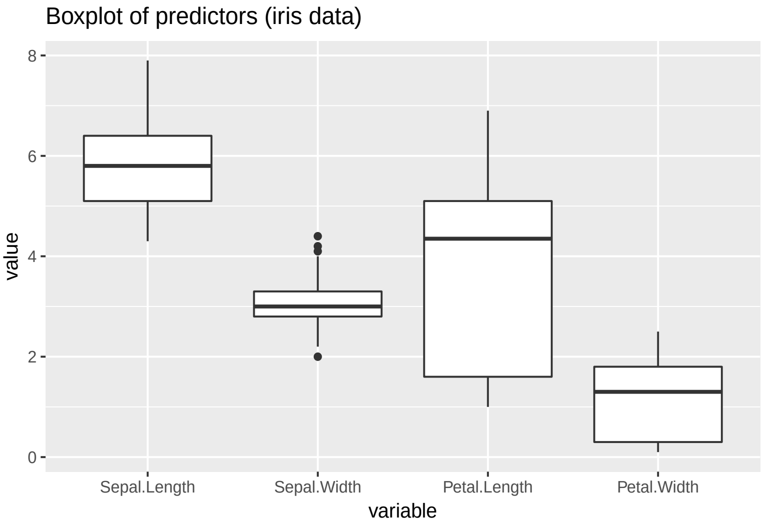
```
head(iris)

  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

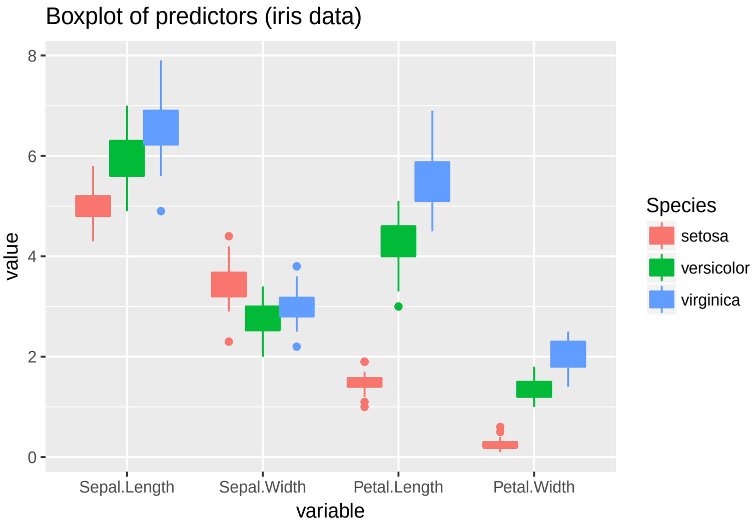
Dataset iris in R

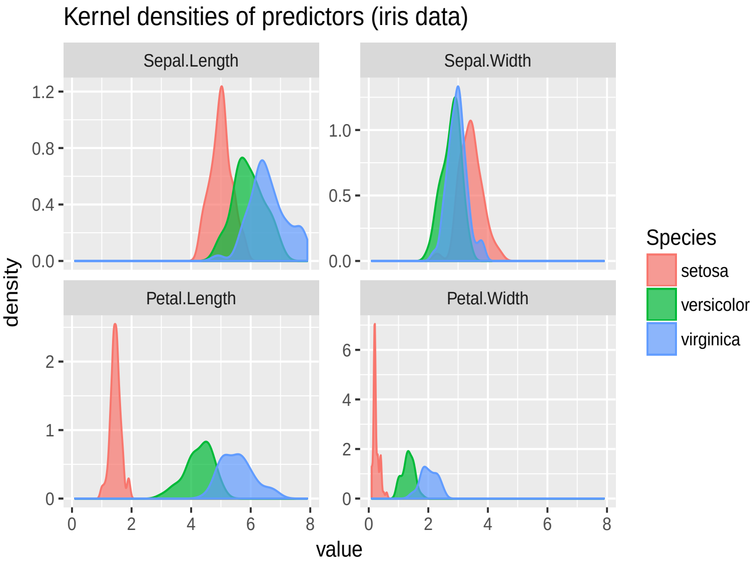
```
summary(iris)
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```


Boxplot of predictors (iris data)
value
8
6
4
2
0
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
variable

Let's take into account the group structure


Boxplot of predictors (iris data)
value
8
6
4
2
0
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
variable
Species
setosa
versicolor
virginica


