|  |  |
| --- | --- |
| **Pima Indians Diabetes Data Set**   **数据来源于国家糖尿病、消化和肾脏疾病研究所，包括成本数据（Peter Turney捐赠）** |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 768 | **Area:** | Life |
| **Attribute Characteristics:** | Integer, Real | **Number of Attributes:** | 8 | **Date Donated** | 1990-05-09 |
| **Associated Tasks:** | Classification | **Missing Values?** | Yes | **Number of Web Hits:** | 191206 |

**数据集信息：**

通过一定约束条件，从较大的数据库中选择了这些数据。特别是，这里的所有患者都是至少21岁的皮马印第安人血统的女性。

**Attribute Information:**

1. Number of times pregnant   
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test 口服葡萄糖耐量试验2小时后测定血浆葡萄糖浓度  
3. Diastolic blood pressure (mm Hg)舒张压  
4. Triceps skin fold thickness (mm) 肱三头肌皮肤褶皱厚度(mm)  
5. 2-Hour serum insulin (mu U/ml) 2小时血清胰岛素  
6. Body mass index (weight in kg/(height in m)^2) 体重指数  
7. Diabetes pedigree function 糖尿病遗传指数  
8. Age (years)   
9. Class variable (0 or 1)   
  
**\*\* UPDATE: Until 02/28/2011 this web page indicated that there were no missing values in the dataset. As pointed out by a repository user, this cannot be true: there are zeros in places where they are biologically impossible, such as the blood pressure attribute. It seems very likely that zero values encode missing data. However, since the dataset donors made no such statement we encourage you to use your best judgement and state your assumptions.**

数据见txt文档 discriminant analysis dataset。

请仔细阅读数据说明、变量说明，了解该数据的背景含义，理解所研究问题的实际意义和实用价值。针对该数据，完成实验内容，并回答以下问题，并给与阐述。

1. 在数据说明中，提出可能数据中可能存在用0代替的缺失值，请根据你的判断来确定是否存在缺失？如果存在缺失请用适当的方法对其进行处理，请给出方案步骤。

**在 Plasma glucose concentration a 2 hours in an oral glucose tolerance test（ 口服葡萄糖耐量试验2小时后测定血浆葡萄糖浓度）、Diastolic blood pressure（舒张压）、Triceps skin fold thickness (肱三头肌皮肤褶皱厚度)、2-Hour serum insulin (mu U/ml) （2小时血清胰岛素）、Body mass index（体重指数）这些数据变量中，出于生理原则考虑不应该为0的地方出现了0值，因此考虑其为缺失值。**

**方案步骤：**

1. **导入数据，查找出相应变量下0的位置并赋值为缺失值**

mydata <- read.csv("~/rstdio\_wsl/discriminant analysis data.txt", header=FALSE)

mydata$V2[which(mydata$V2==0)]<-NA

mydata$V3[which(mydata$V3==0)]<-NA

mydata$V4[which(mydata$V4==0)]<-NA

mydata$V5[which(mydata$V5==0)]<-NA

mydata$V6[which(mydata$V6==0)]<-NA

1. **利用单个插补法补全缺失值，即用该缺失值所在的变量的平均值替代**

library(Hmisc)

mydata$V2=impute(mydata$V2,mean)

mydata$V3=impute(mydata$V3,mean)

mydata$V4=impute(mydata$V4,mean)

mydata$V5=impute(mydata$V5,mean)

mydata$V6=impute(mydata$V6,mean)

mydata$V9=as.factor(mydata$V9)

ceshidata=tail(mydata,20)

xunliandata=mydata[1:748,]

1. 建立判别分析模型，用于对变量Class variable进行判别分析。将数据集的最后20个数据作为测试数据集，其他数据集作为训练数据集。请完成如下几个判别分析方法的计算：

**设有两个总体和，是一个维样品，根据与两个总体的距离和，若样品到总体的距离小于到总体的距离，则认为样品属于样品总体，反之，则认为样品属于总体.即**

**当总体1和2为正态总体且协方差相等时，用马氏距离，即**

**当总体不是正态分布总体时，有时也用马氏距离描述x到总体的远近.**

（Tips

**关于方差分析与判别分析的比较：**虽然两者都有一些假设，但是方差分析是通过检验来判断，所以对假设要求更高。判别分析是结果导向的，是利用误判率、最小损失等结果来评判的，所以对假设要求虽然也有，但是不是那么强烈。

**关于正态检验：**由于多维正态分布一般比较难检验，所以在实际做的时候，经常直接做线性判别或者二次判别，通过结果来判定方法的好坏。可以做正态检验，如果不成立，另外有专门的非参数检验方法[非参数拟合分布下的贝叶斯判别] ）

## 2.1线性判别

若，这时

令

于是判别规则可以表示为

**称为判别函数，且是线性函数.**

> library(MASS)

> attach(xunliandata)

> LD<-lda(V9~.,xunliandata)

> #线性判别

> LDPredict <- predict(LD,newdata=ceshidata)

> #预测的所属类的结果

> newGrop1 <- LDPredict$class

> #显示预测前后分组结果

> cbind(ceshidata$V9,newGrop1)

> #构造混淆矩阵，求出判对率

> tab <- table(ceshidata$V9,newGrop1)

> sum(diag(prop.table(tab)))

[1] 0.8

> LD

Call:

lda(V9 ~ ., data = xunliandata)

Prior probabilities of groups:

0 1

0.6550802 0.3449198

Group means:

V1 V2 V3 V4 V5 V6 V7 V8

0 3.281633 110.7532 70.92592 27.68779 142.4288 30.88269 0.4322776 31.18163

1 4.906977 141.6410 75.13753 31.75091 180.0264 35.39967 0.5533178 36.86047

Coefficients of linear discriminants:

