Linear Model

Fish

2019 年 6 月 22 日

Content

1 What

define the example x described by d property, $x = (x_1; x_2; ...; x_d)$, among the variable x, x_i is decided by ith property of x.

linear model showed as follow:

$$f(x) = w_1x_1 + w_2x_2 + ... + w_dx_d + b \tag{1}$$

whether x has d property

vector as well:

$$f(x) = w^Tx + b \tag{2}$$

2 one sample has one property

to simplify the situation, we assump that the input x propert has only one.

2.1 background

we have dataset: $D = (x_1, y_1)), (x_2, y_2), ..., (x_m, y_m)$, to study a model predict tab of output more correctly.

2.2 analysis

to study model: $f(x) = w^Tx + b$, make the predict result $f(x_i) \cong y_i$, the w, b need to be known.

mean square error is correspondinng to Euclidean distance. The way solve model based on mean square error named least square method. according to linear regression, least square method try to find a line that make the Euclidean distance sum of sample to line be least.

2.3 processing

find least value of equation 3 and 4, get the w^*, b^*

$$(w^*, b^*) = \arg\min_{(w,b)} \sum_{i=1}^{m} (f(x_i) - y_i)^2 \tag{3}$$

$$= \arg\min_{(w,b)} \sum_{i=1}^{m} (wx_i + b - y_i)^2 \tag{4}$$

the progress solving the above equation named least square parameter estimation. there differenetiate w and b, could get following part:

$$\frac{\partial E(w,b)}{\partial w} = \sum_{i=1}^{m} 2(-y_i + b + wx_i)x_i = 2\sum_{i=1}^{m} [wx_i^2 - (y_i - b)x_i] \tag{5}$$

$$\frac{\partial E(w,b)}{\partial b} = \sum_{i=1}^{m} 2(-y_i + b + wx_i) = 2[mb - \sum_{i=1}^{m} (y_i - wx_i)] \tag{6}$$

then optimizer solution fo closed-form is shown:

$$w = \frac{\sum_{i=1}^{m} y_i(x_i - \bar{x})}{\sum_{i=1}^{m} x_i^2 - \frac{1}{m}(\sum_{i=1}^{m} x_i)^2} \tag{7}$$

$$b = \frac{1}{m}\sum_{i=1}^{m} (y_i - wx_i) \tag{8}$$

3 one sample has multiply property

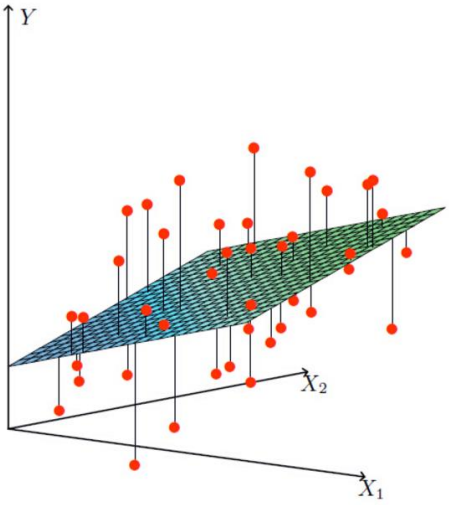

Fig 1: fitting plane in sample space

the common condition is as section one show, dataset D has d property. at present, it should try to study

$$f(\boldsymbol{x}_i) = \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{i}} + b \tag{9}$$

let

$$f(\boldsymbol{x}_i) \cong y_i \tag{10}$$

this condition named multivariate linear regression.

3.1 analysis

similarly, estimate w and b by least square method. conveniently,

$$\hat{\boldsymbol{w}} = (\boldsymbol{w}; b) \tag{11}$$

dataset D show as a matrix $\boldsymbol{X}$, shape $m \times (d+1)$

$$\boldsymbol{X}= \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} & 1 \\ x_{21} & x_{22} & \cdots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} & 1 \end{pmatrix} = \begin{pmatrix} x_1^T & 1 \\ x_2^T & 1 \\ \vdots & \vdots \\ x_m^T & 1 \end{pmatrix} \tag{12}$$

$$\text{tag } \boldsymbol{y} = (y_1; y_2; \cdots; y_m) \tag{13}$$

so fitting plane

$$h_w(x) = w_0 + w_1 x + w_2 x + ... + w_d x, w_0 \text{ is bias} \tag{14}$$

in summary,

$$h_w(x) = \sum_{i=0}^{n} w_i x_i = w^T x \tag{15}$$

3.2 processing

$\boldsymbol{\varepsilon}$ error, represent the difference between true and predict value.

for every sample,

$$y^i = w^T x^i + \varepsilon^i \tag{16}$$

whether ε obey the independent and identical distribution, and observe mean 0 and var θ^2 distribution.

