Advances in Applied Mathematics 应用数学进展, 2022, 11(10), 7097-7108
Published Online October 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.1110753

Hans汉斯

基于判别分析的葡萄酒类型分析

申芮洁

云南财经大学，云南 昆明

收稿日期：2022年9月14日；录用日期：2022年10月5日；发布日期：2022年10月14日

摘　要

基于数据集，使用距离判别，贝叶斯判别和费希尔判别这三种判别分析方法对每个样本来自哪个红酒厂进行判别，与原始数据中所属酒厂进行比较，并用三种非参数估计方法(回代法，划分样本和交叉验证法)计算估计的误判率，比较三种非参数估计方法的优劣，对其他未知产地的红酒判别来自三个红酒厂中的哪一个提供便利。结合三种非参数估计方法的误判率分析得出，一般情况下交叉验证法较其他两种方法效果较好，最值得推荐。

关键词

判别分析，距离判别，贝叶斯判别，费希尔判别，非参数估计方法

Wine Type Analysis Based on Discriminant Analysis

Ruijie Shen

Yunnan University of Finance and Economics, Kunming Yunnan

Received: Sep. 14th, 2022; accepted: Oct. 5th, 2022; published: Oct. 14th, 2022

Abstract

Based on the data set, three discriminant analysis methods, namely distance discrimination, Bayesian discrimination and Fisher discrimination, are used to distinguish which wine factory each sample comes from, compare it with the wine factory in the original data, and calculate the error rate of the estimation with three nonparametric estimation methods (back substitution method, sample division and cross validation method) to compare the advantages and disadvantages of the three nonparametric estimation methods. It is convenient to identify which of the three wineries is the source of red wine from other unknown origins. Combined with the error rate analysis of three nonparametric estimation methods, the cross validation method is generally better than the other two methods, and is most recommended.

文章引用: 申芮洁. 基于判别分析的葡萄酒类型分析[J]. 应用数学进展, 2022, 11(10): 7097-7108.
DOI: 10.12677/aam.2022.1110753

Keywords

Discriminant Analysis, Distance Discriminant, Bayesian Discriminant, Fisher Discriminant, Nonparametric Estimation Method

1. 引言

无论是科研还是生活，我们经常面临根据观察到的数据来区分对象(或样本)的问题。为了有效地区分样本，我们通常依赖于许多监测指标，因此适合在多变量分析中讨论差异分析。在判别分析中，根据样品的 p 维指标值 $x=\left(x_1,x_2,\cdots,x_p\right)'$ 对其的组别归属进行判别，通过图形(通常为二维)方法或代数方法来记述每个样本群的差异，研究了最大限度的分离。

本文首先阐述了距离判别，贝叶斯判别和费希尔判别三种判别分析方法的基本理论，用这些模型分别对数据集建立模型，判别每个葡萄酒样品来自哪个红酒厂，并用非参数估计方法估计误判概率，比较三种非参数估计方法的好坏。

2. 研究目的和意义

使用不一样的判别分析法(距离判别，贝叶斯判别和费希尔判别)对每个样品来自哪个葡萄酒厂进行判别，并与原始数据中所属酒厂进行比较，并用三种非参数估计方法(回代法，划分样本和交叉验证法)计算估计的误判率，比较三种非参数估计方法的优劣，对其他未知产地的葡萄酒新样品判别来自三个葡萄酒厂中的哪一个提供便利。

3. 数据介绍

此数据来自意大利三个不同地区的葡萄酒厂，从三个红酒厂依次取了 59，71 和 48 个共 178 个样本，并对每个样本中的 13 种化学成分进行检测(分别为“酒精”，“灰的碱性”，“苹果酸”，“镁”，“灰”，“总酚”，“类黄酮”，“花青素”，“非黄烷类酚类”，“od280/od315 稀疏葡萄酒”，“色调”，“颜色强度”)。

4. 研究方法概述

4.1. 距离判别

设有 k 个组 $\pi_1,\pi_2,\cdots,\pi_k$，其均值分别是 $\mu_1,\mu_2,\cdots,\mu_k$，协方差矩阵分别是 $\Sigma_1(>0),\Sigma_2(>0),\cdots,\Sigma_k(>0)$，$x$ 到总体 π_i 的平方马氏距离为

$$d^2(\boldsymbol{x},\pi_i)=(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i),\quad i=1,2,\cdots,k \tag{1-1}$$

