Exercises ISLR – Ch.6

Marcelo Previato Simoes Nº 2367070
22/09/2025

Conceptual

Exercises: 7

Ex.7) Bayesian View – Likelihood (a)

(a) Suppose that $y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$ where $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed from a $N(0, \sigma^2)$ distribution. Write out the likelihood for the data.

a) $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j \cdot x_{ij} + \epsilon_i$ $\epsilon_i \perp\!\!\!\perp \epsilon_j,\ i \neq j$ & $\epsilon_i \sim N(0, \sigma^2)$ likelihood function

$$\begin{cases} E(y_i \mid X, \beta) = \beta_0 + \sum_{j=1}^{p} \beta_j \cdot x_{ij} \\ Var(y_i \mid X, \beta) = Var(\epsilon_i) = \sigma^2 \end{cases} \Rightarrow y_i \mid X, \beta \sim N\left(\beta_0 + \sum_{j=1}^{p} \beta_j \cdot x_{ij},\ \sigma^2\right)$$

$\underline{X}$: vector notation

calculating the likelihood function, assuming $y_1, \ldots, y_n$ iid

$$L(\underline{y} \mid \underline{X}, \beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp\left\{ -\frac{1}{2} \frac{\left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \cdot x_{ij}\right)^2}{\sigma^2} \right\}$$

$$\boxed{L(\underline{y} \mid \underline{X}, \beta) = (2\pi\sigma^2)^{-n/2} \cdot \exp\left\{ \frac{-1}{2\sigma^2} \cdot \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \cdot x_{ij} \right)^2 \right\}}$$

Ex.7) Bayesian View – Double Exp. Prior (b)

(b) Assume the following prior for β: $\beta_1, \ldots, \beta_p$ are independent and identically distributed according to a double-exponential distribution with mean 0 and common scale parameter b: i.e. $p(\beta) = \frac{1}{2b}\exp(-|\beta|/b)$. Write out the posterior for β in this setting.

b) Priori $\beta_j \sim$ Double exponencial $(0,b)$ $\Rightarrow$ $p(\beta) = \frac{1}{2b} \cdot \exp(-|\beta|/b)$

The joint distribution of $\beta_1, \ldots, \beta_p$, assuming iid:

$$p(\underline{\beta}) = \prod_{j=1}^{p} \frac{1}{2b} \cdot \exp\left(-\frac{|\beta_j|}{b}\right) = \frac{1}{(2b)^p} \cdot \exp\left(-\frac{1}{b} \cdot \sum_{j=1}^{p} |\beta_j|\right)$$

Applying the Bayes Theorem to find the posterior

$$p(\underline{\beta}|\underline{X},Y) \propto \underbrace{L(\underline{Y}|\underline{X},\underline{\beta})}_{(a)} \cdot \underbrace{p(\underline{\beta}|\underline{X})}_{=p(\underline{\beta})} \text{ assuming } \underline{\beta} \perp\!\!\!\perp X$$

$$\boxed{p(\underline{\beta}|\underline{X},Y) \propto (2\pi\sigma^2)^{-n/2} \cdot (2b)^{-p} \cdot \exp\left\{-\frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij}\right)^2 - \frac{1}{b} \cdot \sum_{j=1}^{p} |\beta_j|\right\}}$$

Ex.7) Bayesian View – Mode & Lasso (c)

(c) Argue that the lasso estimate is the *mode* for β under this posterior distribution.

