多元统计 11

罗震林

1 Question 11.1

1.1 (a)

由题易知，线 $f(x,y)=ax+by+c=0$ 的法向量为 $\overrightarrow{n}=(a,b)$. 设点 (h,k) 为 A，作 $AB\perp f(x,y)$ 交 $f(x,y)$ 于 $B(x_0,y_0)$，则有 $ax_0+by_0+c=0$.

因为 $\overrightarrow{n}$ 和 $\overrightarrow{BA}$ 都垂直于 $f(x,y)$，所以 $\overrightarrow{n}$ 和 $\overrightarrow{BA}$ 共线，从而有

$$
\begin{aligned}
\overrightarrow{n}\cdot\overrightarrow{BA} &= \pm|\overrightarrow{n}||\overrightarrow{BA}| \\
&= (a,b)\cdot(h-x_0,k-y_0) \\
&= a(h-x_0)+b(k-y_0) \\
&= ah+bk-(ax_0+by_0) \\
&= ah+bk+c
\end{aligned}
$$

因此，

$$
\begin{aligned}
|\overrightarrow{BA}| &= \frac{|\overrightarrow{n}\cdot\overrightarrow{BA}|}{|\overrightarrow{n}|} \\
&= \frac{|ah+bk+c|}{\sqrt{a^2+b^2}}
\end{aligned}
$$

所以，点 (h,k) 到线 $f(x,y)$ 的垂直距离为 $|ah+bk+c|/\sqrt{a^2+b^2}$

1.2 (b)

不妨将问题转化为

$$
\begin{aligned}
&\text{minimize} \quad \frac{1}{2}||\mathbf{x}-\mathbf{x}_k||^2 \\
&\text{subject to} \quad \mu(\mathbf{x})=\beta_0+\mathbf{x}^T\boldsymbol{\beta}=0
\end{aligned}
$$

使用拉格朗日乘子法，则

$$L(\mathbf{x},\lambda)=\frac{1}{2}(\mathbf{x}^T\mathbf{x}-2\mathbf{x}^T\mathbf{x}_k+\mathbf{x}_k^T\mathbf{x}_k)+\lambda(\beta_0+\mathbf{x}^T\boldsymbol{\beta})$$

对 $\mathbf{x}$ 求偏导且令其等于 0, 有

$$\frac{\partial L}{\partial\mathbf{x}}=\mathbf{x}-\mathbf{x}_k+\lambda\boldsymbol{\beta}=0$$

从而满足条件的解 $\mathbf{x}^*$，有

$$\mathbf{x}^* = \mathbf{x}_k - \lambda\boldsymbol{\beta}$$

带入约束条件中, 有,

$$\begin{aligned}
&\beta_0 + \mathbf{x}^{*T}\boldsymbol{\beta} = \beta_0 + \mathbf{x}_k^T\boldsymbol{\beta} - \lambda\boldsymbol{\beta}^T\boldsymbol{\beta} = 0 \\
&\lambda = \frac{\beta_0 + \mathbf{x}_k^T\boldsymbol{\beta}}{\boldsymbol{\beta}^T\boldsymbol{\beta}}
\end{aligned}$$

从而 $\mathbf{x}_k$ 到超平面的最短垂直距离为

$$\begin{aligned}
||\mathbf{x} - \mathbf{x}_k|| &= ||\lambda\boldsymbol{\beta}|| = |\lambda|||\boldsymbol{\beta}|| \\
&= \frac{|\mu(\mathbf{x})|}{||\boldsymbol{\beta}||^2}||\boldsymbol{\beta}|| \\
&= \frac{|\mu(\mathbf{x})|}{||\boldsymbol{\beta}||}
\end{aligned}$$

2 Question 11.6

设 $\mathbf{x} = (x_1, y_1)^T, \mathbf{y} = (y_1, y_2)^T$, 则

$$\begin{aligned}
K(\mathbf{x}, \mathbf{y}) &=< \mathbf{x}, \mathbf{y} >^2 = (x_1y_1 + x_2y_2)^2 \\
&= x_1^2y_1^2 + x_2^2y_2^2 + 2x_1x_2y_1y_2
\end{aligned}$$

