判别分析

--方法介绍

判别分析是一种多元统计方法。在判别分析中，往往需要研究考查对象的多个指标或变量，也就是说要有多个判别变量，才能建立合理的判别规则，即判别函数。从统计的角度来看，判别分析可以描述为：已知有个总体，现有样本 y，要根据这 k 个总体和当前样本的特征，判定该样本 y 属于哪一个总体。其主要工作是根据对已知总体的理解，建立判别规则（又称判别函数），然后根据该判别规则对新的样本属于哪个总体做出判断。

最简单的判别分析方法是距离判别法。距离判别法首先根据已知分类的数据，分别计算出各类的重心。再根据新个体到每类的距离（即新个体与各类重心的距离，可采用欧氏距离或者马氏距离等等），根据最短的距离确定分类情况。

设有两个总体（两类）G1 和 G2，从第一个总体 G1 中抽取出 n1 个样本，从第二个总体 G2 中抽取出 n2 个样本，每个样本测量 p 个指标特征。现今任取一个样本 X，实测其 p 个指标特征的值。$X=\left(X_1,X_2,...,X_P\right)^T$，问此样品 X 应被判归为哪一类？

首先计算 X 到 G1,G2 总体的距离，分别为 $D\left(X,G1\right)$ 和 $D\left(X,G2\right)$

若采用欧氏距离计算，则有：

$$D\left(X,G\mathrm{i}\right)=\sqrt{\left(X-\bar{X}^{(\mathrm{i})}\right)^T\left(X-\bar{X}^{(\mathrm{i})}\right)}=\sqrt{\sum_{\mathrm{k}=1}^{p}\left(X_{\mathrm{k}}-\bar{X}_{\mathrm{k}}^{(\mathrm{i})}\right)^2}$$

其中 i=1,2; $\bar{X}^{(\mathrm{i})}$ 表示第 i 类的重心（也就是均值），$\bar{X}_{\mathrm{k}}^{(\mathrm{i})}$ 表示第 i 类第 k 个指标的重心。

若采取马氏距离。则有

$$D\left(X,G\mathrm{i}\right)=\sqrt{\left(X-\bar{X}^{(\mathrm{i})}\right)^{T}\left(\Sigma^{\mathrm{i}}\right)^{(-1)}\left(X-\bar{X}^{(\mathrm{i})}\right)}$$

其中 i=1,2; $\bar{X}^{(\mathrm{i})}$表示第 i 类的重心（也就是均值）,$\left(\Sigma^{\mathrm{i}}\right)^{(-1)}$表示第 i 类的协方差矩阵的逆矩阵。

判别分析在现实中有着广泛的应用。例如，在环境科学上，根据大气中各种颗粒的指标来判断地区是严重污染、中度污染还是无污染等。