Linear Regression

Regression

- Classification model output value is discrete.
- Regression model output value is continuous.

gender

house price

crop yield

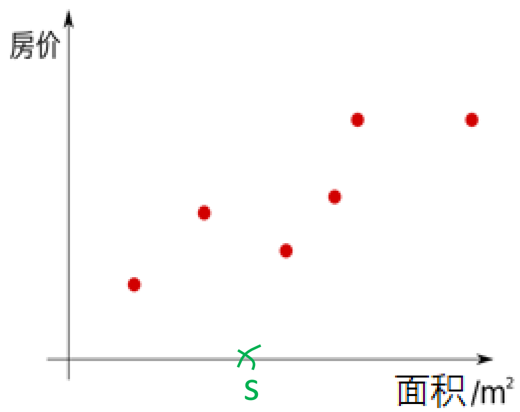
height

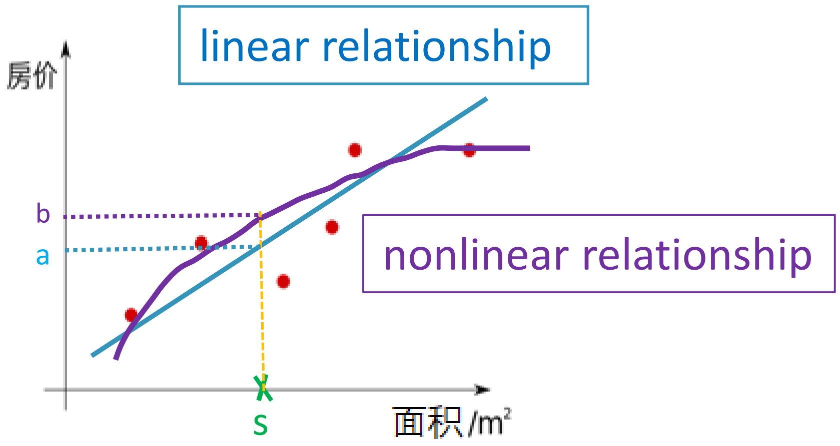
movie type

watermelon: good or bad

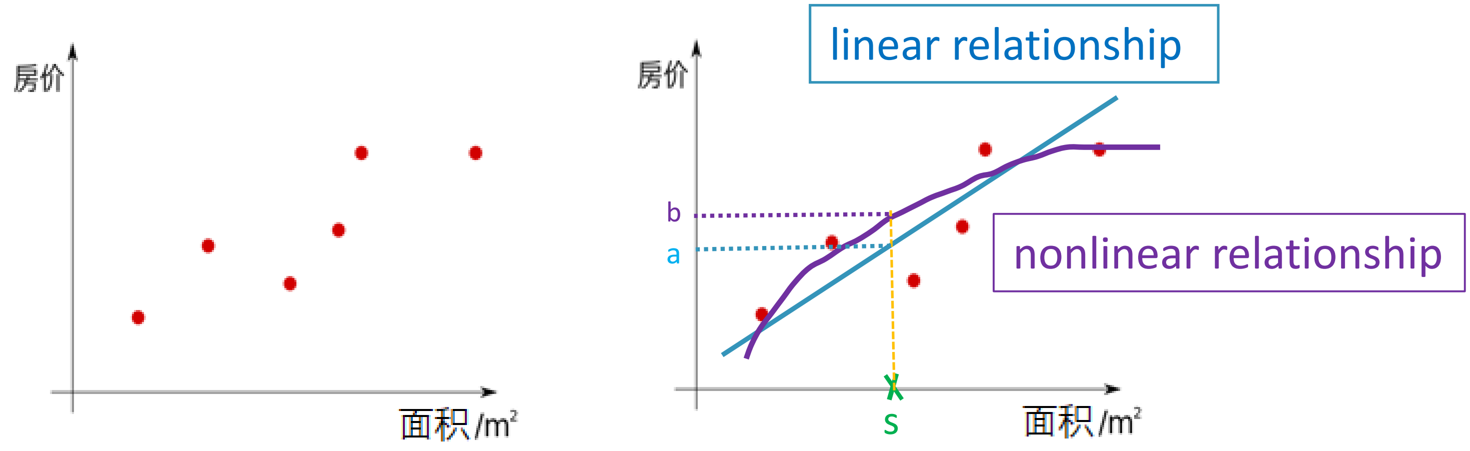
regression	classification
house price crop yield height …	gender:male or female movie type good or bad watermelon …

Regression


房价
S
面积/m²


linear relationship
房价
b
a
nonlinear relationship
S
面积/m²

Regression



- Regression estimates the relationship between input values and output value,and establishes a mathematical model in order to accurately predict the output value of new sample.
- Regression is a supervised learning question.

