Title

subtitle

Name

Email: Name@tju.edu.cn

School or College,Tianjin University(Peiyang University)

June 1, 2016

Contents

Contents

Concept

- Discriminant analysis is a set of methods and tools used to distinguish between groups of populations $\prod_j$ and to determine how to allocate new observations into groups.
判别分析是在已知样品分类的前提下，将给定的新样品按照某种分类准则判入某个类中，它是研究如何将个体"归类"的一种统计分析方法。

Distance

- 距离是判别分析中的基本概念，距离判别法根据一个样品与各个类别距离的远近对该样品所属类别进行判定。常用的距离有欧式距离和马氏距离。

Distance

- 距离是判别分析中的基本概念，距离判别法根据一个样品与各个类别距离的远近对该样品所属类别进行判定。常用的距离有欧式距离和马氏距离。
-

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})}$$

Distance

- 距离类型的选择：
- 在判别分析中，马氏距离更常用。例如，$\mathbf{x}$到G_2的欧式距离小于到G_1的，但是x包含在总体G_1的置信域内，则把$\mathbf{x}$判入总体G_1。

两总体判别的判别函数为：

$$\mathbf{W}^*(\mathbf{x}) = d^2(\mathbf{x}, \mu_2) - d^2(\mathbf{x}, \mu_1) \tag{1.1}$$

Bayes

- Bayes判别对多个总体的判别考虑的不只是建立判别式，还要计算新样品属于各总体的条件概率$p(j/x), j = 1, 2, \cdots k$。比较这$k$个概率的大小，然后将新样品判归为来自概率最大的总体。Bayes判别准则是以个体归属于某类的概率最大或者错判总平均损失最小为标准的。

设有k个总体，各自分布密度函数为$f_j(\mathbf{x})$，判别准则为：

$$R_i = \{\mathbf{x} | p_i f_i(\mathbf{x}) = \max_{1 \le j \le k} p_j f_j(\mathbf{x})\} i = 1, 2, \ldots, k \qquad (1.2)$$

Fisher

- Fisher判别准则要求各类别之间的变异尽可能地大，而各类内部的变异尽可能地小，变异用离均差平方和表示。

$$\mu_y = \frac{1}{2}(\mu_{1y} + \mu_{2y}) = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \tag{1.3}$$

记

$$W(\mathbf{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}\mathbf{x} - \mu_y \tag{1.4}$$

则判别函数等价于

$$\begin{cases} x \in G_1 & \text{if } W(\mathbf{x}) \geq 0 \\ x \in G_2 & \text{if } W(\mathbf{x}) < 0 \end{cases}$$

Contents

Question

- 题目8.1：根据经验，今天与昨天的湿度差X_1及今天的压温差(气压与温度之差)X_2是预报明天下雨或不下雨的两个重要因素。现有一批数据资料，今测得$x_1 = 8.1$，$x_2 = 2.0$，试问预报明天下雨还是预报明天不下雨？分别用距离判别、Bayes判别和Fisher判别来得到所需要的结论。

Data

Table 1: 湿度差和压温差表

雨天		雨天	
X1(湿度差)	X2(压温差)	X1(湿度差)	X2(压温差)
-1.9	3.2	0.2	0.2
-6.9	10.4	-0.1	7.5
5.2	2	0.4	14.6
5	2.5	2.7	8.3
7.3	0	2.1	0.8
6.8	12.7	-4.6	4.3
0.9	-15.4	-1.7	10.9
-12.5	-2.5	-2.6	13.1
1.5	1.3	2.6	12.8
3.8	6.8	-2.8	10

Program

- (1)读入数据表，解析变量，绘制散点图。
  $X = read.table("weatherdata.txt", header = T)$
  $attach(X)$ #解析变量
  $plot(x1, x2);$
  $text(x1, x2, G, adj = -0.8)$

Result

- 从散点图上可以看出，下雨的天数集中在湿度差较小的日子； 可见两类错判的各有1例，判对的共有18例，故判别符合率为18/20=90.0%。线性判别函数
$y = -0.07263524x_1 + 0.14811860x_2$。

Supplement

- SVM & RF
- Bioconductor
- R & Parallel Computing
- Sth.

SVM

支持向量机(Support Vector Machines,SVM)是一种线性和非线性数据分析的分类方法，如线性SVM，非线性SVM，C-SVM(软间隔支持向量机)，ν -SVM，SVR。

Definition 1

考虑训练集$TR = \{(\mathbf{x}_i, y_i), i = 1, \ldots, l\}$，其中$\mathbf{x}_i \in \mathbb{R}^d, y_i \in \{+1, -1\}$。如果存在$\mathbf{w} \in \mathbb{R}^d$和正数$\varepsilon$，使得对所有正类的下标$i$，有$\mathbf{w} \cdot x_i + b \geq \varepsilon$，而对所有负类下标$i$，有$\mathbf{w} \cdot x_i + b \leq -\varepsilon$，则称训练集$TR$ 线性可分。

