目录

随机变量.....2

多元正态分布 2

线性回归.....3

判别分析.....4

聚类分析.....6

主成分分析.....8

随机变量

1. 概率密度函数

概率密度函数非负，且在负无穷到正无穷上的积分为1。

2. 数字特征

设$\mathbf{X}=\left(X_1,X_2,X_3,\cdots,X_p\right)'$，若$E(X_i)\ (i=1,\cdots,p)$存在且有限，则称$E(\mathbf{X})=\left(E(X_1),E(X_2),\cdots,E(X_p)\right)'$为$\mathbf{X}$的均值（向量）或数学期望。

均值向量的性质：

1) $E(\mathbf{AX})=\mathbf{A}E(\mathbf{X})$

2) $E(\mathbf{AXB})=\mathbf{A}E(\mathbf{X})\mathbf{B}$

3) $E(\mathbf{AX}+\mathbf{BY})=\mathbf{A}E(\mathbf{X})+\mathbf{B}E(\mathbf{Y})$

其中，$\mathbf{X}$、$\mathbf{Y}$为随机白能量，$\mathbf{A}$、$\mathbf{B}$为大小合适运算的常熟矩阵。

$Cov(\mathbf{X},\mathbf{Y})\triangleq E(\mathbf{X}-\mathbf{E}(\mathbf{X}))(\mathbf{Y}-\mathbf{E}(\mathbf{Y}))'$，$Cov(\mathbf{X},\mathbf{X})=D(\mathbf{X})$。

若$Cov(\mathbf{X},\mathbf{X})=0$，则称 $\mathbf{X}$ 和 $\mathbf{Y}$ 不相关。

3. 相关系数

若$\mathbf{X}=\left(X_1,X_2,X_3,\cdots,X_p\right)'$的协差阵存在，且每个分量的方差大于零，则称随机向量 $\mathbf{X}$ 的相关阵为$\boldsymbol{R}=Corr(\boldsymbol{X})=(\rho_{ij})_{p\times p}$，其中

$$\rho_{ij}=\frac{Cov\left(X_i,X_j\right)}{\sqrt{D(X_i)}\sqrt{D(X_j)}}=\frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}$$

为X_i与X_j的相关系数。

多元正态分布

1. 多元正态分布的定义

1) 采用特征函数来定义

2) 使用多元正态分布的充要性质来定义

p 维随机向量服从 p 维正态分布的充要条件是其任意线性组合（系数不全为 0）服从以为正态分布。

3) 使用标准正态分布性质来定义

2. 多元正态分布的性质

1) 若$\mathbf{X}=\left(X_1,X_2,X_3,\cdots,X_p\right)'\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$，$\boldsymbol{\Sigma}$是对角阵，则$X_1,\cdots,X_p$相互独立。

2) 若$\mathbf{X}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$，$\boldsymbol{A}$ 为s × p阶常数阵，$\boldsymbol{d}$ 为 s 维常数向量，则

$$\boldsymbol{AX}+\boldsymbol{d}\sim N_s(\boldsymbol{A\mu}+\boldsymbol{d},\boldsymbol{A\Sigma A}')$$

即正态随机向量的线性函数还是正态的。

3) 随机向量$\mathbf{X}=\left(X_1,X_2,X_3,\cdots,X_p\right)'$的协方差阵$D(\mathbf{X})=\boldsymbol{\Sigma}$是对称非负定矩阵。

线性回归

1. 回归拟合优度

TSS(总离差平方和)，RSS(回归平方和)，ESS(残差平方和)，公式：$TSS=\sum_{i=1}^{n}(y_i-\bar{y})^2$，$RSS=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$，$ESS=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$。其中存在$TSS=RSS+ESS$，则拟合优度公式为：$R^2=\frac{RSS}{TSS}$，值越大表明直线拟合程度越好。

2. 逻辑回归

逻辑回归模型是使用一个函数来归一化 y 值，使 y 的取值在区间(0,1)内，这个函数称为 Logistic 函数，它将预测问题转化为了一个概率问题。

判别分析

1. 什么是判别分析

判别分析通过学习以往数据（根据历史上划分类别的有关资料和某种最优准则，确定一种判别方法），判定一个新的样本归属哪一类。

判别分析常用的判别方法：距离判别法、贝叶斯判别法、费歇判别法、逐步判别法。

判别分析分类：

1) 按判别的组数：两组判别分析和多组判别分析

2) 按判别函数的形式：线性判别和非线性判别

3) 按判别时所处理的变量方法不同：逐步判别和序贯判别等

4) 按照判别准则：马氏距离最小准则、Fisher 准则、平均损失最小准则、最大概率准则等

5) 根据资料的性质：定性资料和定量资料的判别分析

2. 距离判别法

1) 马氏距离

设总体 G 为 m 元总体（考察 m 个指标），均值向量为$\mu = (\mu_1, \mu_2, \mu_3, \cdots, \mu_m)'$，协方差阵为$\Sigma = (\sigma_{ij})_{m\times m}$，则样品$X = (x_1, x_2, x_3, \cdots, x_m)'$与总体 G 的马氏距离为