LD1

V1 0.1010892610

V2 0.0287011816

V3 -0.0043015889

V4 0.0018355966

V5 -0.0004101456

V6 0.0643752996

V7 0.5640582855

V8 0.0071340688

> a=colMeans(LD$means)

> a

V1 V2 V3 V4 V5 V6 V7

4.0943047 126.1970835 73.0317231 29.7193496 161.2275907 33.1411792 0.4927977

V8

34.0210489

> b=LD$scaling

> b

LD1

V1 0.1010892610

V2 0.0287011816

V3 -0.0043015889

V4 0.0018355966

V5 -0.0004101456

V6 0.0643752996

V7 0.5640582855

V8 0.0071340688

> a%\*%b

LD1

[1,] 6.364318

**所以W(x)=**

## 2.2 二次判别

当时，可用

**作为判别函数，这时它是二次函数.**

> #二次判别

> QDPredict <- predict(QD,newdata=ceshidata)

> #预测的所属类的结果

> newGrop2 <- QDPredict$class

> #显示预测前后分组结果

> cbind(ceshidata$V9,newGrop2)

> #构造混淆矩阵，求出判对率

> tab <- table(ceshidata$V9,newGrop2)

> sum(diag(prop.table(tab)))

[1] 0.85

2.3 贝叶斯判别

贝叶斯（Bayes）统计的思想是：假定对研究的对象已有一定的认识，常用先验概率分布来描述这种认识，然后我们取得一个样本，用样本来修正已有的认识（先验概率分布），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。将贝叶斯思想用于判别分析，就得到贝叶斯判别。

> #基于两组协差阵相等的贝叶斯判别

> LDB1<-lda(V9~.,prior=c(0.5, 0.5),xunliandata) #先验概率相等的线性判别

> Z<-predict(LDB1) #根据线性判别函数预测所属类别

> Z$posterior #后验概率结果

> newg<-Z$class #预测的所属类别结果

> cbind(xunliandata$V9, round(Z$posterior, 3), newg) #按列合并的结果

> table(xunliandata$V9, newg) #判别情况表

newg

0 1

0 393 97

1 76 182

LDB2<-lda(V9~.,xunliandata) #先验概率不相等的线性判别

Z<-predict(LDB2)

newg<-Z$class #预测的所属类别结果

cbind(xunliandata$V9, round(Z$posterior, 3), newg) #按列合并的结果

table(xunliandata$V9, newg) #判别情况表

newg

0 1

0 434 56

1 114 144

基于两组协差阵不等的贝叶斯判别

QDB1<-qda(V9~., prior=c(0.5, 0.5), xunliandata) #二次判别

Z<-predict(QDB1) #根据二次判别函数预测所属类别

newg<-Z$class #预测的所属类别结果

cbind(xunliandata$V9, round(Z$posterior, 3), newg) #按列合并的结果

table(xunliandata$V9, newg) #判别情况表

newg

0 1

0 392 98

1 80 178

QDB2<-qda(V9~., xunliandata) #二次判别

Z<-predict(QDB2) #根据二次判别函数预测所属类别

newg<-Z$class #预测的所属类别结果

cbind(xunliandata$V9, round(Z$posterior, 3), newg) #按列合并的结果

table(xunliandata$V9, newg) #判别情况表

newg

0 1

0 422 68

1 113 145

**所以，当协差阵相等时，选择先验概率相等。协差阵不相等时，先验概率可以选取按与各组样本容量大小成比例的概率**

## 2.4 Fisher判别

Fisher判别法的思想是用投影，将K组P维数据投影到某个方向，使它们的投影组与组之间的距离尽可能分开。即找到一个线性组合：Z=c1x1+c2x2+L+CPXP,使得综合指标Z在A组的均数与B组的均数的差异 尽可能大，而两组的组内综合指标的变异（S2A+S2B）尽可能小，也就是组间差异尽可能大，组内变异尽可能小.

#fisher线性判别

Z=as.matrix(xunliandata)

fit=lda(factor(Z[,9])~Z[,1]+Z[,2]+Z[,3]+Z[,4]+Z[,5]+Z[,6]+Z[,7]+Z[,8]);#Fisher线性判别

fit; #输出包括先验概率, 各类的均值/重心, 线性判别函数(原始变量无截距),特征值比例

> B=predict(fit,xunliandata)$class #预测的类别

> table(xunliandata$V9,B) # cross-table

B

0 1

0 467 23

1 23 235

请汇总这几种方法的判别结果进行比较。分别给出在训练数据集和测试数据集下各个方法的判别结果比较，说明哪种方法最优。

**在训练数据集下：**

**线性判别，判对率为77.3%**

**二次判别，判对率为75.8%**

**Bayes判别**

**协差阵相等，先验概率相等：判对率为76.9%**

**协差阵相等，先验概率按与各组样本容量大小成比例的概率:判对率为77.3%**

**协差阵不相等，先验概率相等：判对率为76.2%**

**协差阵不相等，先验概率按与各组样本容量大小成比例的概率:判对率为75.8%**

**Fisher判别，判对率为93.9%**

**在测试数集下：**

**线性判别，判对率为80%**

**二次判别，判对率为85%**

**Bayes判别**

**协差阵相等，先验概率相等：判对率为85%**

**协差阵相等，先验概率按与各组样本容量大小成比例的概率:判对率为80%**

**协差阵不相等，先验概率相等：判对率为95%**

**协差阵不相等，先验概率按与各组样本容量大小成比例的概率:判对率为85%**

**Fisher判别，判对率为95%**

综上，fisher判别最优