Lasso: minimize $\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|.$

c) $\hat{\beta}_{Lasso}$ is mode $(P(\beta|X,Y))$ The mode can be find by maximizing $P(\beta|X,Y)$, which means maximizing the exponent, or minimize the expression :

$$\hat{\beta}_{mode} = \underset{}{\arg\min}\left\{ \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \cdot \sum_{j=1}^{p}|\beta_j| \right\}$$

$\Leftrightarrow$ LASSO minimization problem

$$\therefore \boxed{\hat{\beta}_{Lasso} = \text{mode}(P(\beta|X,Y))}$$

Ex.7) Bayesian View – Normal Prior (d)

(d) Now assume the following prior for β: $\beta_1, \dots, \beta_p$ are independent and identically distributed according to a normal distribution with mean zero and variance c. Write out the posterior for β in this setting.

d) Priori $B_j \sim N(0,c)$ $\quad P(B_j) = \frac{1}{\sqrt{2\pi c}} \cdot \exp\left\{-\frac{1}{2} \cdot \frac{B_j^2}{c}\right\}$

The joint distribution $P(\underline{B})$, assuming $B_1, \dots, B_p$ iid

$$P(\underline{B}) = \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi c}} \cdot \exp\left\{-\frac{1}{2} \cdot \frac{B_j^2}{c}\right\}$$

$$= (2\pi c)^{-p/2} \cdot \exp\left\{-\frac{1}{2c} \cdot \sum_{j=1}^{p} B_j^2\right\}$$

Applying the Bayes theorem to find the posterior:

$$P(\underline{B}|\underline{X},\underline{Y}) \propto L(\underline{Y}|\underline{X},\underline{B}) \cdot \underbrace{P(B|X)}_{=P(B)}$$

$$\boxed{P(\underline{B}|\underline{X},\underline{Y}) \propto (2\pi\sigma^2)^{-n/2} \cdot (2\pi c)^{-p/2} \cdot \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(Y_i - B_0 - \sum_{j=1}^{p} B_j \cdot X_{ij}\right)^2 - \frac{1}{2c} \cdot \sum_{j=1}^{p} B_j^2\right\}}$$

since both the likelihood and the priori are normal, the posteriori is normal for a known variance (Normal-normal conjugation)

Ex.7) Bayesian View – Mode, Mean & Ridge (e)

(e) Argue that the ridge regression estimate is both the *mode* and the *mean* for β under this posterior distribution.

Ridge: minimize $$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

e) $\hat{\beta}_{ridge} = mean(P(\underline{\beta}|\underline{X},\underline{Y})) = mode(P(\underline{\beta}|\underline{X},\underline{Y}))$

The mode $(P(\underline{\beta}|\underline{X},\underline{Y}))$ is the value that maximize the posterior, which is the same that minimizes the expression

$$\hat{\beta}_{mode} = \arg\min\left\{\sum_{n=1}^{n}(Y_i-\beta_0-\sum_{j=1}^{p}\beta_j \cdot X_{ij})^2 + \lambda \cdot \sum_{j=1}^{p}\beta_j^2\right\}$$

$\Leftrightarrow$ Ridge minimization problem

$$\boxed{\hat{\beta}_{ridge} = mode(P(\underline{\beta}|\underline{X},\underline{Y}))}$$

Since the posterior is a normal distribution, mean = mode

$$\boxed{\hat{\beta}_{ridge} = mode(P(\underline{\beta}|\underline{X},\underline{Y})) = mean(P(\underline{\beta}|\underline{X},\underline{Y}))}$$

Applied

Exercise 11 – Boston dataset

Ex.11) Crime Rate Regression Models

Full predictors: 13 predictors

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  16.879384   8.215223   2.055  0.04068 *  
zn            0.039526   0.021603   1.830  0.06819 .  
indus        -0.066092   0.102269  -0.646  0.51855    
chas         -0.359512   1.359197  -0.265  0.79155    
nox         -11.391514   6.008571  -1.896  0.05882 .  
rm            0.140098   0.691876   0.202  0.83965    
age           0.004565   0.020183   0.226  0.82121    
dis          -0.925751   0.317961  -2.912  0.00383 ** 
rad           0.539167   0.105247   5.123 5.05e-07 ***
tax          -0.001345   0.006330  -0.213  0.83182    
ptratio      -0.251299   0.219012  -1.147  0.25202    
black        -0.007466   0.004083  -1.828  0.06837 .  
lstat         0.138404   0.086771   1.595  0.11163    
medv         -0.148884   0.066666  -2.233  0.02618 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.19 on 340 degrees of freedom
Multiple R-squared:  0.4709,	Adjusted R-squared:  0.4506 
F-statistic: 23.27 on 13 and 340 DF,  p-value: < 2.2e-16
```

MSE – Test Dataset