判别规则为

$$\boldsymbol{x}\in\pi_l,\ \text{若}\ d^2(\boldsymbol{x},\pi_l)=\min_{1\le i\le k}d^2(\boldsymbol{x},\pi_i) \tag{1-2}$$

4.2. 贝叶斯判别

最大后验概率法

设有 k 个组 $\pi_1,\pi_2,\cdots,\pi_k$，且组 π_i 的概率密度为 $f_i(x)$，样品 x 来自组 π_i 的先验概率为 $p_i, i=1,2,\cdots,k$，满足 $p_1+p_2+\cdots+p_k=1$。则 x 属于 π_i 的后验概率为

$$P(\pi_i \mid \boldsymbol{x})=\frac{p_i f_i(\boldsymbol{x})}{\sum_{j=1}^{k} p_j f_j(\boldsymbol{x})},\quad i=1,2,\cdots,k \tag{2-1}$$

最大后验概率法采用的判别规则如下：

$$\boldsymbol{x}\in\pi_l,\ \text{若}\ P(\pi_l \mid \boldsymbol{x})=\max_{1\le i\le k} P(\pi_i \mid \boldsymbol{x}) \tag{2-2}$$

当 k 组都是正态的，即 $\pi_i \sim Np(\mu_i,\Sigma_i)$，$\Sigma_i>0$，$i=1,2,\cdots,k$。这时，组 π_i 的概率密度为

$$f_i(x)=(2\pi)^{-p/2}|\Sigma_i|^{-1/2}\exp\left[-0.5d^2(x,\pi_i)\right] \tag{2-3}$$

其中

$$d^2(x,\pi_i)=(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i) \tag{2-4}$$

是 x 到 π_i 的平方马氏距离。

当 $p_1=p_2=\cdots=p_k=1/k$，$\Sigma_1=\Sigma_2=\cdots=\Sigma_k=\Sigma$ 时，

$$P(\pi_i \mid \boldsymbol{x})=\frac{\exp\left[-\frac{1}{2}d^2(\boldsymbol{x},\pi_i)\right]}{\sum_{j=1}^{k}\exp\left[-\frac{1}{2}d^2(\boldsymbol{x},\pi_j)\right]}$$

当 $p_1=p_2=\cdots=p_k=1/k$，而 $\Sigma_1,\Sigma_2,\cdots,\Sigma_k$ 不全相等时，

$$P(\pi_i \mid \boldsymbol{x})=\frac{\exp\left\{-\frac{1}{2}\left[d^2(\boldsymbol{x},\pi_i)+\ln|\boldsymbol{\Sigma}_i|\right]\right\}}{\sum_{j=1}^{k}\exp\left\{-\frac{1}{2}\left[d^2(\boldsymbol{x},\pi_j)+\ln|\boldsymbol{\Sigma}_j|\right]\right\}}$$

当 $\Sigma_1=\Sigma_2=\cdots=\Sigma_k=\Sigma$，而 $p_1,p_2,\cdots,p_k$ 不全相等时，

$$P(\pi_i \mid \boldsymbol{x})=\frac{\exp\left\{-\frac{1}{2}\left[d^2(\boldsymbol{x},\pi_i)-2\ln p_i\right]\right\}}{\sum_{j=1}^{k}\exp\left\{-\frac{1}{2}\left[d^2(\boldsymbol{x},\pi_j)-2\ln p_j\right]\right\}}$$

当 $p_1,p_2,\cdots,p_k$ 不全相等，$\Sigma_1,\Sigma_2,\cdots,\Sigma_k$ 也不全相等时，

$$P(\pi_i \mid \boldsymbol{x})=\frac{\exp\left\{-\frac{1}{2}\left[d^2(\boldsymbol{x},\pi_i)+\ln|\boldsymbol{\Sigma}_i|-2\ln p_i\right]\right\}}{\sum_{j=1}^{k}\exp\left\{-\frac{1}{2}\left[d^2(\boldsymbol{x},\pi_j)+\ln|\boldsymbol{\Sigma}_j|-2\ln p_j\right]\right\}}$$

上述各情形的后验概率可统一表达为

$$P(\pi_i \mid \boldsymbol{x})=\frac{\exp\left[-\frac{1}{2}D^2(\boldsymbol{x},\pi_i)\right]}{\sum_{j=1}^{k}\exp\left[-\frac{1}{2}D^2(\boldsymbol{x},\pi_j)\right]},\quad i=1,2,\cdots,k \tag{2-5}$$