所以 $\mathbf{\Phi}_1(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)^T$ 和 $\mathbf{\Phi}_2(\mathbf{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)^T$ 都满足

$$K(\mathbf{x}, \mathbf{y}) =< \mathbf{\Phi}_1(\mathbf{x}), \mathbf{\Phi}_1(\mathbf{y}) >=< \mathbf{\Phi}_2(\mathbf{x}), \mathbf{\Phi}_2(\mathbf{y}) >$$

且 $\mathbf{\Phi}_1(\mathbf{x})$ 和 $\mathbf{\Phi}_1(\mathbf{x})$ 都是 $\mathcal{R}^2$ 到 $\mathcal{R}^3$ 的映射

3 Question 11.11

将 (11.41) 写成矩阵形式有

$$\begin{aligned}
F_D(\theta\boldsymbol{\alpha}) &= \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_jy_iy_j(\mathbf{x}_i^T\mathbf{x_j}) \\
&= \mathbf{1}^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^TQ\boldsymbol{\alpha}
\end{aligned}$$

其中，Q 为对称矩阵，且第 (i,j) 个元素为 $Q_{ij} = y_iy_j\mathbf{x}_i^T\mathbf{x}_j$ 所以，将不等式左右两边依次改写成矩阵形式有

$$\begin{aligned}
&F_D(\theta\boldsymbol{\alpha} + (1-\theta)\boldsymbol{\beta}) = \mathbf{1}^T(\theta\boldsymbol{\alpha} + (1-\theta)\boldsymbol{\beta}) - \frac{1}{2}(\theta\boldsymbol{\alpha} + (1-\theta)\boldsymbol{\beta})^TQ(\theta\boldsymbol{\alpha} + (1-\theta)\boldsymbol{\beta}) \\
&\theta F_D(\boldsymbol{\alpha}) = \theta[\mathbf{1}^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^TQ\boldsymbol{\alpha}] \\
&(1-\theta)F_D(\boldsymbol{\beta}) = (1-\theta)[\mathbf{1}^T\boldsymbol{\beta} - \frac{1}{2}\boldsymbol{\beta}^TQ\boldsymbol{\beta}]
\end{aligned}$$

从而有

$$
\begin{aligned}
&F_D(\theta\boldsymbol{\alpha}+(1-\theta)\boldsymbol{\beta})-[\theta F_D(\boldsymbol{\alpha})+(1-\theta)F_D(\boldsymbol{\beta})]\\
=&\frac{1}{2}[\theta\boldsymbol{\alpha}^TQ\boldsymbol{\alpha}+(1-\theta)\boldsymbol{\beta}^TQ\boldsymbol{\beta}-\theta^2\boldsymbol{\alpha}^TQ\boldsymbol{\alpha}-(1-\theta)^2\boldsymbol{\beta}^TQ\boldsymbol{\beta}-\theta(1-\theta)\boldsymbol{\beta}^TQ\boldsymbol{\alpha}-\theta(1-\theta)\boldsymbol{\alpha}^TQ\boldsymbol{\beta}]\\
=&\theta(1-\theta)\frac{1}{2}[\boldsymbol{\alpha}^TQ\boldsymbol{\alpha}+\boldsymbol{\beta}^TQ\boldsymbol{\beta}-\boldsymbol{\beta}^TQ\boldsymbol{\alpha}-\boldsymbol{\alpha}^TQ\boldsymbol{\beta}]\\
=&\theta(1-\theta)\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\beta})^TQ(\boldsymbol{\alpha}-\boldsymbol{\beta})
\end{aligned}
$$

又因为 $0<\theta<1$, 且

$$
Q=\begin{pmatrix} y_1\mathbf{x}_1^T \\ \vdots \\ y_n\mathbf{x}_n^T \end{pmatrix}\begin{pmatrix} y_1\mathbf{x}_1 & \cdots & y_n\mathbf{x}_n \end{pmatrix}=C'C
$$

其中 C 为 n 阶实矩阵，所以 Q 是半正定的，从而可以得到

$$
F_D(\theta\boldsymbol{\alpha}+(1-\theta)\boldsymbol{\beta})-[\theta F_D(\boldsymbol{\alpha})+(1-\theta)F_D(\boldsymbol{\beta})]\geq 0
$$