Kernel densities of predictors (iris data)
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
density
value
1.2
0.8
0.4
0.0
1.0
0.5
0.0
2
1
0
6
4
2
0
0
2
4
6
8
0
2
4
6
8
Species
🟥 setosa
🟩 versicolor
🟦 virginica

```r
library(reshape2)
library(ggplot2)

iris_melt <- melt(iris, id = "Species")

ggplot(data = iris_melt, aes(x = variable, y = value)) +
  geom_boxplot() +
  ggtitle("Boxplot of predictors (iris data)")

ggplot(data = iris_melt, aes(x = variable, y = value)) +
  geom_boxplot(aes(fill = Species, color = Species)) +
  ggtitle("Boxplot of predictors (iris data)")

ggplot(data = iris_melt, aes(x = value)) +
  geom_density(aes(fill = Species, color = Species),
               alpha = 0.7) +
  facet_wrap(~ variable, scales = 'free_y') +
  ggtitle("Kernel densities of predictors (iris data)")
```

Which predictor provides the
“best” distinction between Species?

One-Way Analysis of Variance

Caveat: messy notation

In regression problems we've been using two indices i and j

- i for objects, $i = 1, \ldots, n$
- j for predictors, $j = 1, \ldots, p$

New index k

Now we have a new index k for groups or classes, $k = 1, \ldots, K$.

First step: focus on group means

- In classification problems, the response variable Y provides a group structure to the data.
- We expect that the predictors will help us to differentiate (i.e. discriminate) between one group and the others.
- The general idea is to look for systematic differences among groups. But how?
- A “natural” way to look for differences is paying attention to group means.

Group (or class) structure

Consider a single predictor X and a categorical response Y measured on n individuals.

Let's take into account the group structure conveyed by Y

- Assume there are K groups
- Let G_k represent the k-th group in Y
- Let n_k be the number of observations in group G_k,
- Then:

$$n = n_1 + n_2 + \cdots + n_K = \sum_{k=1}^{K} n_k$$

Overall Mean and Group Means

The (global) mean value of X is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Each group k will have its mean $\bar{x}_k$:

$$\bar{x}_k = \frac{1}{n_k} \sum_{i \in G_k} x_{ik}$$

Sum of Squares Dispersions

Recall that the (global) dispersion in X is given by the *total sum of squares* (tss):

$$\mathsf{tss} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Each group k will also have its own sum-of-squares ss_k

$$\mathsf{ss}_k = \sum_{i \in G_k} (x_{ik} - \bar{x}_k)^2$$

Focus on Group Means

- One way to look for systematic differences between the groups is to compare their means.
- If there's no group difference in X, then the group means $\bar{x}_k$ should be similar.
- If there is really a diffrence, it is likely that one or more of the mean values will differ.

Between Sum of Squares

A useful measure to compare differences among the k means is the deviation from the overall mean: $\bar{x}_k - \bar{x}$

An effective summary of these deviations is the so-called **between-group sum of squares** (bss) given by:

$$\mathsf{bss} = \sum_{k=1}^{K} n_k (x_k - \bar{x})^2$$

Within Sum of Squares

To assess the relative magnitude of the between sum of squares (bss), we need to compare it to a measure of the "background" variation.

Such a measure of background variation can be formed by combining the group variances into a pooled-estimate called **within-group sum of squares** (wss):

$$\mathsf{wss} = \sum_{k=1}^{K} \sum_{i \in G_k} (x_{ik} - \bar{x}_k)^2 = \mathsf{ss}_1 + \cdots + \mathsf{ss}_K$$

Sums of Squares

So far we have three types of sums of squares:

$$\textit{total} \quad \mathsf{tss} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\textit{between} \quad \mathsf{bss} = \sum_{k=1}^{K} n_k (x_k - \bar{x})^2$$

$$\textit{within} \quad \mathsf{wss} = \sum_{k=1}^{K} \sum_{i \in G_k} (x_{ik} - \bar{x}_k)^2$$

Decomposition of sums-of-squares

An important aspect has to do with looking at the squared deviations: $(x_i - \bar{x})^2$ in terms of the group structure.

A useful trick is to rewrite the deviation terms $x_i - \bar{x}$ as:

$$\begin{aligned} x_i - \bar{x} &= x_i - (\bar{x}_k - \bar{x}_k) - \bar{x} \\ &= (x_i - \bar{x}_k) + (\bar{x}_k - \bar{x}) \end{aligned}$$