Regression

Regression

the number of input values

- single regression(单元回归)
- multiple regression(多元回归)

Regression

the relationship of input values and output value

- linear regression
- nonlinear regression

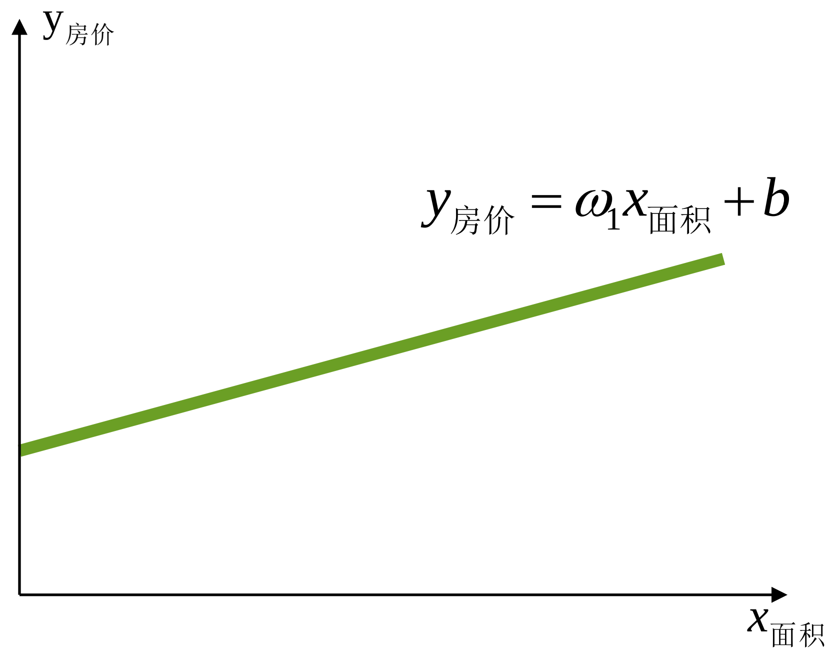
Outline

- linear regression model
  - linear model
  - least square method --> basic linear regression
  - ridge regression(岭回归)
  - lasso regression

Outline

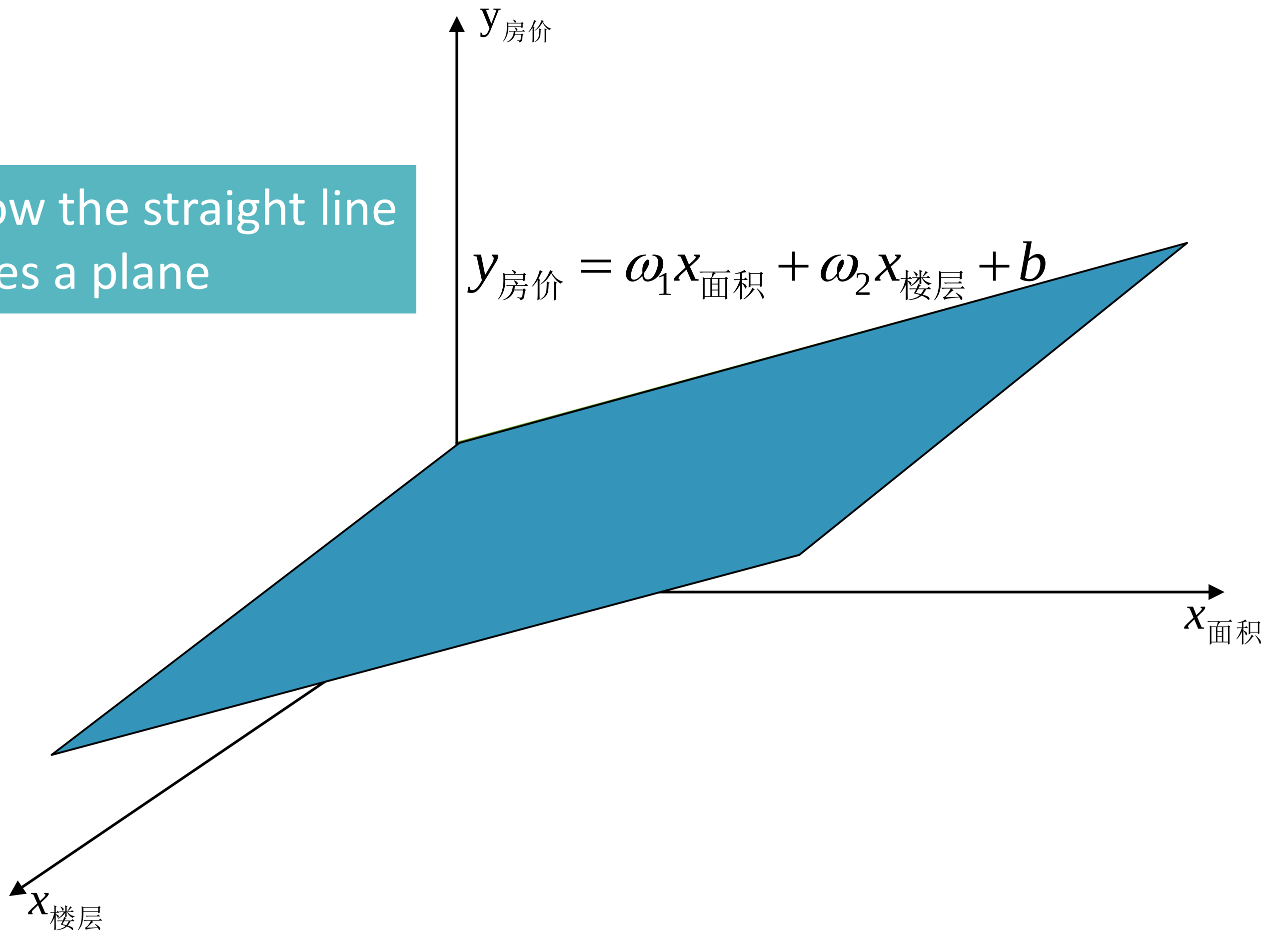
- linear regression model
  - linear model
  - least square method --> basic linear regression
  - ridge regression(岭回归)
  - lasso regression