ε obey gause distribution,

$$p(\varepsilon^i) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(\varepsilon^i)^2}{2\sigma^2}) \tag{17}$$

substitution equation 16 into equation 17, as following:

$$p(y^i|x^i; w) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y^i - w^T x^i)^2}{2\sigma^2}) \tag{18}$$

likelihood function:

$$L(w) = \prod_{i=1}^{m} p(y^i|x^i; w) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y^i - w^T x^i)^2}{2\sigma^2}) \tag{19}$$

logarithmic likelihood function:

$$\log L(w) = \log \prod_{i=1}^{m} p(y^i|x^i;w) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y^i - w^T x^i)^2}{2\sigma^2}) \tag{20}$$

expand equation 20th:

$$\begin{aligned} L(w) &= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y^i - w^T x^i)^2}{2\sigma^2}) \\ &= m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^{m} (y^i - w^T x^i)^2 \end{aligned} \tag{21}$$

target: $L(w)$ more big and more better

$$J(w) = \frac{1}{2} \sum_{i=1}^{m} (y^i - w^T x^i)^2 \quad \text{(least square method)} \tag{22}$$

target function:

$$\begin{aligned} J(w) &= \frac{1}{2} \sum_{i=1}^{m} (h_w(x^i) - y^i)^2 \\ &= \frac{1}{2} (\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y})^T (\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y}) \end{aligned} \tag{23}$$

solve partial derivation:

$$\begin{aligned} \bigtriangledown_w J(w) &= \bigtriangledown_w (\frac{1}{2} (\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y})^T (\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y})) \\ &= \bigtriangledown_w (\frac{1}{2} (\hat{\boldsymbol{w}}^{\boldsymbol{T}} \boldsymbol{X}^{\boldsymbol{T}} - \boldsymbol{y}^{\boldsymbol{T}}) (\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y})) \\ &= \bigtriangledown_w ((\frac{1}{2} (\hat{\boldsymbol{w}}^{\boldsymbol{T}} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X}\boldsymbol{w} - \hat{\boldsymbol{w}}^{\boldsymbol{T}} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{y} - \boldsymbol{y}\boldsymbol{X}\hat{\boldsymbol{w}} + \boldsymbol{y}^{\boldsymbol{T}} \boldsymbol{y})) \\ &= \frac{1}{2} (2\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{y} - \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{y}) \\ &= \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{y} \end{aligned} \tag{24}$$

let partial derivation be equal to 0:

$$\hat{\boldsymbol{w}} = (\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X})^{-\boldsymbol{1}} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{y} \tag{25}$$

estimate method: original estimate part

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y}_i)^2} \tag{26}$$

when R^2 value is very close approximation to 1, the model fits dataset D well.

3.3 deep analysis

if X^TX is irreversible or avoid over fitting, increase λ destabilization

$$\theta = (X^TX + \lambda I)^{-1}X^Ty \tag{27}$$

to remember conclusion simply,

$$X\theta = y \Rightarrow X^TX\theta = X^Ty \Rightarrow \theta = (X^TX)^{-1}X^Ty \tag{28}$$

after add λ destabilization:

- X^TX is positive semidefinite matrix: for any non-zero vector u

$$\theta X^TX\theta = (X\theta)^TX\theta \overset{define\ v=X\theta}{\longrightarrow} v^Tv \geq 0 \tag{29}$$

- for any real $\lambda > 0$, $X^TX + \lambda I$ is positive definite matrix, and is reversible. be sure that regerssion equation must be significative.

$$\theta = (X^TX + \lambda I)^{-1}X^Ty \tag{30}$$

3.4 penalty factor

- the target function of linear regerssion:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2 \tag{31}$$

- increase square sum loss:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2 + \lambda\sum_{j=1}^{m}\theta_j^2 \tag{32}$$

- actually, assume that parameter θ obey Gaussian distribution

3.5 |w|and $|w|^2$

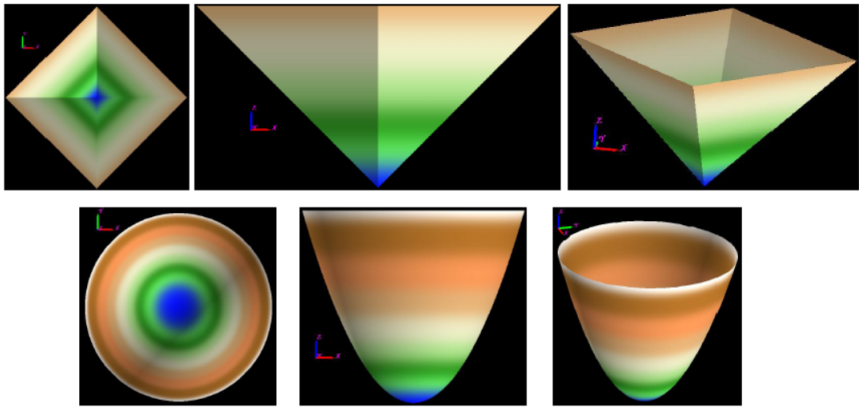

Fig 2: |w| and $|w|^2$

4 machine learning and data analysis

Fig 3: machine learning and data analysis

- cross validation:

  example: 10 folk cross validation