判别分析

徐平峰

长春工业大学

2021/3/30

Boston房价数据

- 响应变量cmedv为数值型变量

```
setwd("E:/teaching_plan_notes/msa11091083/rmd")#设置工作目录
w=read.csv("MVAPureData/BostonHousing2.csv")#读取Boston房价数据
w=w[,-c(1:5)]#去掉前5列变量，只分析后14个变量
head(w)#显示数据的前6行
```

```
#   cmedv    crim zn indus chas   nox    rm  age    dis rad tax ptratio
# 1  24.0 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 3
# 2  21.6 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 3
# 3  34.7 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 3
# 4  33.4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 3
# 5  36.2 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 3
# 6  28.7 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 3
#   lstat
# 1  4.98
# 2  9.14
# 3  4.03
# 4  2.94
# 5  5.33
# 6  5.21
```

水泥强度数据

- 该数据包含了混凝土 7 种成分, 年龄, 以及抗 压强度等 9 个变量. 共有 1030 个观测值.

这些变量为 Cement(水泥), Blast.Furnace.Slag(高炉矿渣), Fly.Ash(粉煤灰), Water(水), Superplasticizer(超塑化剂), Coarse.Aggregate(粗骨料), Fine.Aggregate(细骨料), Age(时间), Compressive.strength(抗压强度). 其中除了 Age(时间) 单位是天, Compressive.strength(抗压强度) 为 MPa(兆帕) 之外全部是在m3 号混合中的 kg(千克) 数.

- 响应变量Compressive.strength为数值型变量

水泥强度数据

- 响应变量Compressive.strength为数值型变量

```
w2=read.csv("E:/teaching_plan_notes/msa11091083/rmd/WuAlmPureData
head(w2)
```

```
#   Cement Blast.Furnace.Slag Fly.Ash Water Superplasticizer Coarse.Aggre
# 1  540.0                0.0       0   162               2.5          10
# 2  540.0                0.0       0   162               2.5          10
# 3  332.5              142.5       0   228               0.0           9
# 4  332.5              142.5       0   228               0.0           9
# 5  198.6              132.4       0   192               0.0           9
# 6  266.0              114.0       0   228               0.0           9
#   Fine.Aggregate Age Compressive.strength
# 1          676.0  28                79.99
# 2          676.0  28                61.89
# 3          594.0 270                40.27
# 4          594.0 365                41.05
# 5          825.5 360                44.30
# 6          670.0  90                47.03
```

鸢尾花

- R 中iris（鸢尾花）数据，三种不同的鸢尾花的 150 个样品的花瓣、花萼长、宽的数据。

```
dim(iris)
```

```
## [1] 150   5
```

```
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  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
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

iris（鸢尾花）

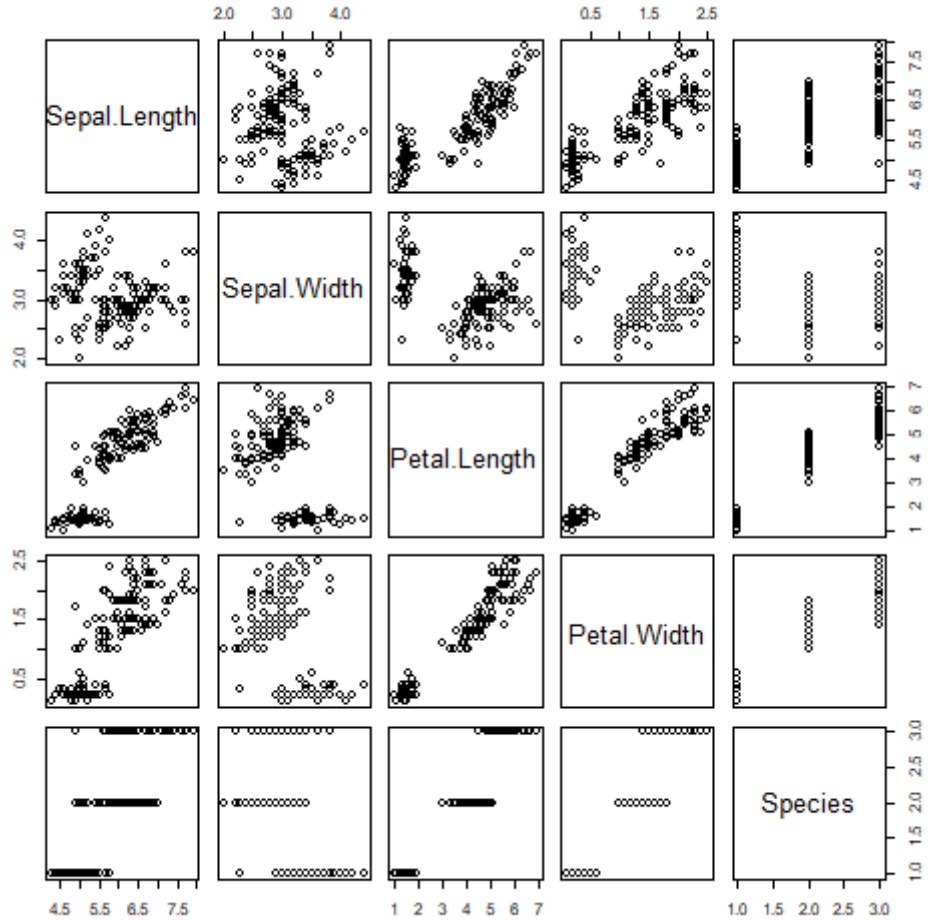
蓝色鸢尾花是拒绝，代表着绝望的爱情，无力支撑的情感关系。也预示着双方游离的情感和已经破碎的激情。

iris（鸢尾花）

明黄色和鲜黄色的鸢尾花适用于友人之间，代表着友谊长存、互相支持之意。

- 数据的散点图矩阵