linear regression model



linear regression model

Note how the straight line becomes a plane



linear regression model

two-dimension: $y_{房价} = \omega_1 x_{面积} + b$

three-dimension: $y_{房价} = \omega_1 x_{面积} + \omega_2 x_{楼层} + b$

four-dimension: $y_{房价} = \omega_1 x_{面积} + \omega_2 x_{楼层} + \omega_3 x_{地段} + b$

linear regression model

general model: $\boxed{f(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d + b}$

$x_1, x_2, \ldots, x_d$: the feature of sample

$w_1, w_2, \ldots, w_d$: weight, represent the importance of corresponding feature

$$f_{房价}(x) = 0.2 \cdot x_{楼层} + 0.5 \cdot x_{地段} + 0.3 \cdot x_{面积} + 1$$

importance: $x_{地段} > x_{面积} > x_{楼层}$

linear model's vector formal

general model: $f(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d + b$

vector formal: $f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x} + b$

$$\boldsymbol{w} = (w_1; w_2; \ldots; w_d)$$

$$\boldsymbol{x} = (x_1; x_2; \ldots; x_d)$$

linear model's vector formal

training set:

S1: 楼层=6，地段=5，面积=120。房价=100
S2: 楼层=3，地段=1，面积=100。房价=200
S3: 楼层=16，地段=8，面积=200。房价=120

$$\begin{cases} 6\omega_1 + 5\omega_2 + 120\omega_3 = 100 \\ 3\omega_1 + \omega_2 + 100\omega_3 = 200 \\ 16\omega_1 + 8\omega_2 + 200\omega_3 = 120 \end{cases} \Rightarrow \begin{aligned} \omega_1 &= a \\ \omega_2 &= b \\ \omega_3 &= c \end{aligned}$$

$$f_{房价}(x) = a \cdot x_{楼层} + \mathrm{b} \cdot x_{地段} + \mathrm{c} \cdot x_{面积}$$

- Solve weigt vector $\boldsymbol{w}$ according to training set.

Which model is better?

