环境学院|环 82 班|何恺洵|2018010412|数据科学导论

HW6 Linear Model (II)

1. Given a Gaussian linear regression model, Maximum likelihood estimation of $\boldsymbol{w}$ under Gaussian noise assumption is equivalent to _________________________. Please prove it.

the least square loss minimization

proof:

For an i.i.d sample $D=\{(\boldsymbol{x}_1,y_1),\dots,(\boldsymbol{x}_N,y_N)\}$ *, assuming that* $p(\boldsymbol{x}_i)=\frac{1}{N}$ *, likelihood of D is*

$$p(D;\widehat{\boldsymbol{\theta}})=\prod_{i=1}^{N}p(\boldsymbol{x}_i,y_i;\ \widehat{\boldsymbol{\theta}})=\frac{1}{N}\prod_{i=1}^{N}p(y_i\mid\boldsymbol{x}_i,\widehat{\boldsymbol{\theta}})$$

$$\ln p(D;\widehat{\boldsymbol{\theta}})=\sum_{i=1}^{N}\ln p(y_i\mid\boldsymbol{x}_i,\widehat{\boldsymbol{\theta}})-\ln N$$

For a Gauss linear regression model $h(\boldsymbol{x}_i)=\boldsymbol{w}^T\boldsymbol{x}_i$*, according to Gauss noise assumption,* $\varepsilon_i=y_i-\boldsymbol{w}^T\boldsymbol{x}_i$ *follows Gaussian distribution.* $\varepsilon_i\sim N(0,\sigma^2)$*.*

Therefore $y_i\mid\boldsymbol{x}_i,\boldsymbol{w}\sim N(\boldsymbol{w}^T\boldsymbol{x}_i,\sigma^2)$

$$p(y_i\mid\boldsymbol{x}_i,\boldsymbol{w})=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(y_i-\boldsymbol{w}^T\boldsymbol{x}_i)^2\right\}$$

So the log-likelihood in this situation is

$$\ln p(D;\widehat{\boldsymbol{\theta}})=\sum_{i=1}^{N}\ln p(y_i\mid\boldsymbol{x}_i,\boldsymbol{w})-\ln N=-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\boldsymbol{w}^T\boldsymbol{x}_i)^2-\frac{N}{2}\ln 2\pi\sigma^2-\ln N$$

Maximum likelihood estimation of $\boldsymbol{w}$ *is*

$$\arg\max-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\boldsymbol{w}^T\boldsymbol{x}_i)^2=\arg\min\sum_{i=1}^{N}(y_i-\boldsymbol{w}^T\boldsymbol{x}_i)^2$$

Notice that $\min\sum_{i=1}^{N}(y_i-\boldsymbol{w}^T\boldsymbol{x}_i)^2$ *is the least square loss minimization.* ■

2. Given a Laplacian linear regression model, Maximum likelihood estimation of $\boldsymbol{w}$ under Laplacian noise assumption is equivalent to ___________________________. Please prove it.

least absolute deviations

proof:

For a Laplacian linear regression model $h(\boldsymbol{x}_i)=\boldsymbol{w}^T\boldsymbol{x}_i$ *, according to Laplacian noise assumption,* $\varepsilon_i=y_i-\boldsymbol{w}^T\boldsymbol{x}_i$ *follows Laplacian distribution.* $\varepsilon_i\sim Laplace(0,b)$*.*

Therefore $y_i\mid\boldsymbol{x}_i,\boldsymbol{w}\sim Laplace(\boldsymbol{w}^T\boldsymbol{x}_i,b)$

$$p(y_i\mid\boldsymbol{x}_i,\boldsymbol{w})=\frac{1}{2b}\exp\left\{-\frac{1}{b}|y_i-\boldsymbol{w}^T\boldsymbol{x}_i|\right\}$$

So the log-likelihood in this situation is

$$\ln p\big(D;\widehat{\boldsymbol{\theta}}\big)=\sum_{i=1}^{N}\ln p(y_i \mid \boldsymbol{x}_i,\boldsymbol{w})-\ln N=-\frac{1}{b}\sum_{i=1}^{N}|y_i-\boldsymbol{w}^T\boldsymbol{x}_i|-N\ln 2b-\ln N$$