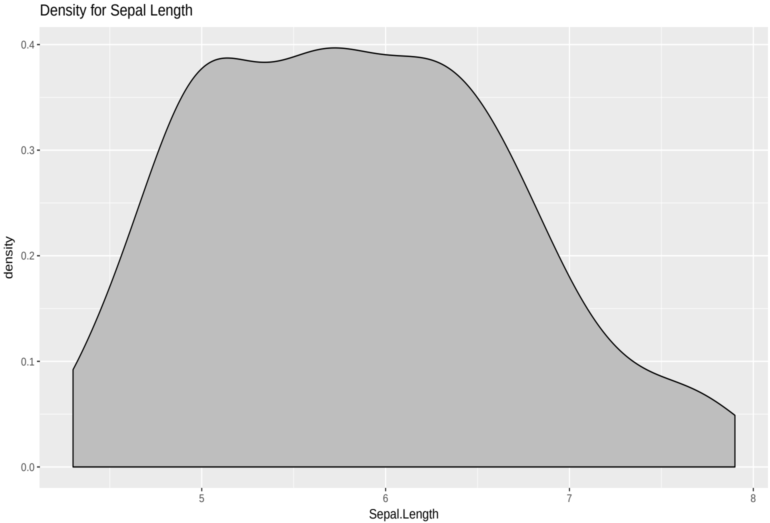
Sum of Squares Decomposition

We can decompose tss in terms of bss and wss:

$$\underbrace{\sum_{k=1}^{K} \sum_{i \in G_k} (x_{ik} - \bar{x})^2}_{\text{tss}} = \underbrace{\sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})^2}_{\text{bss}} + \underbrace{\sum_{i \in G_k} (x_{ik} - \bar{x}_k)^2}_{\text{wss}}$$

In summary:

$$\text{tss} = \text{bss} + \text{wss}$$



```
ggplot(data = iris, aes(x = Sepal.Length)) +
geom_density(fill = 'gray') +
ggtitle('Density for Sepal Length')
```

TSS for Sepal.Length

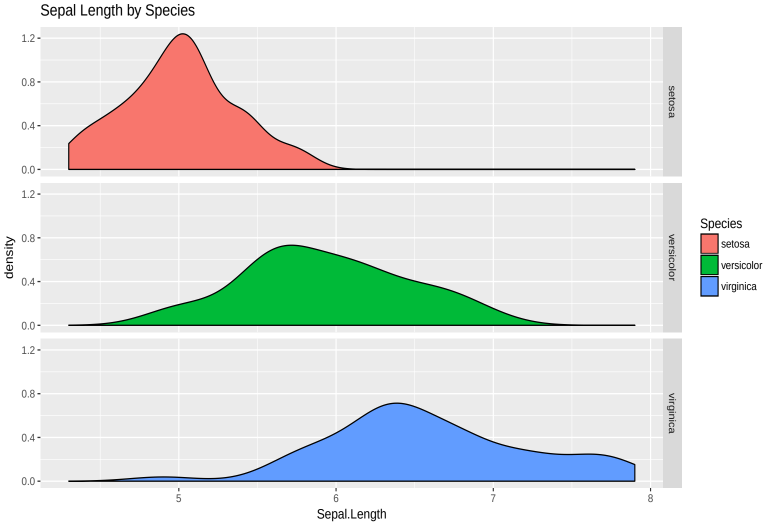
```r
x <- iris$Sepal.Length

# overall mean
x_bar <- mean(x)
x_bar
```

```
## [1] 5.843333
```

```r
# total sums-of-squares
tss <- sum((x - x_bar)^2)
tss
```

```
## [1] 102.1683
```



```
ggplot(data = iris, aes(x = Sepal.Length, group = Species)) +
  geom_density(aes(fill = Species)) +
  facet_grid(Species ~ .) +
  ggtitle('Sepal Length by Species')
```

BSS for Sepal.Length

```r
# Sepal Length group means
x_means <- tapply(x, iris$Species, mean)

# between sums-of-squares
bss <- sum(50 * (x_means - x_bar)^2)
bss
```

```
## [1] 63.21213
```

WSS for Sepal.Length

```r
# Sepal Length group sum of squares
w1 <- sum((x[1:50] - x_means[1])^2)
w2 <- sum((x[51:100] - x_means[2])^2)
w3 <- sum((x[101:150] - x_means[3])^2)

# within sums-of-squares
wss <- w1 + w2 + w3
wss


## [1] 38.9562
```