$$f_{房价}(x) = 0.2 \cdot x_{楼层} + 0.5 \cdot x_{地段} + 0.3 \cdot x_{面积} + 1$$

$$g_{房价}(x) = 0.1 \cdot x_{楼层} + 0.4 \cdot x_{地段} + 0.5 \cdot x_{面积} + 3$$

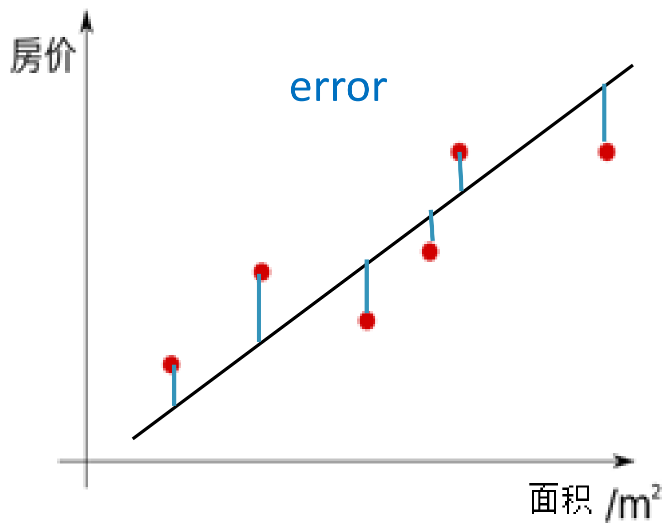
a known data:楼层=6，地段=5，面积=120，房价=100
model f predictive value: 0.2*6+0.5*5+0.3*120+1=40.7
model g predictive value: 0.1*6+0.4*5+0.5*120+3=65.6

model f error: 100-40.7=59.3
model g error: 100-65.6=34.4
34.4<59.3
model g is better

- the model which can minimize error between model predictive value and true value.

least square method

•In regression task, the least square method is often used measuring error between model predictive value and true value.



$$\text{error}=\sum_{i=1}^{m}\left(f\left(x_i\right)-y_i\right)^2$$

predictive value

true value

least square method:

$$min\sum_{i=1}^{m}\left(f\left(x_i\right)-y_i\right)^2$$

basic linear regression

- solve w and b

single linear regression:

model: $f(x_i) = wx_i + b$

loss function: $E_{(w,b)} = \sum_{i=1}^{m} (f(x_i) - y_i)^2$

$$= \sum_{i=1}^{m} (y_i - wx_i - b)^2$$

$$min\, E_{(w,b)}$$

basic linear regression

- solve $\boldsymbol{w}$ and $\boldsymbol{b}$

single linear regression: $E_{(w,b)} = \sum_{i=1}^{m}(y_i - wx_i - b)^2$

$$\frac{\partial E_{(w,b)}}{\partial w} = 2\left(w\sum_{i=1}^{m}x_i^2 - \sum_{i=1}^{m}(y_i - b)\,x_i\right) = 0$$

$$\frac{\partial E_{(w,b)}}{\partial b} = 2\left(mb - \sum_{i=1}^{m}(y_i - wx_i)\right) = 0$$

$$w = \frac{\sum_{i=1}^{m} y_i(x_i - \bar{x})}{\sum_{i=1}^{m} x_i^2 - \frac{1}{m}\left(\sum_{i=1}^{m} x_i\right)^2}$$

$$b = \frac{1}{m}\sum_{i=1}^{m}(y_i - wx_i) \qquad \bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

new sample

model: $f(x_i) = wx_i + b$

dataset: $D = \{(x_i, y_i)\}_{i=1}^{m}$

basic linear regression

•solve $\boldsymbol{w}$ and b

multiple linear regression:

model: $f(\boldsymbol{x}_i) = w_1 x_{i1} + w_2 x_{i2} + \ldots + w_d x_{id} + b$

$\Updownarrow$

vector model: $f(\boldsymbol{x}_i) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i + b$

$\Updownarrow * m$

$$f(\boldsymbol{x}) = \mathbf{X}\hat{\boldsymbol{w}}$$

$\hat{\boldsymbol{w}} = (\boldsymbol{w}; b)$ $(d+1)$ column vector

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1d} & 1 \\ x_{21} & x_{22} & \ldots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \ldots & x_{md} & 1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} & 1 \\ \boldsymbol{x}_2^{\mathrm{T}} & 1 \\ \vdots & \vdots \\ \boldsymbol{x}_m^{\mathrm{T}} & 1 \end{pmatrix}$$ $m \times (d+1)$ matrix

$\boldsymbol{y} = (y_1; y_2; \ldots; y_m)$ m column vector,true values of samples

basic linear regression

- solve $\hat{\boldsymbol{w}}$

multiple linear regression:

model: $f(\boldsymbol{x}) = \mathbf{X}\hat{\boldsymbol{w}}$

least square method

loss function: $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})$

$\hat{\boldsymbol{w}} = (\boldsymbol{w}; b)$

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1d} & 1 \\ x_{21} & x_{22} & \dots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{md} & 1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} & 1 \\ \boldsymbol{x}_2^{\mathrm{T}} & 1 \\ \vdots & \vdots \\ \boldsymbol{x}_m^{\mathrm{T}} & 1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} = 1^2 + 2^2 + 3^2 + 4^2$$