其中，

$$D^2(x,\pi_i)=d^2(x,\pi_i)+g_i+h_i$$

$$g_i=\begin{cases}\ln|\boldsymbol{\Sigma}_i|, & \text{若}\ \boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\cdots,\boldsymbol{\Sigma}_k\ \text{不全相等}\\ 0, & \text{若}\ \boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\cdots=\boldsymbol{\Sigma}_k=\boldsymbol{\Sigma}\end{cases}$$

$$h_i=\begin{cases}-2\ln p_i, & \text{若}\ p_1,p_2,\cdots,p_k\ \text{不全相等，}\ i=1,2,\cdots,k\\ 0, & \text{若}\ p_1=p_2=\cdots=p_k=\dfrac{1}{k}\end{cases}\tag{2-6}$$

称 $D_2(x,\pi_i)$为 x 到 π_i 的广义平方距离。在正态性假定下，上述判别规则可表达如下：

$$\boldsymbol{x}\in\pi_l,\ \text{若}\ D^2(\boldsymbol{x},\pi_l)=\min_{1\le i\le k}D^2(\boldsymbol{x},\pi_i)\tag{2-7}$$

当 $\Sigma_1=\Sigma_2=\cdots=\Sigma_k=\Sigma$ 时，上述后验概率公式可简化为

$$P(\pi_i\mid\boldsymbol{x})=\frac{\exp(\boldsymbol{I}_i'\boldsymbol{x}+c_i+\ln p_i)}{\sum_{j=1}^{k}\exp(\boldsymbol{I}_j'\boldsymbol{x}+c_j+\ln p_j)},\quad i=1,2,\cdots,k\tag{2-8}$$

其中 $I_i=\Sigma^{-1}\mu_i, c_i=-\frac{1}{2}\mu_i'\Sigma^{-1}\mu_i, i=1,2,\cdots,k$。此时，判别规则等价于

$$\boldsymbol{x}\in\pi_l,\ \text{若}\ \boldsymbol{I}_l'\boldsymbol{x}+c_l+\ln p_l=\max_{1\le i\le k}(\boldsymbol{I}_i'\boldsymbol{x}+c_i+\ln p_i)\tag{2-9}$$

如果我们对 x 来自哪一组的先验信息一无所知或难以确定，则一般可取 $p_1=p_2=\cdots=p_k=1/k$。这时，判别规则简化为

$$\boldsymbol{x}\in\pi_l,\ \text{若}\ \boldsymbol{I}_l'\boldsymbol{x}+c_l=\max_{1\le i\le k}(\boldsymbol{I}_i'\boldsymbol{x}+c_i)\tag{2-10}$$

4.3. 费希尔判别

4.3.1. 费希尔判别的基本思想

费希尔判别(或称典型判别)的基本思想是投影(或降维)：用 p 维向量的少数几个线性组合(称为费希尔判别函数或典型变量) $y_1=\boldsymbol{a}_1'\boldsymbol{x},y_2=\boldsymbol{a}_2'\boldsymbol{x},\cdots,y_r=\boldsymbol{a}_r'\boldsymbol{x}$ (一般 r 明显小于 p)来代替原始的 p 个变量 $x_1,x_2,\cdots,x_p$，以达到降维的目的，并根据这 r 个判别函数 $y_1,y_2,\cdots,y_r$ 对样品的归属作出判别或将各组分离。

4.3.2. 费希尔判别函数

设来自组 π_i 的 p 维观测值为 x_{ij}，$j=1,2,\cdots,n_i$，$i=1,2,\cdots,k$，将它们共同投影到某一 p 维常数向量 a 上，得到的投影点可分别对应线性组合 $y_{ij}=a'x_{ij}$，$j=1,2,\cdots,n_i$，$i=1,2,\cdots,k$。[1]

$$\overline{y}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}y_{ij}=\boldsymbol{a}'\overline{\boldsymbol{x}}_i$$

$$\overline{y}=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}y_{ij}=\frac{1}{n}\sum_{i=1}^{k}n_i\overline{y}_i=\boldsymbol{a}'\overline{\boldsymbol{x}}$$

式中 $n=\sum_{i=1}^{k}n_i$，$\overline{\boldsymbol{x}}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}$，$\overline{\boldsymbol{x}}=\frac{1}{n}\sum_{i=1}^{k}n_i\overline{\boldsymbol{x}}_i$。

