《机器学习》课程第 4 次作业

姓名：刘哲　　学号：2022103691

1　第5题

a

岭回归最优化问题的目标函数为

$$Obj_{ridge} = \sum_{i=1}^{2} \left(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}\right)^2 + \lambda \left(\beta_1^2 + \beta_2^2\right) \tag{1}$$

b

令

$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} = \begin{bmatrix} a & a \\ -a & -a \end{bmatrix}, \qquad Y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} b \\ -b \end{bmatrix} \tag{2}$$

其中，$a \neq 0$，$b \neq 0$。岭回归的参数估计应满足使目标函数 Obj_{ridge} 达到最小。令 Obj_{ridge} 分别对β_0、β_1和β_2求偏导，使偏导数等于零，得到

$$\begin{cases} \frac{\partial Obj_{ridge}}{\partial \beta_0} = -2\left(b - \beta_0 - a\beta_1 - a\beta_2\right) - 2\left(b - \beta_0 + a\beta_1 + a\beta_2\right) = 0 \\ \frac{\partial Obj_{ridge}}{\partial \beta_1} = -2a\left(b - \beta_0 - a\beta_1 - a\beta_2\right) + 2a\left(b - \beta_0 + a\beta_1 + a\beta_2\right) + 2\lambda\beta_1 = 0 \\ \frac{\partial Obj_{ridge}}{\partial \beta_2} = -2a\left(b - \beta_0 - a\beta_1 - a\beta_2\right) + 2a\left(b - \beta_0 + a\beta_1 + a\beta_2\right) + 2\lambda\beta_2 = 0 \end{cases} \tag{3}$$

解得

$$2\lambda\beta_1 = 2\lambda\beta_2 \tag{4}$$

因为 $\lambda \neq 0$，所以有 $\hat{\beta}_1 = \hat{\beta}_2$。

c

LASSO回归最优化问题的目标函数为

$$Obj_{LASSO} = \sum_{i=1}^{2} \left(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}\right)^2 + \lambda \left(|\beta_1| + |\beta_2|\right) \tag{5}$$

d

LASSO回归的参数估计应满足使目标函数 Obj_{LASSO} 达到最小。令 Obj_{LASSO} 分别对β_0、β_1和β_2求偏导，使偏导数等于零，得到

$$\begin{cases} \frac{\partial Obj_{LASSO}}{\partial \beta_0} = -2\left(b-\beta_0-a\beta_1-a\beta_2\right)-2\left(b-\beta_0+a\beta_1+a\beta_2\right)=0 \\ \frac{\partial Obj_{LASSO}}{\partial \beta_1} = -2a\left(b-\beta_0-a\beta_1-a\beta_2\right)+2a\left(b-\beta_0+a\beta_1+a\beta_2\right)+\lambda\frac{\partial|\beta_1|}{\partial\beta_1}=0 \\ \frac{\partial Obj_{LASSO}}{\partial \beta_2} = -2a\left(b-\beta_0-a\beta_1-a\beta_2\right)+2a\left(b-\beta_0+a\beta_1+a\beta_2\right)+\lambda\frac{\partial|\beta_2|}{\partial\beta_2}=0 \end{cases} \tag{6}$$

因为 $\frac{\partial|\beta_1|}{\partial\beta_1}$ 和 $\frac{\partial|\beta_2|}{\partial\beta_2}$ 的值取决于 β_1 和 β_2 的正负号，所以 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 不唯一，且满足 $\left|\hat{\beta}_1\right|=\left|\hat{\beta}_2\right|$。