```
pairs(iris)
```



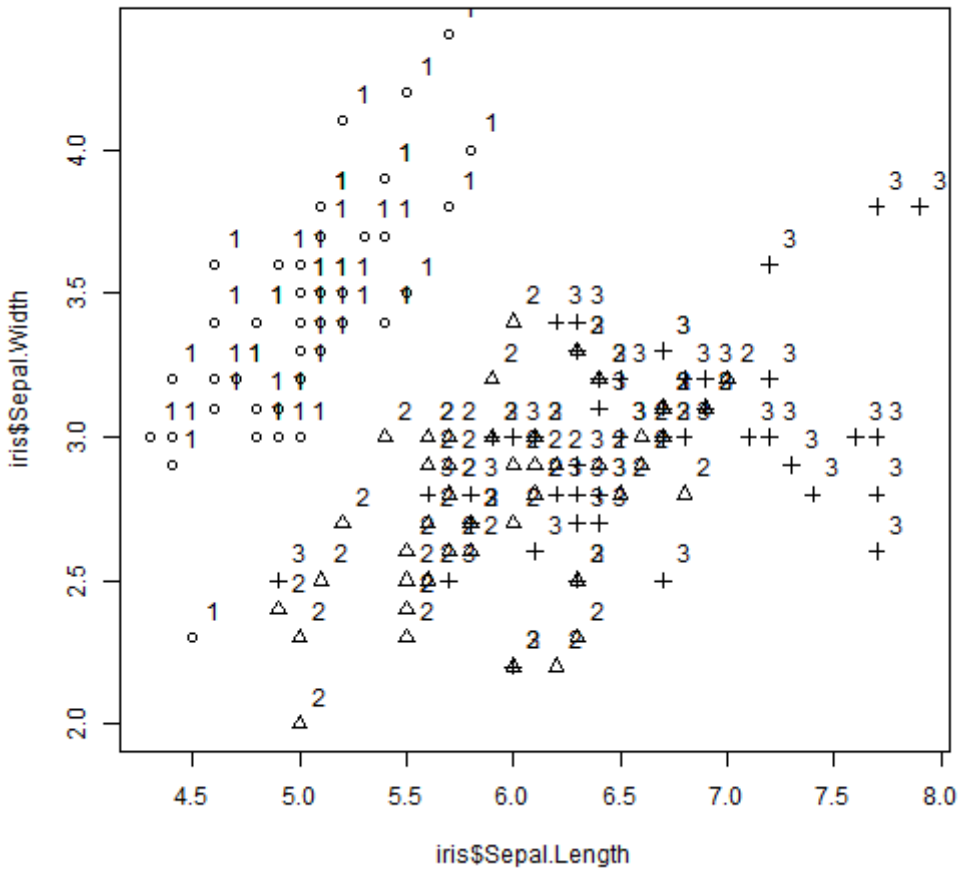
- 将花名替换为数值1，2，3

```
index1<-which(iris$Species=="setosa")
index2<-which(iris$Species=="versicolor")
index3<-which(iris$Species=="virginica")
type<-rep(0, nrow(iris))
type[index1]<-1
type[index2]<-2
type[index3]<-3
```

- 三种鸢尾花的Sepal(萼片)的长、宽

```
