北京航空航天大學
经济管理学院
BEIHANG UNIVERSITY
SCHOOL OF ECONOMICS AND MANAGEMENT

第 6 讲：判别分析

康雁飞

2020-05-31

Outline

1. 为什么判别分析?
2. 距离判别法
3. Fisher 判别法
4. R 中进行判别分析
5. 作业

Outline

判别分析（Discriminant Analysis）的目的

- 已知某客观事物按照某种标准可分为 k 个总体 $G_1, G_2, \cdots, G_k$
- 根据已掌握的各个总体的样本信息，总结事物分类的规律
- 建立合理有效的判别规则

例如：

- 根据病人的诸项检验指标，进行疾病诊断
- 根据已有的气象资料来进行气象预报
- 根据心理测试问题，判断受试者的基本心理特征

例：根据个人信用资料，做违约风险评估

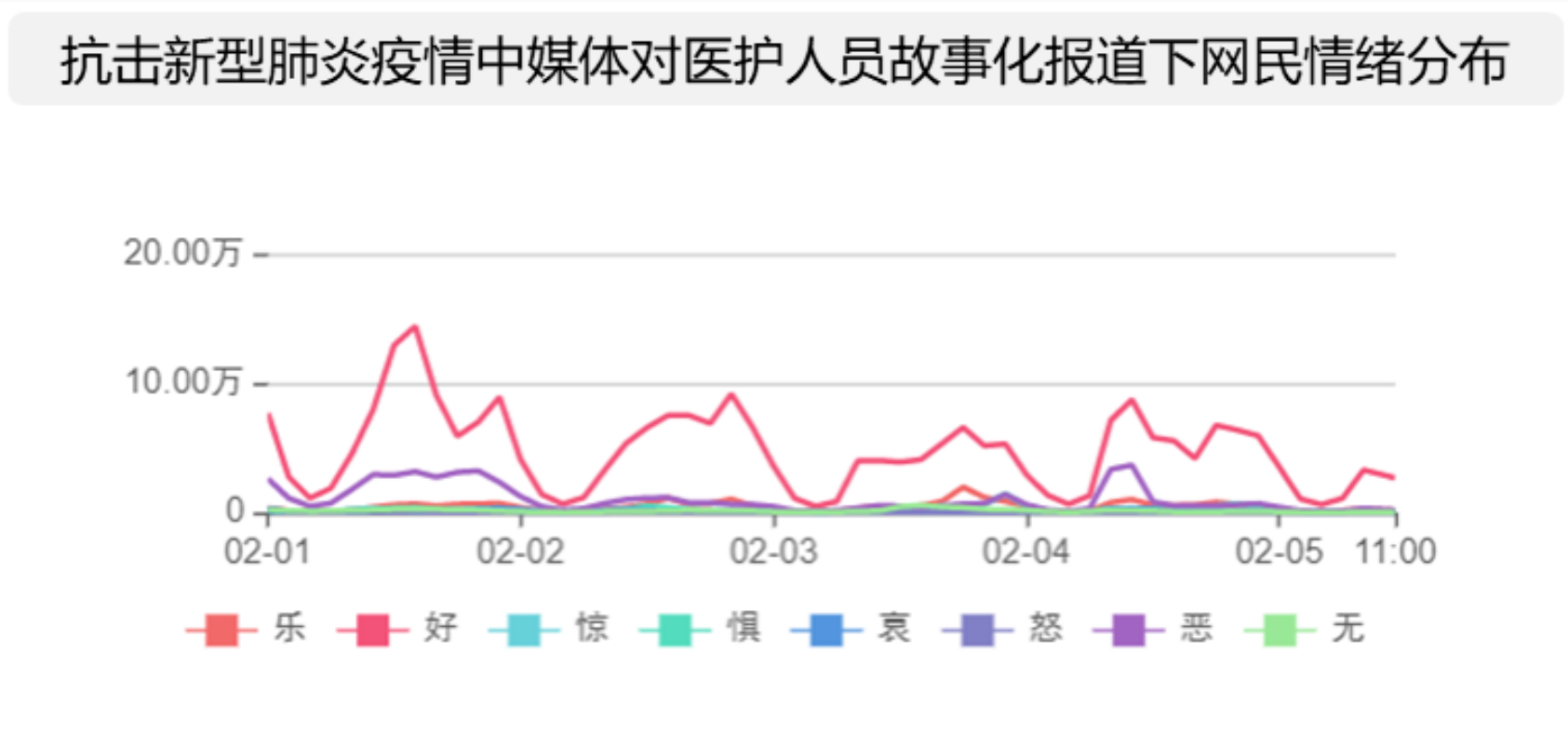
数据：

- 个人基本资料（性别、年龄、学历、婚姻）
- 实物资产（房、车），收入（工资、股票、配偶收入）
- 社交资产（微信好友个数、电话本好友数、QQ 好友数...）
- 贷款模式（贷款总额、还款年限、月付）
- 违约记录

目的：

- 识别：人群类型 + 贷款模式 $\Rightarrow$ 违约风险
- 向客户建议贷款总额和贷款模式（还款年限、月付）

例：舆情分析



来源：https://www.eefung.com/daily-report/20200205163459

例：垃圾邮件分类（Spam e-mail classification）

- 来源：ewlett-Packard Labs
- 4601 封邮件，57 个变量
- 判别：每封邮件是否为垃圾邮件（spam or non-spam）

数据样例

```
##      charSemicolon charRoundbracket charSquarebracket charExclamation
## 4430         0.000             0.000                 0           0.000
## 1211         0.000             0.000                 0           0.327
## 917          0.000             0.000                 0           0.168
## 3730         0.073             0.048                 0           0.024
## 304          0.000             0.000                 0           0.551
## 2298         0.000             0.000                 0           0.000
## 