$$D^2(X, G) = (X - \mu)'\Sigma^{-1}(X - \mu)$$

总体 G 内两点 X 与 Y 之间的马氏距离为

$$D^2(X, Y) = (X - Y)'\Sigma^{-1}(X - Y)$$

2) 欧式距离：当$\Sigma = I$（单位矩阵）时，为欧式距离

3) 欧式距离与马氏距离的优缺点

缺点：就大部分统计问题而言，欧式距离时不能令人满意的。每个坐标对欧式距离的贡献是同等的。当坐标表示测量值时，它们往往带有大小不等的随机波动，在这种情况下，合理的方法是对坐标加权，使变化较大的坐标比变化较小的坐标有较小的权系数，这就产生了各种距离。当各个分量为不同性质的量时，距离的大小与指标的单位有关。它将样品的不同属性之间的差别等同看待，这一点有时不能满足实际要求。没有考虑到总体变异对距离远近的影响。

马氏距离表示数据的协方差距离。为两个服从同一分布并且其协方差矩阵为Σ的随机变量的差异长度：如果协方差矩阵为单位阵，那么马氏距离就简化为欧式距离，如果协方差矩阵为对角阵，则其也可称为正规化的欧式距离。

优点：它不受量纲的影响，两点之间的马氏距离与原始数据的测量单位无关。由标准化数据和中心化数据计算出的二点之间的马氏距离相同。马氏距离还可以排除变量之间的相关性的干扰。

3. 费歇（Fisher）判别法

1) 基本思想

费歇判别的基本思想是将高维数据点投影到低维空间上来，然后利用方差分析的思想选出一个最优的投影方向。

2) 实现方法

选择投影方向 a 应使投影的 k 个一元总体能尽量分开（同一总体中的样品的投影尽量靠近），k 要做到这一点，只要投影后的 k 个

一元总体均值由显著差异，即可利用方差分析的方法使组间平房和尽可能的大。则选取投影方向 a 使$\Delta\alpha$达到极大即可。

4. 贝叶斯（Bayes）判别法

1) 基本思路

贝叶斯判别法的思路使先假定对研究对象已有一定的认识，常用先验概率分布来描述这种认识，然后取得一个样本，用样本来修正已有的认识，得到后验概率分布，各种统计推断都通过后验概率分布来进行。

2) 不足之处

判别法与总体各自出现的概率的大小无关；判别方法与错判之后造成的损失无关。

聚类分析

1. 聚类分析方法

系统聚类法、模糊聚类分析、K 均值聚类分析。

2. 聚类分析根据分类对象的不同分为 R 型和 Q 型聚类

对样品的分类称为 Q 型聚类，对变量的分类称为 R 型聚类。

3. 聚类分析根据聚类方法的不同分为快速聚类、谱系聚类、模糊聚类法、逐步聚类法，最优分割法，分解法，加入法等。

4. 聚类分析的基本思想

其基本思想是研究的样品或指标之间存在着程度不同的相似性，于是根据一批样品的多个观测指标，具体找出一些能够度量样品或指标至今啊的相似度从统计量，以这些统计量作为划分类型的已经，把

一些相似程度较大的样品聚类合为一类，把另外一些彼此之间相似度较大的样品又聚合为另外一类，知道把所有的样品聚合完毕，形成一个由小到大的分类系统。

5. 系统聚类法的原理和具体步骤

原理：先将各个样品个堪称一类，然后规定类与类之间的距离，选择距离最小的一堆合并成新的一类，计算新类与其他类之间的距离，再将距离最近的两类合并，这样每次减少一类，直至所有的样品合为一类为止。

步骤：对数据进行变换处理；构造 n 个类，每个类只包含一个样本；计算 n 个样本两两之间的距离；合并距离最近的两类为一新类；计算新类与当前各类的距离；画聚类图；决定类的个数，从而得出分类结果。

6. K-均值聚类的方法原理

将样本分为 K 个初始类，通过将各个样本分配到最近均值的类中后重新计算接受了新样品的类和事情样品的类的均值，不断迭代，直到各个类，直到各类无元素进出。

7. 模糊聚类的思想

模糊聚类分析是根据客观事物间的特征、亲疏、相似性，通过建立模糊相似关系对客观事物进行聚类的分析方法。

8. 关联测度

关联测度用来衡量定性变量的相似性，常见的有简单匹配系数、雅科比系数、匹配系数，其中匹配系数、雅科比系数只适用于二分类

变量。

主成分分析

1. 什么是主成分分析

在力求数据信息丢失最少的原则下，对多维变量进行加工将为，即研究多维变量的线性组合，使它们构成的综合变量尽可能多地保留原始变量的信息，这些综合变量就称为主成分。

2. 主成分与原始变量的关系

每个主成分都是各原始变量的线性组合；主成分的数目大大少于原始变量的数目；主成分保留了原始变量绝大多数信息；各主成分之间互不相关。