费希尔判别需假定 $\Sigma_1=\Sigma_2=\cdots=\Sigma_k=\Sigma$。

y_{ij}的组间平方和及组内平方和分别为

$$SSTR=\sum_{i=1}^{k}n_i\left(\overline{y}_i-\overline{y}\right)^2=\sum_{i=1}^{k}n_i\left(\boldsymbol{a}'\overline{\boldsymbol{x}}_i-\boldsymbol{a}'\overline{\boldsymbol{x}}\right)^2=\boldsymbol{a}'\boldsymbol{H}\boldsymbol{a}$$

$$SSE=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}_i\right)^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\boldsymbol{a}'\boldsymbol{x}_{ij}-\boldsymbol{a}'\overline{\boldsymbol{x}}_i\right)^2=\boldsymbol{a}'\boldsymbol{E}\boldsymbol{a}$$

式中

$$\boldsymbol{H}=\sum_{i=1}^{k}n_i\left(\overline{\boldsymbol{x}}_i-\overline{\boldsymbol{x}}\right)\left(\overline{\boldsymbol{x}}_i-\overline{\boldsymbol{x}}\right)'$$

$$\boldsymbol{E}=\sum_{i=1}^{k}\left(n_i-1\right)\boldsymbol{S}_i=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}}_i\right)\left(\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}}_i\right)'$$

可用来反映 y_{ij} 的组之间分离程度的一个量是

$$\Delta\left(\boldsymbol{a}\right)=\frac{SSTR}{SSE}=\frac{\boldsymbol{a}'\boldsymbol{H}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{E}\boldsymbol{a}} \tag{3-1}$$

在约束条件 $a'S_pa=1$ 下，寻找 a，使得 $\Delta(a)$达到最大，其中 $S_p=\frac{1}{n-k}E$，是 Σ 的联合无偏估计。

设 $E^{-1}H$ 的全部非零特征值依次为 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_s>0$，这里 $s=\mathrm{rank}(H)$，且有

$$s\leq\min\left(k-1,p\right) \tag{3-2}$$

相应的特征向量依次记为 $t_1,t_2,\cdots,t_s$ (标准化为 $t_i'S_pt_i=1,i=1,2,\cdots,s$)。

$\Delta(a_1)$的最大值 λ_1 在 $a_1=t_1$ 时达到。所以，选择投影到 t_1 上能使各组的投影点最大限度地分离，称 $y_1=t_1'x$ 为费希尔第一线性判别函数，简称第一判别函数。

在很多情况下(如 k 或 p 是大的)，只使用第一判别函数可能会不够，应考虑建立 $y_2=a_2'x$，且满足

$$\mathrm{Cov}\left(y_1,y_2\right)=\mathrm{Cov}\left(\boldsymbol{t}_1'\boldsymbol{x},\boldsymbol{a}_2'\boldsymbol{x}\right)=\boldsymbol{t}_1'\boldsymbol{\Sigma}\boldsymbol{a}_2=0$$

用 S_p 代替未知的 Σ，于是在约束条件 $t_1'S_pa_2=0$ (或 $t_1'Ea_2=0$)下寻找 a_2，使得 $\Delta(a_2)$达到最大。当 $a_2=t_2$ 时 $\Delta(a_2)$达到最大值 λ_2，$y_2=t_2'x$ 称为第二判别函数。一般情况下要求第 i 个线性组合 $y_i=a_i'x$ 与前 $i-1$ 个判别函数中的信息不重复，即

$$\mathrm{Cov}\left(y_j,y_i\right)=\mathrm{Cov}\left(\boldsymbol{t}_j'\boldsymbol{x},\boldsymbol{a}_i'\boldsymbol{x}\right)=\boldsymbol{t}_j'\boldsymbol{\Sigma}\boldsymbol{a}_i=0,\ \ j=1,2,\cdots,i-1$$

用 S_p 替代 Σ，上式变为

$$\boldsymbol{t}_j'\boldsymbol{S}_p\boldsymbol{a}_i=0\ \ \left(\text{或}\,\boldsymbol{t}_j'\boldsymbol{E}\boldsymbol{a}_i=0\right),\ \ \ j=1,2,\cdots,i-1$$

在上述约束条件下寻找 a_i，使得 $\Delta(a_i)$达到最大。$\Delta(a_i)$达到最大值 λ_i 时取 $a_i=t_i$，称 $y_i=t_i'x$ 为第 i 判别函数，$i=2,3,\cdots,s$。