TSS Decomposition

Let’s check that we have:

$$\text{tss} = \text{bss} + \text{wss}$$

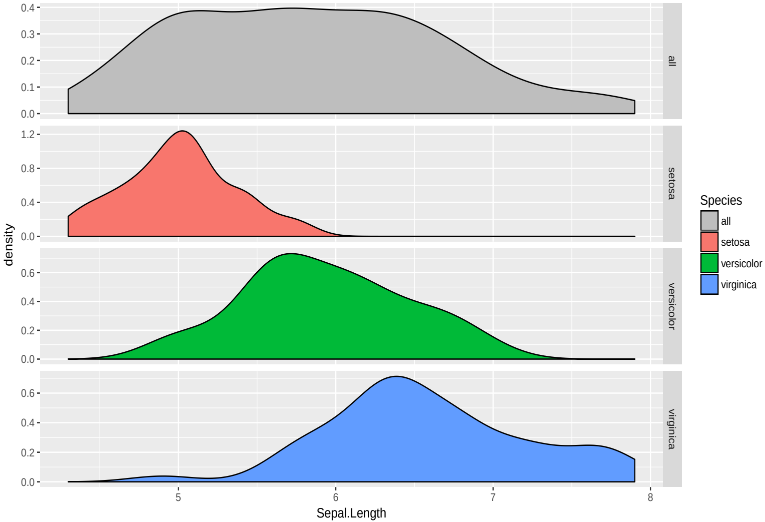
```
# tss
tss

## [1] 102.1683

# bss + wss
bss + wss

## [1] 102.1683
```

Dispersion in Sepal.Length



Derived Ratios from $\text{tss} = \text{bss} + \text{wss}$

Correlation Ratio

Correlation ratio η^2 (proposed by K. Pearson):

$$\eta^2(X, Y) = \frac{\mathsf{bss}}{\mathsf{tss}}$$

- η^2 takes vaues between 0 and 1.
- $\eta^2 = 0$ represents the special case of no dispersion among the means of the different groups.
- $\eta^2 = 1$ refers to no dispersion within the respective groups.

The correlation ratio is a measure of the relationship between the dispersion within groups and the dispersion across all individuals.

F Ratio

With $\text{tss} = \text{bss} + \text{wss}$, we can also calculate:

F ratio (proposed by R.A. Fisher):

$$F = \frac{\text{bss}/(K-1)}{\text{wss}/(n-K)}$$

The larger the value of both ratios, the more variability is there between groups than within groups.

Ratios for Sepal.Length

```
# correlation ratio
eta_sqr <- bss / tss
eta_sqr

## [1] 0.6187057

# F ratio
F_ratio <- (bss / (3 - 1)) / (wss / (150 - 3))
F_ratio

## [1] 119.2645
```

Ratios for all Variables

Let's compute the decompositions for all predictors, and obtain the correlation ratios and F ratios

```
etas

## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##    0.6187057    0.4007828    0.9413717    0.9288829

Fs

## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##    119.26450     49.16004   1180.16118    960.00715
```

F Ratio statistic

The F ratio can be used for hypothesis testing purposes. More formally, a null hypothesis postulates that the population means do not differ $(H_0 : \mu_1 = \mu_2 = \cdots = \mu_K = \mu)$ versus the alternative hypothesis H_1 that one or more population means differ among the K normally distributed populations.

Assuming or knowing that the variances of each sampled population are the same σ^2, a test statistic to assess the null hypothesis is:

$$F = \frac{\mathsf{bss}/(K-1)}{\mathsf{wss}/(n-K)}$$

which has an F-distribution with $K-1$ and $n-K$ degrees of freedom under the null hypothesis.

Next slides

In the next slides we'll:

- expand the notation for more than one predictor
- look at the geometric perspective
- consider dispersion in terms of matrices
  - sus-of-squares matrix
  - variance matrices

Bibliography

- **Principles of Multivariate Analysis: A User's Perspective** by W.J. Krzanowski (1988). *Chapter 11: Incorporating group structure: descriptive methods*. Wiley.
- **Practical Biostatistical Methods** by Steve Selvin (1995). *Chapter 1: General Concepts*. Duxbury Press.
- **Data Mining and Statistics for Decision Making** by Stephane Tuffery (2011). *Chapter 11: Classification and prediction methods*.
- **Multivariate Analysis** by Maurice Tatsuoka (1988). *Chapter 7: Discriminant Analysis and Canonical Correlation*.

French Literature

- **Statistique Exploratoire Multidimensionnelle** by Lebart et al (2004). *Chapter 3, section 3: Analyse factorielle discriminante*. Dunod, Paris.
- **Probabilites, analyse des donnees et statistique** by Gilbert Saporta (2011). *Chapter 18: Analyse discriminante et regression logistique*. Editions Technip, Paris.
- **Statistique explicative appliquee** by Nakache and Confais (2003). *Chapter 1: Analyse discriminante sur variables quantitatives*. Editions Technip, Paris.
- **Statistique: Methodes pour decrire, expliquer et prevoir** by Michel Tenenhaus (2008). *Chapter 10: L'analyse discriminante*. Dunod, Paris.