Maximum likelihood estimation of $\boldsymbol{w}$ *is*

$$\arg\max-\frac{1}{b}\sum_{i=1}^{N}|y_i-\boldsymbol{w}^T\boldsymbol{x}_i|=\arg\min\sum_{i=1}^{N}|y_i-\boldsymbol{w}^T\boldsymbol{x}_i|$$

Notice that $\min\sum_{i=1}^{N}|y_i-\boldsymbol{w}^T\boldsymbol{x}_i|$ *is the least absolute deviations.* ■

3. Given a linear regression model, please write down the Tikhonov Form and Ivanov Form of Ridge Regression, and these two forms of Lasso Regression as well.

According to PPT:

Ridge regression (Tikhonov Form):

The ridge regression solution for regularization parameter $\lambda \geq 0$ is

$$\widehat{\boldsymbol{w}}=\arg\min\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i-y_i)^2+\lambda\|\boldsymbol{w}\|_2^2,\qquad \boldsymbol{w}\in R^d$$

Where $\|\boldsymbol{w}\|_2^2=\boldsymbol{w}_1^2+\cdots+\boldsymbol{w}_d^2$ is the square of l_2-norm

Ridge regression (Ivanov Form):

The ridge regression solution for complexity parameter $r \geq 0$ is

$$\widehat{\boldsymbol{w}}=\arg\min\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i-y_i)^2,\qquad \|\boldsymbol{w}\|_2^2\leq r$$

Lasso Regression (Tikhonov Form):

The lasso regression solution for regularization parameter $\lambda \geq 0$ is

$$\widehat{\boldsymbol{w}}=\arg\min\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i-y_i)^2+\lambda\|\boldsymbol{w}\|_1,\qquad \boldsymbol{w}\in R^d$$

Where $\|\boldsymbol{w}\|_1=|\boldsymbol{w}_{(1)}|+\cdots+|\boldsymbol{w}_{(d)}|$ is the square of l_1-norm

Lasso regression (Ivanov Form):

The lasso regression solution for complexity parameter $r \geq 0$ is

$$\widehat{\boldsymbol{w}}=\arg\min\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i-y_i)^2,\qquad \|\boldsymbol{w}\|_1\leq r$$

4. By adding a Ridge Regression in the linear regression model of Question 4 in hw5-linear-model, can we get a lower generalization error? If yes, use cross validation to attain the best regularization parameter λ, whose possible values are [1.e-06, 1.e-05, 1.e-04, 1.e-03, 1.e-02, 1.e-01, 1.e+00, 1.e+01, 1.e+02, 1.e+03, 1.e+04, 1.e+05, 1.e+06]. If no, please explain why. See the tutorial of linear model in sklearn: https://scikit-learn.org/stable/modules/linear_model.html if you need some help.

It's believed that a lower generation error will be achieved with a Ridge Regression. By using it, we can have smaller $\boldsymbol{w}$ and avoid overfitting.

In the program, I use 'sklearn' to find the best regularization parameter λ. For the cross validation, it's necessary to decide how much data should be held out for validation. Considering that the amount of data is relatively small and that we must use enough data to train the model, the 'cv' attribute is decided to be 10.

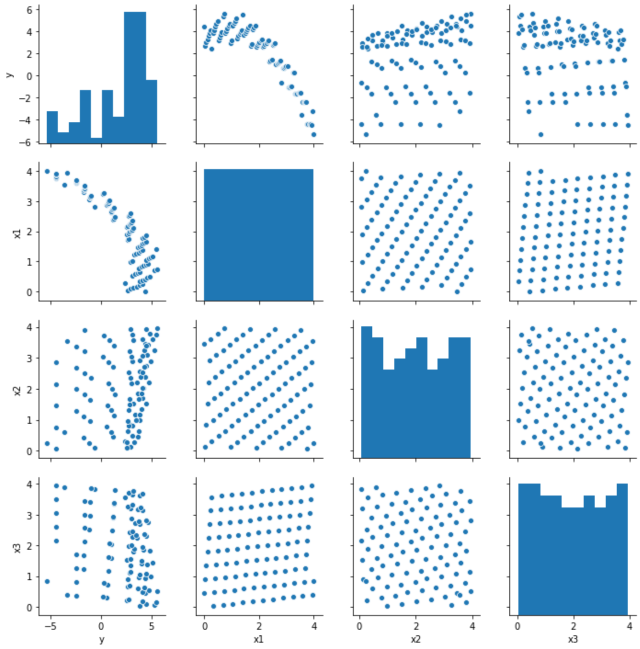
And finally the result is $\boldsymbol{\lambda = 10}$, which also indicates a Ridge Regression is effective.