$\Delta\left(t_i\right)=\lambda_i$ 表明了 y_i 对分离各组的贡献大小，y_i 在所有 s 个判别函数中的贡献率为

$$\lambda_i\Big/\sum_{j=1}^{s}\lambda_j \tag{3-3}$$

而前 $r(\leq s)$个判别函数 $y_1,y_2,\cdots,y_r$ 的累计贡献率为

$$\sum_{i=1}^{r}\lambda_i\Big/\sum_{i=1}^{s}\lambda_i \tag{3-4}$$

它表明了 $y_1,y_2,\cdots,y_r$ 的判别能力。

4.3.3. 判别规则

因为各判别函数都有单位方差并且彼此不相关，所以此时的马氏距离等同于欧氏距离。我们采用距离判别法，依据$(y_1, y_2, \cdots, y_r)$值，判别新样品归属离它最近的那一组。判别规则为

$$\boldsymbol{x} \in \pi_l，\ 若 \sum_{j=1}^{r}\left(y_j - \overline{y}_{lj}\right)^2 = \min_{1\le i\le k}\sum_{j=1}^{r}\left(y_j - \overline{y}_{ij}\right)^2 \tag{3-5}$$

其中 $\overline{y}_{ij} = t_j'\overline{x}_i$，$\overline{x}_i = \frac{1}{n_i}\sum_{j}^{n_i} x_{ij}$，$i = 1, 2, \cdots, k$。该判别规则也可表达为

$$\boldsymbol{x} \in \pi_l，\ 若 \sum_{j=1}^{r}\left[\boldsymbol{t}_j'\left(\boldsymbol{x} - \overline{\boldsymbol{x}}_l\right)\right]^2 = \min_{1\le i\le k}\sum_{j=1}^{r}\left[\boldsymbol{t}_j'\left(\boldsymbol{x} - \overline{\boldsymbol{x}}_i\right)\right]^2 \tag{3-6}$$

若在判别过程中只是用一种判别函数(即 $r = 1$)，则以上判别规则可简化为

$$\boldsymbol{x} \in \pi_l，\ 若 \left|y - \overline{y}_l\right| = \min_{1\le i\le k}\left|y - \overline{y}_i\right| \tag{3-7}$$

式中 y 和 $\overline{y}_i\left(i = 1, 2, \cdots, k\right)$ 分别是前面判别规则中的 y_1 和 $\overline{y}_{i1}\left(i = 1, 2, \cdots, k\right)$。

4.4. 误判概率的非参数估计

如果两组无法假设成正态组，可以用样本中样品的误判比例来估计 $P(2|1)$和 $P(1|2)$，一般情况下非参数估计方法有三种：

4.4.1. 回代法

令 $n(2|1)$为 π_1 的样本被误判为 π_2 的个数，$n(1|2)$为 π_1 的样本被误判为 π_2 的个数，那么 $P(2|1)$和 $P(1|2)$可估计为

$$\hat{P}\left(2|1\right) = \frac{n\left(2|1\right)}{n_1},\quad \hat{P}\left(1|2\right) = \frac{n\left(1|2\right)}{n_2} \tag{4-1}$$

4.4.2. 划分样本

将整个样本分为两部分，一部分是训练样本以构造判别函数，另一部分是验证样本以评估该判别函数。用验证样本的被误判比例来估计误判概率，其估计是无偏的。

4.4.3. 交叉验证法(或称刀切法)

从 π_1 组中提取 x_{1j}，判别函数通过 π_2 组的 n_2 个观测值和 π_1 组剩余的 $n_1 - 1$ 个观测值构造，然后对 x_{1j} 进行判别，$j = 1, 2, \cdots, n_1$。同理，从 π_2 组中取出 x_{2j}，判别函数通过 π_1 组的 n_1 个观测值和 π_2 组剩余的 $n_2 - 1$ 个观测值构造，再对 x_{2j} 作出判别，$j = 1, 2, \cdots, n_2$。

令

$n^*(2|1)$——π_1 的样本被误判为 π_2 的个数

$n^*(1|2)$——π_2 的样本被误判为 π_1 的个数

误判概率 $P(2|1)$和 $P(1|2)$的估计量分别为

$$\hat{P}\left(2|1\right) = \frac{n^*\left(2|1\right)}{n_1},\quad \hat{P}\left(1|2\right) = \frac{n^*\left(1|2\right)}{n_2} \tag{4-2}$$