$\boldsymbol{y} = (y_1; y_2; \dots; y_m)$

basic linear regression

- solve $\hat{\boldsymbol{w}}$

multiple linear regression:

loss function: $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})$

$$= y^T y - y^T X\hat{w} - \hat{w}^T X^T y + \hat{w}^T X^T X\hat{w}$$

$$- 2X^T y + 2X^T X\hat{w}$$

向量求导公式：

$$\frac{\partial(X\theta)}{\partial\theta} = X^T$$

$$\frac{\partial(\theta^T X)}{\partial\theta} = X$$

$$\frac{\partial(\theta^T X\theta)}{\partial\theta} = (X^T + X)\theta$$

令 $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})$, 对 $\hat{\boldsymbol{w}}$ 求导得到

$$\frac{\partial E_{\hat{\boldsymbol{w}}}}{\partial \hat{\boldsymbol{w}}} = 2\,\mathbf{X}^{\mathrm{T}} (\mathbf{X}\hat{\boldsymbol{w}} - \boldsymbol{y}) = 0$$

$$\downarrow$$

analytical solution: $\hat{\boldsymbol{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$

解析解

basic linear regression

- multiple linear regression:

dataset: $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$

1. substitute

$$\hat{\boldsymbol{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$$

2. get

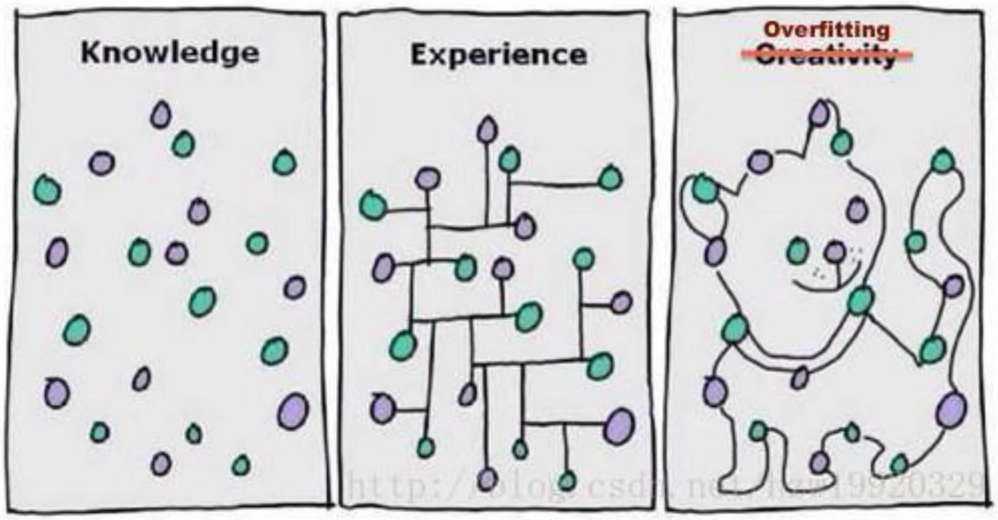
model: $f(\boldsymbol{x}) = \mathbf{X}\hat{\boldsymbol{w}}$

new sample

least square method's drawback

loss function: $E_{(w,b)} = \sum_{i=1}^{m} (f(x_i) - y_i)^2$

(1)最小二乘法的损失函数有可能会造成模型过拟合。



least square method's drawback

loss function: $E_{(w,b)} = \sum_{i=1}^{m} \left(f\left(x_i\right) - y_i\right)^2$

analytical solution:

$$\hat{\boldsymbol{w}}^* = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$$

(2)解析解中$\mathbf{X}^{\mathrm{T}}\mathbf{X}$有可能不是满秩矩阵，那样$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$不存在，

- 样本的特征数远远超过样本数

这样,$\hat{\boldsymbol{w}}$会有多个解。

least square method's application

1.已有样本中的值比较准确
2.样本的特征数小于样本数

Outline

- linear regression model
  - linear model
  - least square method --> basic linear regression
  - ridge regression(岭回归)
  - lasso regression

ridge regression & lasso regression

model: $f(\boldsymbol{x}_i) = w_1 x_{i1} + w_2 x_{i2} + \ldots + w_d x_{id} + b$

vector model: $f(\boldsymbol{x}_i) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i + b$

$$f(\boldsymbol{x}) = \mathbf{X}\hat{\boldsymbol{w}}$$

loss function:

basic linear regression: $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})$

ridge regression: $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}) + \lambda \|\hat{\boldsymbol{w}}\|_2^2 \quad \lambda > 0$

lasso regression: $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}) + \lambda \|\hat{\boldsymbol{w}}\|_1 \quad \lambda > 0$

ridge 回归和lasso回归可以防止模型过拟合，也可以解决样本的特征数远超过样本数的问题。

Norm

Norm
范数

Vector norm

Matrix norm

- to mesure the size of vector and matrix
- Norm is a value.