The program is attached.

5. By adding a Lasso Regression in the linear regression model of Question 4 in hw5-linear-model, can we get a lower generalization error? If yes, use cross validation to attain the best regularization parameter λ, whose possible values are [1.e-06, 1.e-05, 1.e-04, 1.e-03, 1.e-02, 1.e-01, 1.e+00, 1.e+01, 1.e+02, 1.e+03, 1.e+04, 1.e+05, 1.e+06]. If no, please explain why. See the tutorial of linear model in sklearn: https://scikit-learn.org/stable/modules/linear_model.html if you need some help.

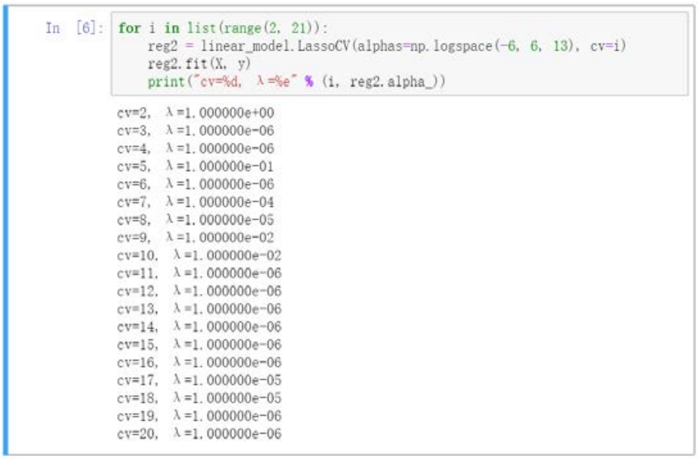
I think the answer is **no**.

A Lasso Regression wants to reduce features that have few contributions to the result and maintain the important ones. However, this time we have only three features and we don't need to reduce features. Also, we can see from the pairplot (maybe it is not so clear) that all these three features (x_1, x_2, x_3) have something to do with y. If we use Lasso Regression Model, maybe features x_2 and x_3 will be left out and we can not have so good a model as before.



pairplot for x_1, x_2, x_3, y

We can prove this conclusion by calculating λ. When choosing different values of 'cv', λ is always very small. In most cases, it's 1.e-06, which is the smallest one of the given possible values.



```
for i in list(range(2, 21)):
    reg2 = linear_model.LassoCV(alphas=np.logspace(-6, 6, 13), cv=i)
    reg2.fit(X, y)
    print("cv=%d, λ=%e" % (i, reg2.alpha_))
```

```
cv=2,  λ=1.000000e+00
cv=3,  λ=1.000000e-06
cv=4,  λ=1.000000e-06
cv=5,  λ=1.000000e-01
cv=6,  λ=1.000000e-06
cv=7,  λ=1.000000e-04
cv=8,  λ=1.000000e-05
cv=9,  λ=1.000000e-02
cv=10,  λ=1.000000e-02
cv=11,  λ=1.000000e-06
cv=12,  λ=1.000000e-06
cv=13,  λ=1.000000e-06
cv=14,  λ=1.000000e-06
cv=15,  λ=1.000000e-06
cv=16,  λ=1.000000e-06
cv=17,  λ=1.000000e-05
cv=18,  λ=1.000000e-05
cv=19,  λ=1.000000e-06
cv=20,  λ=1.000000e-06
```

best λ