plot(iris$Sepal.Length, iris$Sepal.Width, pch=type)
text(x=iris$Sepal.Length+0.1, y=iris$Sepal.Width+0.1,labels=as.ch
```

```
plot(iris$Sepal.Length, iris$Sepal.Width, pch=type)
text(x=iris$Sepal.Length+0.1, y=iris$Sepal.Width+0.1,labels=as.ch
```



天气数据

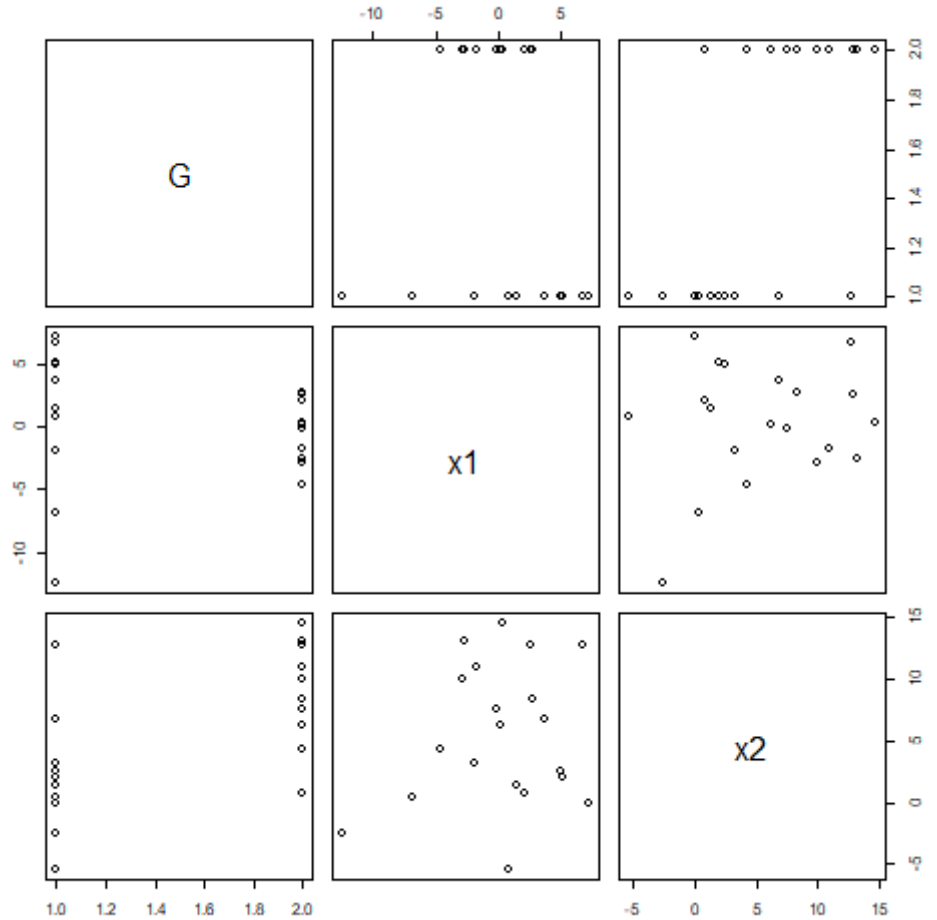
- 利用今天和昨天湿度差x1, 气温差x2, 预报今天x1=8.1, x2=2.0是否下雨
- 1 为下雨，2为晴天