它们都是接近无偏的估计量。

5. 实证分析

5.1. 各种判别分析方法的实证分析

5.1.1. 用 SPSS 进行数据的描述性分析

Table 1. Data descriptive analysis
表 1. 数据描述性分析

描述统计						
	个案数	最小值	最大值	平均值	标准差	方差
酒精	178	11.03	14.83	13.0006	0.81183	0.659
苹果酸	178	0.74	5.80	2.3363	1.11715	1.248
灰	178	1.36	3.23	2.3665	0.27434	0.075
灰的碱性	178	10.6	30.0	19.495	3.3396	11.153
镁	178	70	162	99.74	14.282	203.989
总酚	178	0.98	3.88	2.2951	0.62585	0.392
类黄酮	178	0.34	5.08	2.0293	0.99886	0.998
非黄烷类酚类	178	0.13	0.66	0.3619	0.12445	0.015
花青素	178	0.41	3.58	1.5909	0.57236	0.328
颜色强度	178	1.28	13.00	5.0581	2.31829	5.374
色调	178	0.480	1.710	0.95745	0.228572	0.052
od280od315 稀疏葡萄酒	178	1.27	4.00	2.6117	0.70999	0.504
脯氨酸	178	278	1680	746.89	314.907	99166.717
有效个案数(成列)	178					

从表 1 的数据描述性分析可知，所选数据无缺失判别变量。

5.1.2. K 最近邻算法分析

1) 基本思想

kNN 算法的核心思想是：如果一个数据在特征空间中最相邻的 k 个数据中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性。通俗地说，对于给定的测试样本和基于某种度量距离的方式，通过最靠近的 k 个训练样本来预测当前样本的分类结果。[2]

2) 实验结果分析

通过 R.编程实现我们得到如下结果：见图 1 所示。

```
> table(test$Type,fit$fitted.values) #混淆矩阵

   1  2  3
1 25  0  0
2  1 18  2
3  0  0 14
> mean(test$Type!=fit$fitted.values) #错误率
[1] 0.05
```

Figure 1. Operation results
图 1. 运行结果

编程过程中使用了划分样本法，178 个样品中选取 1/3 共 60 个作为测试集，其余 2/3 共 118 个作为训练集，判别 60 个葡萄酒测试样本所属类别，从图我们可以看出，其中有 1 个来自第二类误判为第一类，有 2 个来自第二类误判为第三类，误判率为 5.0%。

5.1.3. 贝叶斯判别分析

使用 SPSS 进行以下分析：

1) 如表 2 先验概率所示，我们对 x 来自哪一组的先验信息一无所知或难以确定，则一般可取 $p_1 = p_2 = p_3 = 1/3$。

Table 2. Prior probability
表 2. 先验概率

组的先验概率			
葡萄酒类型	先验	在分析中使用的个案	
		未加权	加权
1	0.333	59	59.000
2	0.333	71	71.000
3	0.333	48	48.000
总计	1.000	178	178.000

2) 通过表 3 分类函数系数写出判别函数，计算每个样品分别落入 3 个葡萄酒类型的概率，概率最大就被归为一类。

Table 3. Classification function coefficient
表 3. 分类函数系数

分类函数系数			
	葡萄酒类型		
	1	2	3
酒精	57.351	52.373	54.127
苹果酸	0.854	0.134	2.099
灰	39.031	28.029	35.906
灰的碱性	−0.662	0.465	0.554
镁	0.502	0.496	0.485
总酚	−3.261	−1.061	1.531
类黄酮	3.579	0.075	-9.235
非黄烷类酚类	39.626	41.418	28.223
花青素	1.243	2.970	2.317
颜色强度	−3.988	−3.856	−1.266
色调	27.600	31.177	21.434
od280od315 稀疏葡萄酒	22.527	18.445	13.554
脯氨酸	0.021	0.000	0.000
(常量)	−523.438	−427.379	−453.571