```
[1] 48.97129
```

Ex.11) Crime Rate Regression Models

Forward Stepwise – BIC: 2 predictors

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.15947    0.67206  -6.189 1.69e-09 ***
rad          0.50743    0.04405  11.518  < 2e-16 ***
lstat        0.23315    0.05422   4.300 2.22e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.264 on 351 degrees of freedom
Multiple R-squared:  0.4407,	Adjusted R-squared:  0.4375 
F-statistic: 138.3 on 2 and 351 DF,  p-value: < 2.2e-16
```

MSE – Test Dataset

```
[1] 51.84761
```

Ex.11) Crime Rate Regression Models

Backward Stepwise – BIC: 4 predictors

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.89923    1.53327   3.195  0.00152 ** 
zn           0.05177    0.01997   2.593  0.00991 ** 
dis         -0.72881    0.22543  -3.233  0.00134 ** 
rad          0.49013    0.04602  10.651  < 2e-16 ***
medv        -0.16820    0.04014  -4.190 3.53e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.226 on 349 degrees of freedom
Multiple R-squared:  0.4506,	Adjusted R-squared:  0.4443 
F-statistic: 71.56 on 4 and 349 DF,  p-value: < 2.2e-16
```

MSE – Test Dataset

```
[1] 50.44479
```

Ex.11) Crime Rate Regression Models

Ridge with Cross Validation: 8 out of 13 predictors reduced towards 0

	Ridge	OLS
(Intercept)	9.318869511	16.879384
zn	0.030951874	0.039526
indus	-0.078975739	-0.066092
chas	-0.438115276	-0.359512
nox	-6.467581357	-11.391514
rm	0.094753774	0.140098
age	0.004213388	0.004565
dis	-0.683172006	-0.925751
rad	0.390201979	0.539167
tax	0.004881571	-0.001345
ptratio	-0.114310672	-0.251299
black	-0.008290456	-0.007466
lstat	0.145087084	0.138404
medv	-0.103703109	-0.148884

MSE – Test Dataset

[1] 49.96527

Ex.11) Crime Rate Regression Models

Lasso with Cross Validation: all predictors reduced towards zero, 3 of 13 eliminated

	Lasso	OLS
(Intercept)	11.823219576	16.879384
zn	0.032730795	0.039526
indus	-0.050384551	-0.066092
chas	-0.274353526	-0.359512
nox	-7.031684679	-11.391514
rm	.	0.140098
age	.	0.004565
dis	-0.724610951	-0.925751
rad	0.494368414	0.539167
tax	.	-0.001345
ptratio	-0.141945061	-0.251299
black	-0.007385504	-0.007466
lstat	0.134998778	0.138404
medv	-0.111911665	-0.148884

MSE – Test Dataset

[1] 49.45321

Ex.11) Crime Rate Regression Models

Principal Components: 5 components

	crim
zn	0.55318815
indus	0.66942812
chas	-0.21981627
nox	0.49299837
rm	0.15032635
age	0.01629888
dis	-0.03084570
rad	1.60206772
tax	1.53063157
ptratio	0.50106858
black	-1.46788298
lstat	0.44810566
medv	-0.46831414

MSE – Test Dataset

[1] 53.95436

Ex.11) Crime Rate Regression Models

Full comparison based on MSE test data

Model <chr>	Test_MSE <dbl>
OLS: 13 predictors	48.97129
Lasso: 10 predictors	49.45321
Ridge: 13 predictors	49.96527
Backward: 4 predictors	50.44479
Forward: 2 predictors	51.84761
PCR: 5 components	53.95436

- **OLS with all predictors:** best performance even in test dataset, which means that the model is not suffering of overfitting despite using all predictors

- **Models with fewer predictors:** some of them presented a very good performance and could be selected to reduce the risk of overfitting in a different test data