2 第9题

a

将 College 数据集中的 Private 变量转化为哑变量，令1表示Yes，0表示No。按 3:1 的比例将数据集划分为训练集和测试集，其中训练集包含583行数据，测试集包含194行数据。

b

通过逐步选择法筛选模型的解释变量，结果如下：

```
##
## Call:
## lm(formula = Apps ~ Accept + Top10perc + Expend + Outstate +
##     Enroll + Top25perc + Private + Grad.Rate + Room.Board + PhD,
##     data = college.train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4892.7  -439.3   -31.5   341.5  7762.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -92.09498  317.72417  -0.290 0.772030
## Accept         1.59928    0.04630  34.539  < 2e-16 ***
## Top10perc     52.90957    6.70444   7.892 1.53e-14 ***
## Expend         0.08475    0.01401   6.051 2.60e-09 ***
## Outstate      -0.10791    0.02200  -4.904 1.23e-06 ***
## Enroll        -0.65861    0.13671  -4.818 1.86e-06 ***
## Top25perc    -16.06054    5.31078  -3.024 0.002605 **
## Private     -581.38858  161.01377  -3.611 0.000332 ***
## Grad.Rate      8.42036    3.29619   2.555 0.010890 *
## Room.Board     0.13713    0.05596   2.450 0.014567 *
## PhD           -8.13331    3.72712  -2.182 0.029501 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1070 on 572 degrees of freedom
## Multiple R-squared:  0.9266, Adjusted R-squared:  0.9254
## F-statistic: 722.5 on 10 and 572 DF,  p-value: < 2.2e-16
```

选取显著性水平最高的5个解释变量 Accept、Top10perc、Expend、Outstate、 Enroll 建立线性回归模型，结果如下：

```
##
## Call:
## lm(formula = Apps ~ Accept + Top10perc + Expend + Outstate +
##     Enroll, data = college.train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5113.8  -459.6    -6.7   309.7  7458.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -490.17524  145.73590  -3.363 0.000821 ***
## Accept         1.61352    0.04641  34.770  < 2e-16 ***
## Top10perc     35.69107    3.71731   9.601  < 2e-16 ***
## Expend         0.09573    0.01379   6.940 1.05e-11 ***
## Outstate      -0.11359    0.01761  -6.449 2.39e-10 ***
## Enroll        -0.61441    0.13242  -4.640 4.32e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1096 on 577 degrees of freedom
## Multiple R-squared:  0.9224, Adjusted R-squared:  0.9217
## F-statistic:  1372 on 5 and 577 DF,  p-value: < 2.2e-16
```

其中模型参数均显著。用该模型对测试集数据进行预测，得到测试误差为1232.63。

c

使用10折交叉验证法确定岭回归的最佳 $\lambda = 368.2153$，建立岭回归模型，系数估计结果如下：

```
## 18 x 1 sparse Matrix of class "dgCMatrix"
##                       s0
## (Intercept) -1.286037e+03
## Private     -4.969173e+02
## Accept       1.033328e+00
## Enroll       5.026189e-01
## Top10perc    2.804957e+01
## Top25perc   -8.025723e-01
## F.Undergrad  5.118720e-02
## P.Undergrad  3.058901e-02
## Outstate    -4.542984e-02
## Room.Board   1.846988e-01
## Books        1.817234e-01
## Personal    -8.251005e-02
## PhD         -1.838533e+00
## Terminal    -6.072487e+00
## S.F.Ratio    1.010473e+01
## perc.alumni -3.771864e+00
## Expend       8.273436e-02
## Grad.Rate    1.045583e+01
```

用该模型对测试集数据进行预测，得到测试误差为3763.05。

d

使用10折交叉验证法确定LASSO回归的最佳 $\lambda = 29.1804$，建立LASSO回归模型，系数估计结果如下：

```
## 18 x 1 sparse Matrix of class "dgCMatrix"
##                       s0
## (Intercept) -4.785867e+02
## Private     -4.152613e+02
## Accept       1.473349e+00
## Enroll      -2.544750e-01
## Top10perc    3.628616e+01
## Top25perc   -4.183875e+00
## F.Undergrad  .
## P.Undergrad  .
## Outstate    -7.242817e-02
## Room.Board   1.061217e-01
## Books        .
## Personal    -9.168276e-03
## PhD         -2.395752e+00
## Terminal    -4.980854e+00
## S.F.Ratio    .
## perc.alumni  .
## Expend       7.957515e-02
## Grad.Rate    5.465301e+00
```

用该模型对测试集数据进行预测，得到测试误差为1405.266。