1178         0.000             0.000                 0           0.503
## 2517         0.000             0.293                 0           0.000
## 2188         0.000             0.000                 0           0.000
## 1940         0.000             0.214                 0           0.214
##      charDollar charHash capitalAve capitalLong capitalTotal    type
## 4430      0.000    0.000      1.000           1           16 nonspam
## 1211      1.357    0.046      5.769          72          450    spam
## 917       0.336    0.000      4.576          17          119    spam
## 3730      0.000    0.000      5.150          82          582 nonspam
## 304       0.459    0.000      2.333          22          119    spam
## 2298      0.000    0.000      1.666           7           25 nonspam
## 1178      0.062    0.000      1.820          12           91    spam
## 2517      0.000    0.000      3.968          34          127 nonspam
```

判别分析

问题：判别分析的输入是什么？

1. $X_{n\times p}$ 分为 K 类：$G_1, \cdots, G_K$。
2. 每行样本都有类标签（有监督学习，supervised learning)。
3. 假设第 k 个类的均值为 $\mu_k = \frac{1}{n_k} \sum_{x\in G_k} x$，方差协方差矩阵为 Σ_k。

Outline

距离判别法

离哪个类的距离最近，就属于哪一类。

距离如何定义？马氏距离

$$d^2(x, G_k) = (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)$$

两个类别情形

可根据以下判别函数建立判别规则：

$$W(x) = d^2(x, G_1) - d^2(x, G_2).$$

如何建立判别规则?