证毕

4 Question 11.12

选取的 C 和 $\gamma=\frac{1}{\sigma^2}$ 如下

$$
C=10,80,100,200,500,1000
$$
$$
\gamma=0.00001,0.0001,0.002,0.01,0.04
$$

4.1 数据预处理

同书上的处理方式一样，对 0 值赋值 0.001，然后再取对数；同时设置好 C 和 γ 的值

```r
library(MASS)
library(dplyr)
library(rpart)
library(e1071)

wdbc <- read.table("wdbc.txt",sep=",",stringsAsFactors = TRUE)
wdbc[wdbc==0] <- 0.001
wdbc[,c(-1,-2)] <- apply(X = wdbc %>% select(-V1, -V2),
                         FUN = log,MARGIN = 2)
wdbc <- wdbc[,-1]

C <- c(10,80,100,200,500,1000)
gam <- c(0.00001,0.0001,0.001,0.01,0.04)
```

4.2 SVM 交叉验证

```
set.seed(1234)
N <- nrow(wdbc)
index <- sample(1:N, N, replace = F)

CV <- matrix(0, nrow = 6, ncol = 5) # 初始化，储存错误率
rownames(CV) <- c(10,80,100,200,500,1000)
colnames(CV) <- c(0.00001,0.0001,0.001,0.01,0.04)
CV <- as.data.frame(CV)

for (k in 1:6) {
  for (t in 1:5) {
    for (i in 1:10) {
    id <- (57*(i-1)+1):(57*i) # 分割数据集
    valid_index <- index[id]
    train_index <- index[-id]

    train <- wdbc[train_index,]
    valid <- wdbc[valid_index,]
    valid <- na.omit(valid) # 忽略NA

    out <- svm(V2~., data = train, kernel="radial",
        cost = C[k], gamma = gam[t])

    pred <- predict(out, valid)

    A <- table(valid$V2, pred)
    CV[k,t] <- CV[k,t] +(sum(A)-sum(diag(A)))/nrow(valid)
    }
  CV[k,t] <- CV[k,t]/10
  }
}
```

表 1: 交叉验证结果

C \ γ	0.00001	0.0001	0.002	0.01	0.04
10	0.15648496	0.04746241	0.02108396	0.02462406	0.02459273
80	0.05272556	0.02283835	0.02462406	0.02108396	0.02634712
100	0.04746241	0.01932957	0.02459273	0.01929825	0.02634712
200	0.02991855	0.01932957	0.02459273	0.03164160	0.02634712
500	0.02459273	0.01757519	0.02634712	0.02988722	0.02634712
1000	0.01932957	0.02108396	0.02634712	0.02988722	0.02634712

通过表1，很容易看到，当 $(C,\gamma)=(500,0.0001)$ 时，误判率最小，为 0.01757519. 进一步，我们使用 LDA 和分类树进行交叉验证，比较三种方法的优劣

4.3 LDA 交叉验证

```
err_lda <-0
for (i in 1:10) {
  id <- (57*(i-1)+1):(57*i)
  valid_index <- index[id]
  train_index <- index[-id]

  train <- wdbc[train_index,]
  valid <- wdbc[valid_index,]
  valid <- na.omit(valid)
  out <- lda(V2~., data = train)

  pred <- predict(out, valid)

  A <- table(valid$V2, pred$class)
  err_lda<- err_lda +sum(A)-sum(diag(A))
}
err_lda<- err_lda/569
```

经计算，使用 LDA 进行交叉验证得到误判率为 0.04570802

4.4 分类树交叉验证

```
err_t <-0
for (i in 1:10) {
  id <- (57*(i-1)+1):(57*i)
  valid_index <- index[id]
  train_index <- index[-id]

  train <- wdbc[train_index,]
  valid <- wdbc[valid_index,]
  valid <- na.omit(valid) # delete NA

  out <- rpart(V2~., method = "class",data = train)

  pred <- predict(out, valid, type="class")

  A <- table(valid$V2, pred)
  err_t<- err_t +(sum(A)-sum(diag(A)))/nrow(valid)
}
err_t <- err_t/10
```

分类树进行交叉验证得到的误判率为 0.08436717

4.5 总结

通过自行划分数据，使上述三种方法进行交叉验证所使用的训练集和测试集都相同，最后计算得到的结果是：SVM 的误判率最低，效果最好，其次是 LDA，最差的是分类树，且差异较明显；但是 SVM 要得到较好的结果，需要调参，会运行更长的时间