判别函数为

$$y_1 = 57.351x_1 + 0.854x_2 + 39.031x_3 - 0.662x_4 + 0.502x_5 - 3.261x_6 + 3.579x_7 + 39.626x_8 + 1.243x_9 - 3.988x_{10} + 27.6x_{11} + 22.527x_{12} + 0.021x_{13} - 523.438$$

$$y_2 = 52.373x_1 + 0.134x_2 + 28.029x_3 + 0.465x_4 + 0.496x_5 - 1.061x_6 + 0.75x_7 + 41.418x8 + 2.97x_9 - 3.856x_{10} + 31.177x_{11} + 18.445x_{12} + 0.000x_{13} - 427.379$$

$$y_3 = 54.127x_1 + 2.099x_2 + 35.906x_3 + 0.554x_4 + 0.485x_5 - 1.531x_6 - 9.235x_7 + 28.223x_8 + 2.317x_9 - 1.266x_{10} + 21.434x_{11} + 13.551x_{12} + 0.000x_{13} - 453.571$$

3) 从表 4 分类结果可以看出：

a) 使用回代法进行估计时，正确地对 100.0%个原始已分组个案进行了分类，误判率为 0。

b) 仅针对分析中的个案进行交叉验证。在交叉验证中，每个个案都由那些从该个案以外的所有个案派生的函数进行分类。

c) 使用交叉验证法进行估计时，正确地对 98.9%个进行了交叉验证的已分组个案进行了分类，误判率为 1.1%。

Table 4. Classification results
表 4. 分类结果

分类结果 [a,c]						
		葡萄酒类型	预测组成员信息			总计
			1	2	3	
原始	计数	1	59	0	0	59
		2	0	71	0	71
		3	0	0	48	48
	%	1	100.0	0.0	0.0	100.0
		2	0.0	100.0	0.0	100.0
		3	0.0	0.0	100.0	100.0
交叉验证 [b]	计数	1	59	0	0	59
		2	1	69	1	71
		3	0	0	48	48
	%	1	100.0	0.0	0.0	100.0
		2	1.4	97.2	1.4	100.0
		3	0.0	0.0	100.0	100.0

5.1.4. 费希尔判别分析

1) 首先，查看表 5“威尔克 Lambda”，是对函数判定有无价值的检验。可以看到，函数 1 和 2 的 Lambda 检验显著性 P = 0.000 < 0.05，提示采用函数 1 得到的判别结果是有效的，效能很高。

2) 进一步，由表 6 特征值表反映了典型函数所能解释的方差变异程度，即贡献率。由函数 1 的方差百分比值为 68.7%，说明函数 1 贡献率达到 68.7%，解释能力较高；函数 1 的方差百分比值为 31.3%，贡献率达到 31.3%，解释能力次之。

3) 结合查看表 7“典则判别函数系数(未标准化系数)”和图 2“典则判别函数”，图 2 中给出的是利用样本数据结合函数 1 和函数 2 判别出样本数据的所属类别。

Table 5. Wilke Lambda
表 5. 威尔克 Lambda

威尔克 Lambda				
函数检验	威尔克 Lambda	卡方	自由度	显著性
1 直至 2	0.019	666.795	26	0.000
2	0.195	276.282	12	0.000

Table 6. Characteristic values
表 6. 特征值

特征值				
函数	特征值	方差百分比	累计百分比	典型相关性
1	9.082[a]	68.7	68.7	0.949
2	4.128[a]	31.3	100.0	0.897

[a]在分析中使用了前 2 个典则判别函数。

Table 7. Coefficient of canonical discriminant function
表 7. 典则判别函数系数

典则判别函数系数		
	函数	
	1	2
酒精	0.403	0.872
苹果酸	−0.165	0.305
灰	0.369	2.346
灰的碱性	−0.155	−0.146
镁	0.002	0.000
总酚	−0.618	−0.032
类黄酮	1.661	−0.492
非黄烷类酚类	1.496	-1.631
花青素	−0.134	−0.307
颜色强度	−0.355	0.253
色调	0.818	−1.516
od280od315 稀疏葡萄酒	1.158	0.051
脯氨酸	0.003	0.003
(常量)	−9.231	−14.642
未标准化系数		

判别函数为

$$y_1 = 0.403x_1 - 0.165x_2 + 0.369x_3 - 0.155x_4 + 0.002x_5 - 0.618x_6 + 1.661x_7 + 1.496x_8 - 0.134x_9 - 0.355x_{10} + 0.818x_{11} + 1.158x_{12} + 0.003x_{13} - 9.231$$

$$y_2 = 0.872x_1 + 0.305x_2 + 2.346x_3 - 0.146x_4 + 0.000x_5 - 0.032x_6 - 0.492x_7 - 1.631x_8 - 0.307x_9 + 0.253x_{10} - 1.516x_{11} + 0.051x_{12} + 0.003x_{13} - 14.642$$