只有一个判别变量（$p = 1$）

假设两个类别总体方差相等，$\sigma_1^2 = \sigma_2^2 = \sigma^2$，用 s^2 估计。

我们可以写出判别函数：

$$W(x) = \frac{(x - \mu_1)^2 - (x - \mu_2)^2}{s^2}.$$

当 $\mu_1 > \mu_2$ 时，判断规则为：

$$x \in \begin{cases} G_1, & x > \frac{\mu_1 + \mu_2}{2} \\ G_2, & x < \frac{\mu_1 + \mu_2}{2} \\ \text{Unknown}, & x = \frac{\mu_1 + \mu_2}{2} \end{cases}$$

只有一个判别变量（$p = 1$）

假设两个类别总体方差不等，$\sigma_1^2 \neq \sigma_2^2$，分别用 s_1^2 和 s_2^2 估计。判别阈值为：

$$\mu^* = \frac{s_1\mu_2 + s_2\mu_1}{s_1 + s_2}$$

当 $\mu_1 > \mu_2$ 时，判断规则为：

$$x \in \begin{cases} G_1, & x > \mu^* \\ G_2, & x < \mu^* \\ \text{Unknown}, & x = \mu^* \end{cases}$$

多个总体的情形（更一般的情况）

- 假设有 K 个总体：$G_1, \cdots, G_K$。
- 第 k 个类的均值为 $\mu_k = \frac{1}{n_k} \sum_{x \in G_k} x$，方差协方差矩阵为 Σ_k。

问题：$\forall x \in R^p$，x 属于哪一类?

- 计算 $d^2(x, G_k)$，求最小值。
- $\arg\min_k d^2(x, G_k)$。

多个总体的情形

当 $\Sigma_1 = \cdots = \Sigma_K = \Sigma$ 时，

$$
\begin{aligned}
d^2(x, G_k) =& (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \\
=& x^T \Sigma^{-1} x - 2[x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k].
\end{aligned}
$$

多个总体的情形

定义一个线性函数 $f_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k$, 那么

$$
\begin{aligned}
d^2(x, G_1) =& x^T\Sigma^{-1}x - 2f_1(x) \\
d^2(x, G_2) =& x^T\Sigma^{-1}x - 2f_2(x) \\
&\vdots \\
d^2(x, G_K) =& x^T\Sigma^{-1}x - 2f_K(x)
\end{aligned}
$$

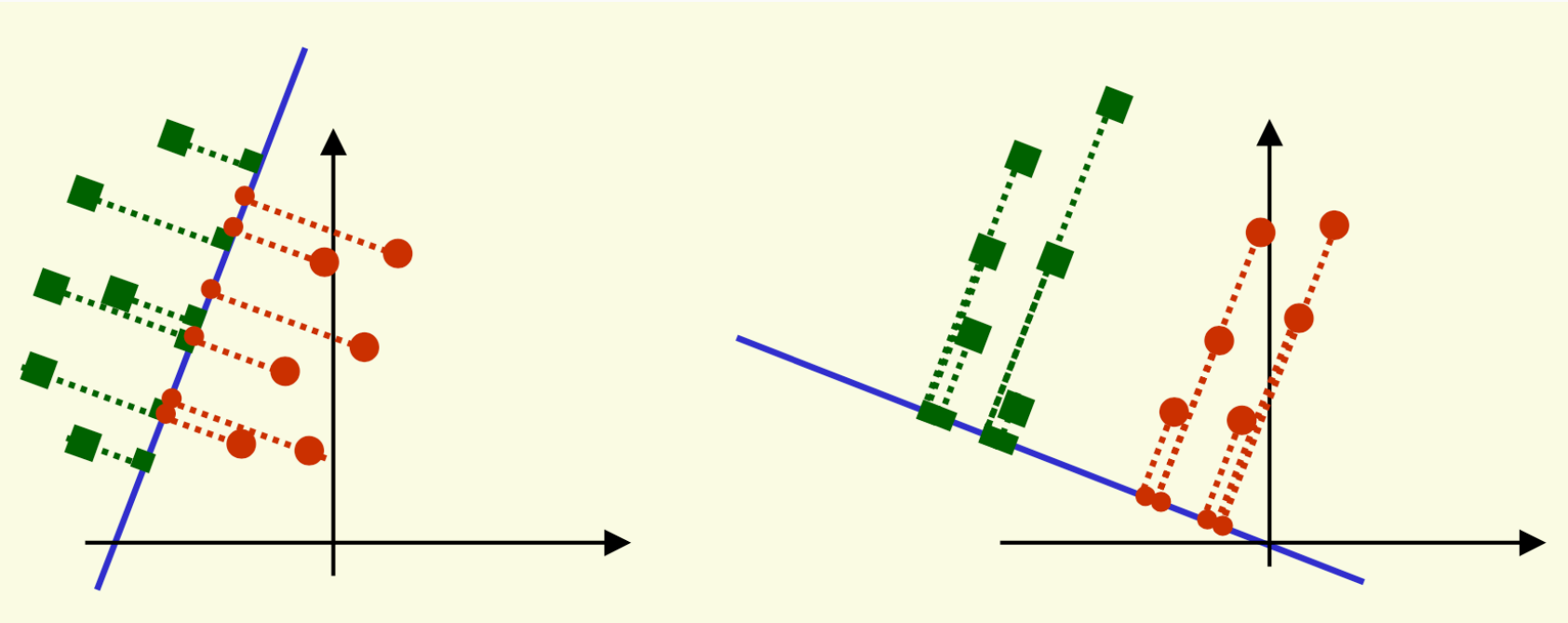
$$
x \in G_i \Leftrightarrow d^2(x, G_i) = \min_{k=1,\cdots,K}\{d^2(x, G_k)\}
$$