Vector norm

$\boldsymbol{w} \epsilon \boldsymbol{R}^d$

L_0-norm: the number of non-zero elements in the vector

L_1-norm: $\|\boldsymbol{w}\|_1 = |w_1| + |w_2| + \ldots + |w_d| = \sum_{i=1}^{d} |w_i|$

L_2-norm: $\|\boldsymbol{w}\|_2 = \sqrt{w_1^2 + w_2^2 + \ldots + w_d^2} = (\sum_{i=1}^{d} w_i^2)^{\frac{1}{2}}$

L∞-norm: $\|\boldsymbol{w}\|_\infty = \mathbf{max}\{|w_1|, |w_2|, \ldots, |w_d|\} = \mathbf{max}_{1 \le i \le d}\{|w_i|\}$

$\boldsymbol{w} = \begin{bmatrix} 1 & 2 & -2 \end{bmatrix}$

$||\boldsymbol{w}||_0 = 3 \quad ||\boldsymbol{w}||_1 = 5$

$||\boldsymbol{w}||_2 = 3 \quad ||\boldsymbol{w}||_\infty = 2$

Outline

- linear regression model
  - linear model
  - least square method --> basic linear regression
  - ridge regression(岭回归)
  - lasso regression

ridge regression

model: $f(\boldsymbol{x}_i) = w_1 x_{i1} + w_2 x_{i2} + \ldots + w_d x_{id} + b$

$\Updownarrow$

$f(\boldsymbol{x}_i) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i + b$

loss function: $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}) + \lambda \|\hat{\boldsymbol{w}}\|_2^2 \quad \lambda > 0$

$$= (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}) + \lambda \hat{\boldsymbol{w}}^{\mathrm{T}} \hat{\boldsymbol{w}}$$

$$= (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}) + \lambda \hat{\boldsymbol{w}}^{\mathrm{T}} \boldsymbol{I} \hat{\boldsymbol{w}}$$

$\boldsymbol{I}$:identity matrix

向量求导公式：

$$\frac{\partial (X\theta)}{\partial \theta} = X^T$$

$$\frac{\partial (\theta^T X)}{\partial \theta} = X$$

$$\frac{\partial (\theta^T X \theta)}{\partial \theta} = (X^T + X)\theta$$

$$\frac{\partial E_{\hat{\boldsymbol{w}}}}{\partial \hat{\boldsymbol{w}}} = 2\,\mathbf{X}^{\mathrm{T}} (\mathbf{X}\hat{\boldsymbol{w}} - \boldsymbol{y}) + \lambda \boldsymbol{I} \hat{\boldsymbol{w}} = 0$$

$\Downarrow$

analytical solution: $\hat{\boldsymbol{w}} = (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^T \boldsymbol{y}$

解析解

ridge regression

basic linear regression:

loss function: $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})$

analytical solution: $\hat{\boldsymbol{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$

ridge regression:

loss function: $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}) + \lambda\|\hat{\boldsymbol{w}}\|_2^2 \quad \lambda > 0$

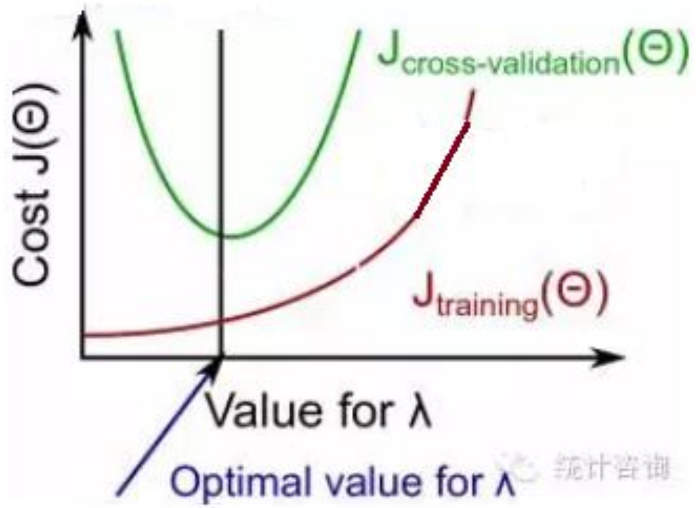
analytical solution: $\hat{\boldsymbol{w}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

$(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})$ is full rank matrix, $\hat{\boldsymbol{w}}$ can get analytical solution.

