

基于加权多元密度核估计的 Bayes 判别分析

1、摘要

Bayes 判别分析通常需要假定数据服从正态分布，而实际样本未必服从正态分布，因此本文考虑在 Bayes 判别法中不假设总体满足任何分布，而是通过非参数多元密度核估计的方法直接估计总体中各类的密度函数，同时采用随机森林为各特征变量赋权，建立了基于加权多元密度核估计的 Bayes 判别法。而后考虑到核函数可以将数据映射到高维空间，使得不可分数据转变为高维空间中的线性可分数据。因此可以将核函数引入距离判别分析来提高其在非线性可分数据上的判别准确率，建立了基于核函数相似性度量的判别分析算法。

2、改进措施

(1) 多元密度核函数估计

设 p 维连续随机变量 X ，具有未知密度函数 $f(x)$ ，抽取样本 $X_1, X_2, \dots, X_n$ ，其中 $X_i = (X_{i1}, \dots, X_{ip})^T, i = 1, 2, \dots, n$ ，则 $f(x)$ 的密度核估计为

$$\hat{f}(x) = \frac{1}{nh_1 \dots h_p} \sum_{i=1}^n \left[\prod_{j=1}^p K\left(\frac{x_j - x_{ij}}{h_j}\right) \right]$$

$K(\cdot)$ 为核函数， $h_i, i=1, 2, \dots, p$ 为各维窗宽。

(2) 多元密度核估计的 Bayes 判别法

设总体样本分布为 $x_i^1 = (x_{i1}^1, x_{i2}^1, \dots, x_{ip}^1)^T, i = 1, 2, \dots, n_1$ ，与 $x_i^2 = (x_{i1}^2, x_{i2}^2, \dots, x_{ip}^2)^T, i = 1, 2, \dots, n_2$ ，待判别样本 $x = (x_1, \dots, x_p)^T$ ，则是平均误判损失极小的划分为：

$$\begin{cases} R_1 = \{x: \omega(x) \geq \beta\} \\ R_2 = \{x: \omega(x) < \beta\} \end{cases}$$

其中

$$\begin{aligned} \omega(x) &= \ln \left\{ \sum_{i=1}^{n_1} \left[\prod_{j=1}^p K\left(\frac{x_j - x_{ij}^1}{h_j^1}\right) \right] \right\} - \ln \left\{ \sum_{i=1}^{n_2} \left[\prod_{j=1}^p K\left(\frac{x_j - x_{ij}^2}{h_j^2}\right) \right] \right\} \\ \beta &= \ln \left[\frac{L(1|2) p_2}{L(2|1) p_1} \cdot \frac{n_1}{n_2} \frac{h_1^1}{h_1^2} \dots \frac{h_p^1}{h_p^2} \right] \end{aligned}$$

本文使用高斯核函数 $K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$ ，且设所有窗宽相等，即为高斯核函数的参数 σ ，于是最终的判别函数形式为：

$$\begin{aligned} \omega(x) &= \ln \left\{ \sum_{i=1}^{n_1} e^{-\frac{1}{2} \sum_{j=1}^p \left(\frac{x_j - x_{ij}^1}{\sigma}\right)^2} \right\} - \ln \left\{ \sum_{i=1}^{n_2} e^{-\frac{1}{2} \sum_{j=1}^p \left(\frac{x_j - x_{ij}^2}{\sigma}\right)^2} \right\} \\ \beta &= \ln \left[\frac{L(1|2) p_2}{L(2|1) p_1} \cdot \frac{n_1}{n_2} \right] \end{aligned}$$

(3) 基于核函数相似性度量的判别法

在距离判别算法中，对于高维数据来说，传统的距离计算方法往往需要较大的计算量，除了计算距离，也可以计算两个样本的相似度来判断其是否属于同一类别，因此考虑基于核函数计算向量夹角来度量样本间的相似度，而后进行判别分析。判别规则如下：

$$\begin{aligned} W(x) &= S(x, \mu_1) - S(x, \mu_2) \\ \begin{cases} x \in X_1, & \text{when } W(x) \geq 0 \\ x \in X_2, & \text{when } W(x) < 0 \end{cases} \end{aligned}$$

$$\text{其中 } S(x_i, x_j) = \frac{K(x_i, x_j)}{\sqrt{K(x_i, x_i)K(x_j, x_j)}}。$$

3、数据来源和采集时间

本次案例使用西南证券软件提供的数据，采集航空板块 30 支股票和酿酒板块 36 支股票 2021 年 4 月 10 日的相关指标进行分析，选取涨幅、涨速、换手、量比、振幅、强弱度、均涨幅和实体涨幅共 8 个指标，选取第 26、27、28、29、30、61、62、63、64、65、66 个数据为测试集。

4、R 程序分析结果

(1) 基于加权多元密度核估计的 Bayes 判别分析

首先对数据构建随机森林模型，得到各个特征的重要性排序，然后以 Gini 指数作为衡量特征权重的指标。基于训练集以本文 2.2 节提出的基于加权多元密度核估计的 Bayes 判别法构建判别函数 (σ 取 1.5)，并于传统 Bayes 判别法进行对比，判别正确率见表 1，预测结果见表 2。

(2) 基于核函数相似性度量的判别法

按 2.3 节中构建的判别规则在训练集上构建判别函数 (σ 取 0.5)，并于传统的基于马氏距离的距离判别法进行对比，判别正确率见表 1，预测结果见表 2。

表 1 两总体不同判别分析方法的正确率

方法	训练集正确率	测试集正确率
传统距离判别	0.8909091	0.8181818
基于核函数相似性度量判别	0.6545455	0.8181818
Bayes 判别	0.9090909	0.8181818
基于加权多元密度核估计的 Bayes 判别	0.7090909	0.8181818

表 2 不同判别方法在测试集上的预测结果

方法	航天电子	航发动力	纵横股份	行亚科技	江航装备	会稽山	迎驾贡酒	今世缘	口子窖	ST 威龙	金徽酒
传统距离判别	1	2	1	1	1	2	2	2	2	1	2
基于核函数相似性度量判别	1	2	2	1	1	2	2	2	2	2	2
Bayes 判别	1	2	1	1	1	2	2	2	2	1	2
基于加权多元密度核估计的 Bayes 判别	1	1	1	1	1	2	1	2	2	1	2

5、结论与展望

从判别准确率可以看出，基于核函数改进后的距离判别和 Bayes 判别效果并没有更好，虽然测试集上的正确率相同，但在训练集上的表现却不如传统判别方法。考虑可以从以下几个方面来改进：(1) 通过增加参数等方法改进高斯核函数；(2) 尝试其他核函数看是否有所改进；(3) 尝试其他基于核函数度量样本相似性的公式。

6、参考文献

[1] 高全力,高岭,石美红,朱欣娟,陈锐,赵雪青.多特征的核线性判别分析推荐方法[J].东南大学学报(自然科学版),2019,49(05):883-889.
[2] 陈相兵.基于多元密度核估计的 Bayes 判别法[J].数学的实践与认识,2015,45(10):185-189.