$$
\Leftrightarrow f_i(x) = \max_{k=1,\cdots,K}\{f_k(x)\}.
$$

Outline

Fisher 判别法

- PCA：通过平移 + 旋转省去数据变异不大方向的信息。
- Fisher 判别法：将 K 组的 p 维数据投影在某一个方向，对投影点来说能使得组与组之间尽可能地分开。



Fisher 判别法

- 假设有 K 个总体：$G_1, \cdots, G_K$。
- 第 k 个类的均值为 μ_k，方差协方差矩阵为 Σ_k。
- $w^T x \in R^1$ 为 x 在 w 方向上的投影。
- 寻找方向 $w \in R^p$，使得类与类之间的分辨度尽可能大，而类内的点尽可能聚合。

Fisher 判别法

- 组间离差

$$\begin{aligned}\tilde{S}_B &= \sum_{k=1}^{K} n_k (w^T \mu_k - w^T \mu)^2 \\ &= w^T [\sum_{k=1}^{K} n_k (\mu_k - \mu)(\mu_k - \mu)^T] w \\ &= w^T S_B w.\end{aligned}$$

- 组内离差

$$\begin{aligned}\tilde{S}_W &= \sum_{k=1}^{K} \sum_{x \in G_k} (w^T x - w^T \mu_k)^2 \\ &= w^T [\sum_{k=1}^{K} \sum_{x \in G_k} (x - \mu_k)(x - \mu_k)^T] w \\ &= w^T S_W w\end{aligned}$$

Fisher 判别法

Fisher 判别法的思想是最大化：

$$J(w) = \frac{w^T S_B w}{w^T S_W w}.$$

求导得到：

$$S_W^{-1} S_B w = \lambda w.$$

将 $S_W^{-1} S_B$ 特征值排序 $\lambda_1 \geq \cdots \geq \lambda_p$，最大的特征值对应的特征向量记为 w_1，记为判别效率最高的方向。