SVM

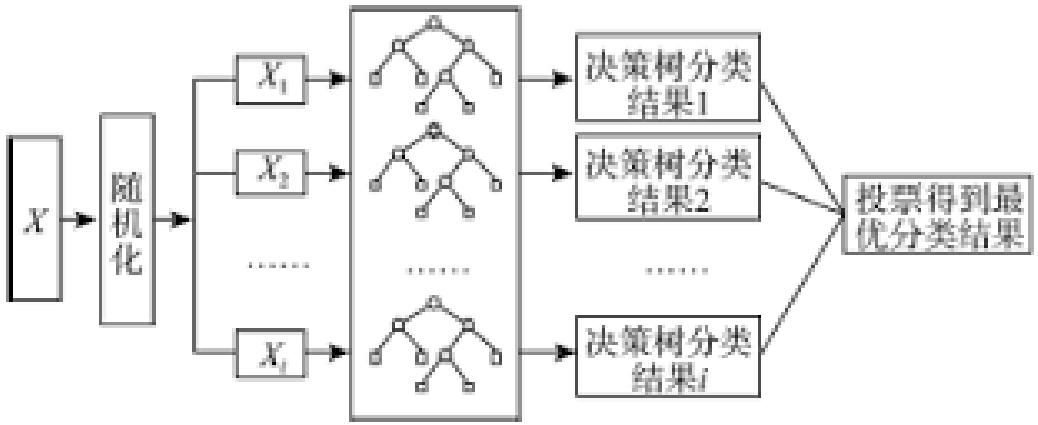
在近似线性可分的情形下，线性分类支持向量机的决策函数可描述为：

$$f(\mathbf{x}) = sgn(\mathbf{w}^* \cdot \mathbf{x} + b^*) = sgn(\sum_{i=1}^{l} \alpha_i^* y_i \mathbf{x}_i \cdot \mathbf{x} + b^*) \qquad (3.1)$$

SVM

- k-折交叉验证(k-fold cross validation)中，首先将数据集大致均分为k个子集，比如k=5,10,20 等。每次将其中一个子集作为测试集，其他的(k-1)个子集合作为训练集，轮流进行。

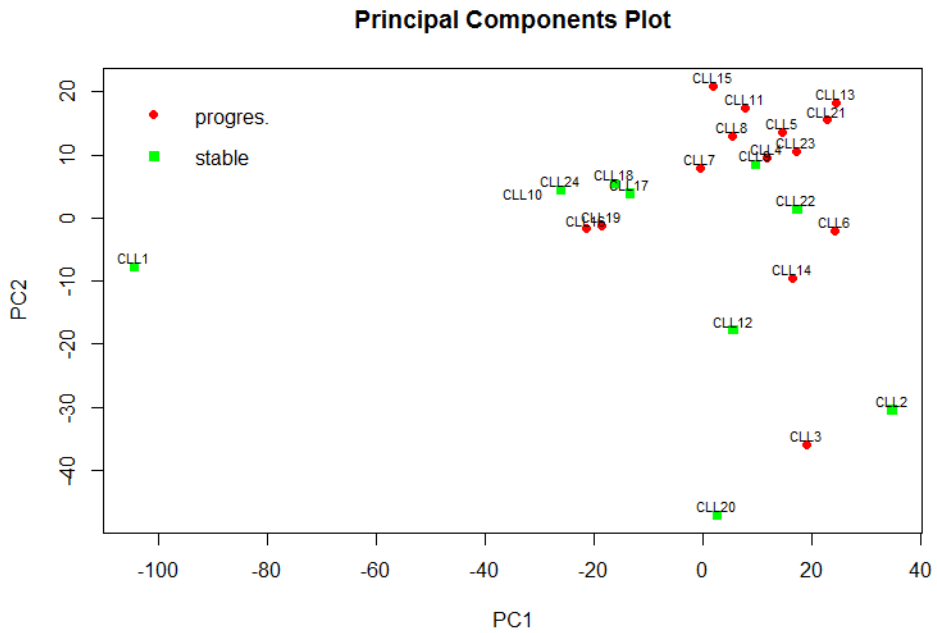
SVM与随机森林(RF)的比较



- BREIMAN L. Random Forests[J]. Machine Learning, 2001, 45: 5-32.
- 黄衍，查伟雄. 随机森林与支持向量机分类性能比较[J].软件，2012年，第33卷，第6期

Bioconductor

Bioconductor用结构化的数据来表示生物概念。以慢性淋巴细胞白血病为例，使用聚类方法和主成分分析方法分类观察结果。根据由相关系数矩阵导出的距离矩阵进行聚类。稳定组和恶化分组没能很好分开，这是因为如果总体上两组数据是分开的，那么说明是我们关心的因素起主导作用；如果不是，则可能有其他的主导作用。



R & parallel computing

- 1 R是运行时解释的(interpreted when begin to run)
  2 R是单线程的(single thread)
  3 R需要将全部数据加载到内存(whole data onto memory)
  4 算法设计影响时间和空间复杂度(algorithm)

R & parallel computing

- 1 R是运行时解释的(interpreted when begin to run)
  2 R是单线程的(single thread)
  3 R需要将全部数据加载到内存(whole data onto memory)
  4 算法设计影响时间和空间复杂度(algorithm)
- 可使用计算机集群执行多个任务(solving:cluster)
- *R High Performance Programming*

RStudio

ggplot2

R主要支持四套图形系统：基础图形(base)、网格图形(grid)、lattice图形和ggplot2。2005年出现ggplot2，使用grammer of graphics的语言来描述如何画图。以下是图形语法中的各个成分：

Data	Mapping
Geometric Object	Aesthitic Property
Scale	Coordiante
Statistical Trasfromation	Facet
Positional adjustment	etc.

ggplot2:Elegant Graphics for Data Analysis。

R的编程标准：https://google-style.googlecode.com/svn/trunk/google-r-style.html

Template

LaTeX 2_ε template, such as TJU Thesis, TJU Beamer, etc.

- https://www.github.com/6gbluewind

Reference

- 多元统计分析-基于R 主编费宇，北京，中国人民大学出版社，2014
- 多元数据分析及其R实现，肖枝洪[等]编著，北京，科学出版社，2013
- R in Nutshell,2nd Edition,Joseph Adler,2014.7

Q&A

? & !

Acknowledgement