```
(w3=read.csv("E:/teaching_plan_notes/msa11091083/rmd/MVAPureData/
```

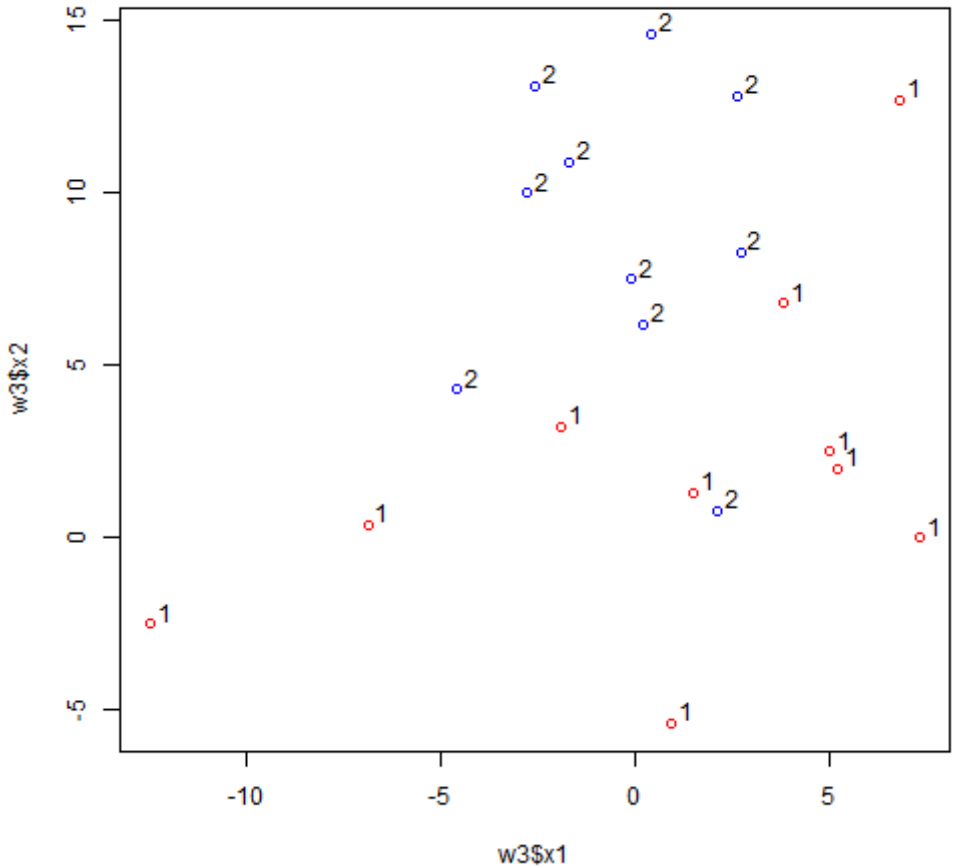
```
#    G    x1   x2
# 1  1  -1.9  3.2
# 2  1  -6.9  0.4
# 3  1   5.2  2.0
# 4  1   5.0  2.5
# 5  1   7.3  0.0
# 6  1   6.8 12.7
# 7  1   0.9 -5.4
# 8  1 -12.5 -2.5
# 9  1   1.5  1.3
# 10 1   3.8  6.8
# 11 2   0.2  6.2
# 12 2  -0.1  7.5
# 13 2   0.4 14.6
# 14 2   2.7  8.3
# 15 2   2.1  0.8
# 16 2  -4.6  4.3
# 17 2  -1.7 10.9
# 18 2  -2.6 13.1
# 19 2   2.6 12.8
```

- 两两散点图

```
pairs(w3)
```



```
plot(w3$x1, w3$x2, col=c("red", "blue")[w3$G])
text(w3$x1+0.4, w3$x2+0.4, labels=as.character(w3$G))
```



因变量为多分类变量

- 车牌号的识别
- 手写字的识别
- 喜怒哀乐的判别等
- 这些问题的因变量为多分类的，或者称为定性的，属性的

这些问题称为 判别分析

经典判别分析方法

- 距离判别
- 贝叶斯判别
- Fisher判别

机器学习方法

- 决策树
- 随机森林
- adaboost

蟹蟹

本幻灯片由 R 包 **xaringan** 生成；

查克拉来自于 remark.js、**knitr**、以及 R Markdown。