判别效率

判别效率：

$$\frac{\lambda_1}{\sum_{i=1}^{p} \lambda_i}.$$

若使用一维判别函数判别效率太低，可采用 m 个特征向量，然后按 $p > 1$ 的情形使用距离判别法，对应的判别效率为：

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i}.$$

Outline

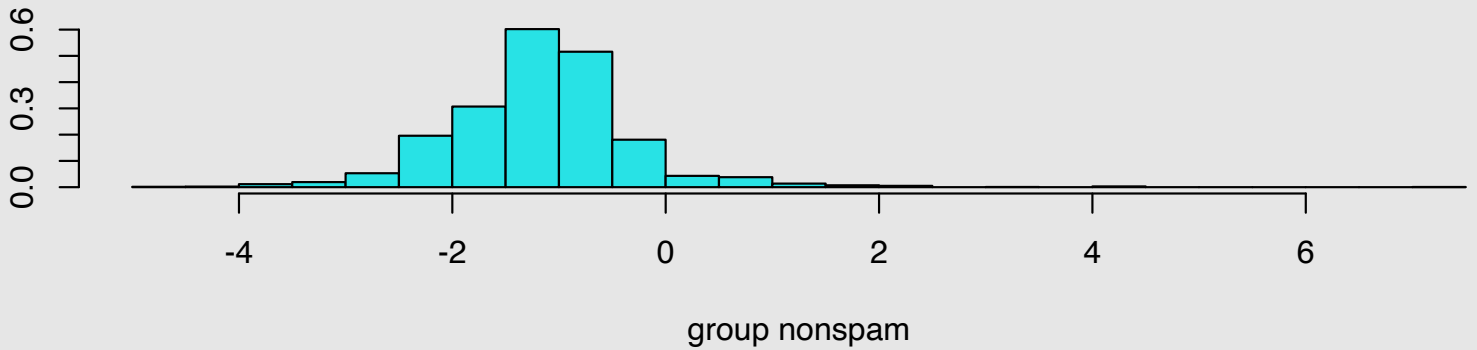
R 中进行判别分析

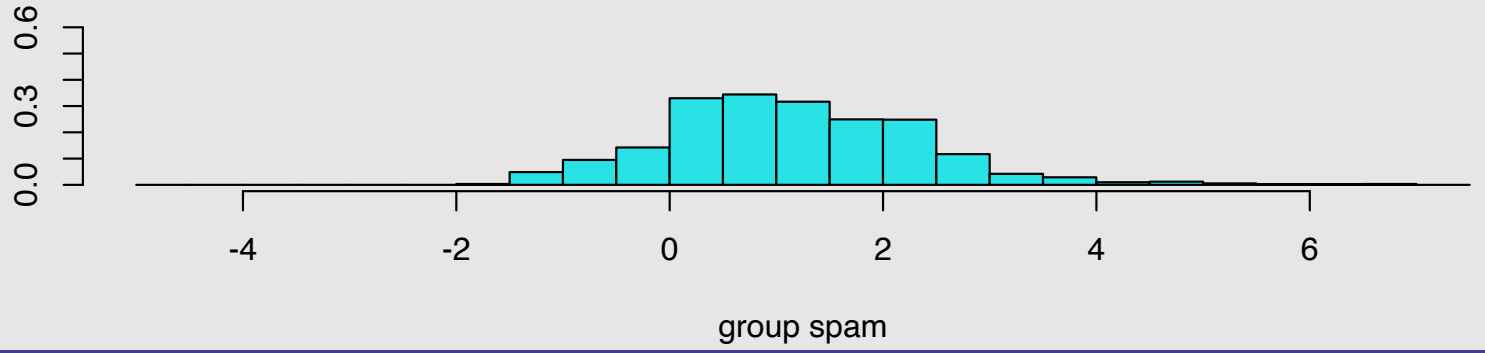
```
library(MASS)
spam.lda <- lda(type~., data = spam)
table(predict(spam.lda)$class, spam$type)
```

```
##
##               nonspam spam
##    nonspam       2663  387
##    spam           125 1426
```

R 中进行判别分析

```
plot(spam.lda)
```





其他常用的判别模型

1. 贝叶斯判别模型
2. k 近邻
3. 支持向量机
4. 决策树
5. 随机森林
6. 等等

Outline

利用 Fisher 判别法使用 `spam` 的前 4101 样本点进行建模分析，并用 `predict()` 对后 500 个样本点进行预测，然后请给出分析报告。