可以解决特征数大于样本数的问题

ridge regression

loss function: $E_{\hat{w}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}) + \lambda\|\hat{\boldsymbol{w}}\|_2^2 \quad \lambda > 0$



Outline

- linear regression model
  - linear model
  - least square method --> basic linear regression
  - ridge regression(岭回归)
  - lasso regression

lasso regression

- Least Absolute Shrinkage and Selection Operator,最小绝对收敛算子

model: $f(\boldsymbol{x}_i) = w_1 x_{i1} + w_2 x_{i2} + \ldots + w_d x_{id} + b$

vector model: $f(\boldsymbol{x}_i) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i + b$

$$f(\boldsymbol{x}) = \mathbf{X}\hat{\boldsymbol{w}}$$

loss function: $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}) + \lambda \|\hat{\boldsymbol{w}}\|_1 \quad \lambda > 0$

$$\|\boldsymbol{w}\|_1 = |w_1| + |w_2| + \ldots + |w_d| = \sum_{i=1}^{d} |w_i|$$

lasso regression

loss function: $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}) + \lambda\|\hat{\boldsymbol{w}}\|_1 \quad \lambda > 0$

函数在某点可导条件：
1）函数在该点连续
2）函数在该点左右两侧导数都存在并且相等。

$$y = |\omega|, \omega \in R$$

$$y'(0^-) = -1, y'(0^+) = 1$$

lasso regression

loss function: $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}) + \lambda\|\hat{\boldsymbol{w}}\|_1 \quad \lambda > 0$

•Solving loss function

Duing to $\|\boldsymbol{w}\|_1$ having not derivative(导数), we use proximal gradient descent method(近端梯度下降法) to minimize the loss function of lasso regression.

analytical solution:

$$w_{k+1}^i = \begin{cases} z^i - \lambda/L, \lambda/L < z^i; \\ 0, |z^i| \le \lambda/L; \\ z^i + \lambda/L, z^i < -\lambda/L, \end{cases} \qquad L > 0 \text{ 的常数}$$

$$\boldsymbol{z} = \boldsymbol{w}_k - 1/L \nabla f(\boldsymbol{w}_k)$$

可以解决特征数大于样本数的问题

lasso regression

Lasso regression is easier to get sparse solution $\hat{\boldsymbol{w}}$.

sparse solution(稀疏解):

- $\boldsymbol{w}$ contains less non-zero elements.
- It can use feature selection.

$w_d = 0$ means the feature x_{id} is not important to the task.

lasso regression

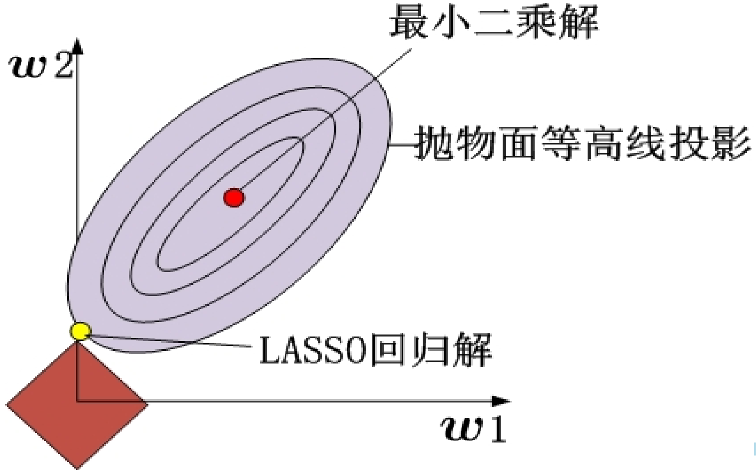
model: $f(\boldsymbol{x}_i) = w_1 x_{i1} + w_2 x_{i2} + \ldots + w_d x_{id} + b$

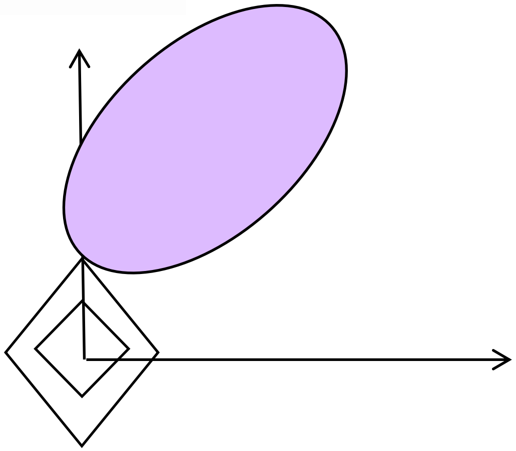
$$f(\boldsymbol{x}_i) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i + b$$

loss function: $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}) + \lambda \|\hat{\boldsymbol{w}}\|_1 \quad \lambda > 0$