如图 2 两个判别函数得分的散点图所示，通过判别函数得分做出判别函数得分的散点图。三个类型的分离较大程度上显现在函数 y_1 上，而在函数 y_2 上第一类与第三类的分离效果不太理想，这与表 5 威尔克 Lambda 中的方差百分比反映了典型函数所能解释的方差变异程度一致。

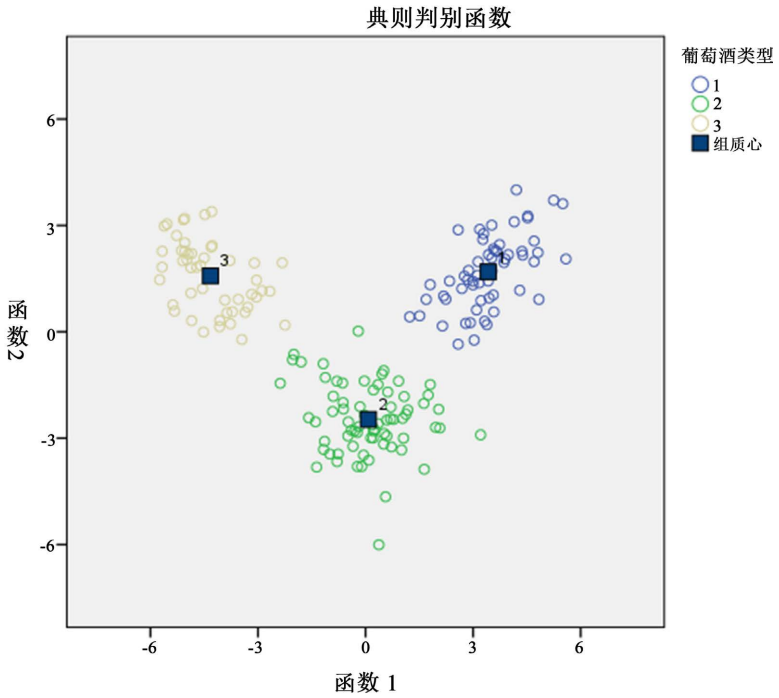

Figure 2. Scatter diagram of scores of two discriminant functions
图 2. 两个判别函数得分的散点图

5.2. 比较误判概率的非参数估计方法

分析表 8 可以发现，回代法是一种简单、直观且易于计算的方法。但不幸的是，误判率通常偏低，同一样本数据不仅用于构造判别函数，还被用于评价该判别函数，信息重复使用是不合理的，此实验中误判率为 0.0%。与使用所有样本数据相比，只使用部分样本数据会丢失许多有价值的信息，其有用性自然低于后者，这表明与前者误判率通常高于后者，此实验中误判率为 5.0%，而后者的误判概率才是我们真正关心的。当样本信息足够大时，回代法和划分样本这两种方法误判率的偏差可以忽略。交叉验证法较前两种方法效果较好，误判率为 1.1%，一般情况下最值得推荐。

Table 8. False judgment rate
表 8. 误判率

方法	误判率
回代法	0.0%
划分样本	5.0%
交叉验证	1.1%

6. 结论

通过上述实验分析我们可以发现，在知道其葡萄酒新样品来自三个产地之一，却不确定来自哪一个

产地时，将未知产地的葡萄酒新样品通过两个判别函数计算出两个判别函数得分，带入散点图中，通过目测散点图走势对新样品点所属类别进行主观判断，如果通过散点图走势无法判断，就采用距离判别法判别，判别新样本归属离它最近的那一组。在选择误判概率的非参数估计方法时，交叉验证方法是最推荐的方法，它不仅避免了同一样本数据不仅用于构造判别函数，还被用于评价该判别函数，从而导致不合理地重复使用信息，还在构造判别函数时几乎避免了样本信息的损失(只损失了一个样本观测)。通过分析可知，该判别方法对葡萄酒新样品所属产地的判别是快速有效的，在生产生活中，类似的实际分类判别问题很多，可以考虑使用此方法。此外，分类判别考虑多方面因素，方法之间各有优略，快速有效的方法还有很多等待我们结合实际问题去探索。

参考文献

[1] 王学民. 应用多元统计分析[M]. 上海: 上海财经大学出版社, 2017.

[2] 王晓华. AI 制胜[M]. 北京: 清华大学出版社, 2020.