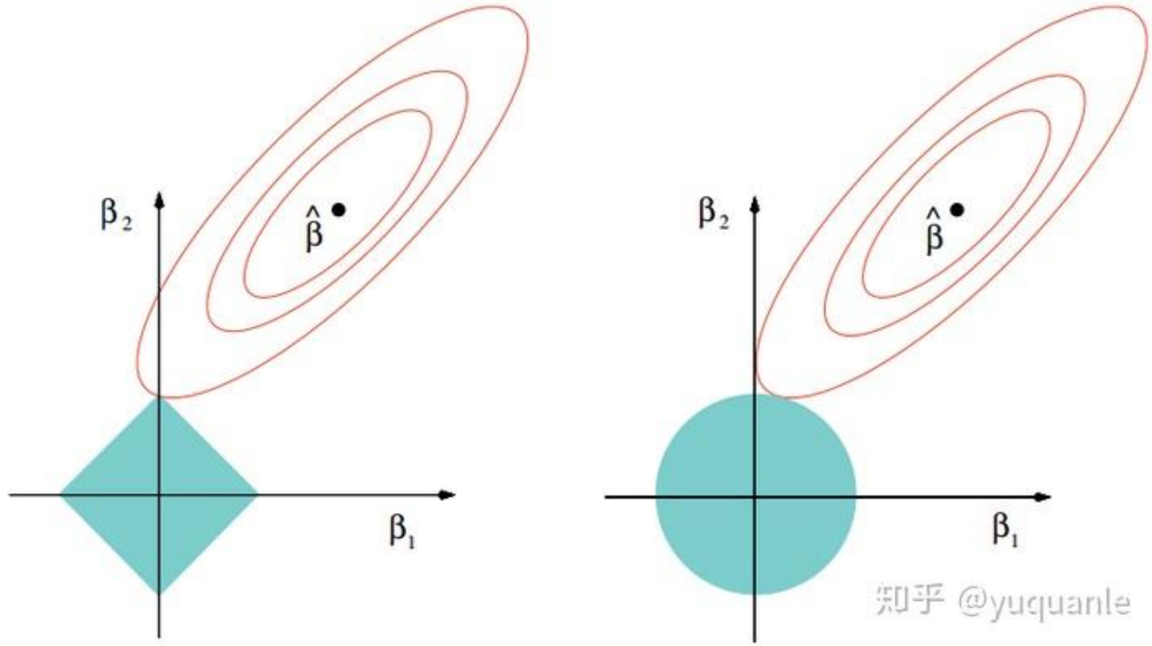
$$\|\boldsymbol{w}\|_1 = |w_1| + |w_2| + \ldots + |w_d|$$

basic linear regression: $E_{(w,b)} = \sum_{i=1}^{m} (f(\boldsymbol{x}_i) - y_i)^2$





lasso vs ridge regression



Lasso和岭回归的区别很好理解，在优化过程中，最优解为函数等值线与约束空间的交集，正则项可以看作是约束空间。可以看出二范的约束空间是一个球形，一范的约束空间是一个方形，这也就是二范会得到很多参数接近0的值，而一范会尽可能非零参数最少。

lasso vs ridge regression

1)Prevent model overfitting.

2)solve the problem that the number of features is larger than the number of samples.

ridge regression

lasso regression

It is easier to get sparse solution w.

function

matlab:

```
b = ridge(y,X,k);  %ridge regression
B = lasso(X,Y);  %lasso regression
```

Summary

- linear regression model
  - linear model

    model: $f(\boldsymbol{x}_i) = w_1 x_{i1} + w_2 x_{i2} + \ldots + w_d x_{id} + b$

    vector model: $f(\boldsymbol{x}_i) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i + b$

    $*\, m$

    $f(\boldsymbol{x}) = \mathbf{X}\hat{\boldsymbol{w}}$

  - least square method --> basic linear regression

    loss function: $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})$

    analytical solution: $\hat{\boldsymbol{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} \boldsymbol{y}$

Summary

- linear regression model
  - ridge regression(岭回归)

    loss function: $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}) + \lambda\|\hat{\boldsymbol{w}}\|_2^2 \quad \lambda > 0$

    analytical solution: $\hat{\boldsymbol{w}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

  - lasso regression

    loss function: $E_{\hat{\boldsymbol{w}}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}) + \lambda\|\hat{\boldsymbol{w}}\|_1 \quad \lambda > 0$

    Lasso regression is easier to get sparse solution $\hat{\boldsymbol{w}}$.

Reference

[1]周志华. 机器学习[M]. 北京：清华大学出版社，2016

[2]李航. 统计学习方法[M]. 北京：清华大学出版社，2012

[3]于剑. 机器学习：从公理到算法[M]. 北京：清华大学出版社，2017

Thank you！