Chapter 3 Linear Model

Jinghao.Zhao

Abstract

This note mainly recode the formula derivation of linear model and logistic regression.

1 Basic Form

Linear model try to learn a linear function to predict samples which consists of d features. That is

$$f(\mathbf{x}) = \omega_1 x_1 + \omega_2 x_2 + ... + \omega_d x_d + b, \tag{1}$$

which write in vector format is

$$f(\mathbf{x}) = \boldsymbol{\omega}^T \mathbf{x} + b, \tag{2}$$

where $\boldsymbol{\omega} = (\omega_1; \omega_2; ...; \omega_d)$. The model will be defined after learned $\boldsymbol{\omega}$ and b.

2 Linear Regression

2.1 Linear regression

Support we have

$$f(x_i) = \omega x_i + b \rightarrow f(x_i) \simeq y_i. \tag{3}$$

So we use **square loss** to measure the model. We want to get

$$(w^*, b^*) = \arg\min_{(\omega,b)} \sum_{i=1}^{m} (f(x_i) - y_i)^2 \ = \ \arg\min_{(\omega,b)} \sum_{i=1}^{m} (y_i - \omega x_i - b)^2. \tag{4}$$

We want to solve the ω and b to minimize the square loss. In linear regression, the **parameter estimation** is the process of minimizing $E_{(\omega,b)} = \sum_{i=1}^{m} (y_i - \omega x_i - b)^2$. First we get derivatives for ω and b.

$$\frac{\partial E_{(\omega,b)}}{\partial \omega} = 2(\omega \sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m} (y_i - b) x_i), \tag{5}$$

$$\frac{\partial E_{(\omega,b)}}{\partial b} = 2(mb - \sum_{i=1}^{m}(y_i - \omega x_i)), \tag{6}$$

make (5) and (6) equal to 0, then

$$b = \frac{1}{m}\sum_{i=1}^{m}(y_i - \omega x_i) \tag{7}$$

Plug(7) into (5), and make $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$

$$\begin{aligned}
\frac{\partial E_{(\omega,b)}}{\partial \omega} &= 2(\omega\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m}(y_i - b)x_i) = 0 \\
&\Rightarrow \omega\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m} x_i y_i + \sum_{i=1}^{m} x_i b = 0 \\
&make \sum_{i=1}^{m} x_i b = m \cdot \frac{1}{m}\sum_{i=1}^{m} x_i b, \\
&\Rightarrow \omega\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m} x_i y_i + mb\bar{x} = 0 \\
\Rightarrow \omega\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m} x_i y_i + m \cdot \frac{1}{m}\sum_{i=1}^{m}(y_i - \omega x_i) &= 0 \ (use\ (7)) \\
\Rightarrow \omega(\sum_{i=1}^{m} x_i^2 - \bar{x}\sum_{i=1}^{m} x_i) &= \sum_{i=1}^{m} x_i y_i - \bar{x}\sum_{i=1}^{m} y_i \\
\omega &= \frac{\sum_{i=1}^{m} x_i y_i - \bar{x}\sum_{i=1}^{m} y_i}{\sum_{i=1}^{m} x_i^2 - (\sum_{i=1}^{m} x_i)^2}
\end{aligned} \tag{8}$$

2.2 Multivariate linear regression(多元线性回归)

Generally, when we have dataset with more than one feature, we want to learn

$$f(\mathbf{x}_i) = \boldsymbol{\omega}^T\mathbf{x}_i + b \tag{9}$$

so we set $\hat{\boldsymbol{\omega}} = (\boldsymbol{\omega}; b)$ represent a (d × 1) vector, and make dataset D into a (m × (d + 1)) vector $\mathbf{X}$, also make y as (m × 1) vector as$(y_1; y_2; ...; y_m)$, then we want to get

$$\begin{aligned}
\hat{\boldsymbol{\omega}}^* &= \arg\min_{\hat{\omega}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\omega}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\omega}}) = E_{\hat{\omega}} \\
E_{\hat{\omega}} &= y^T y - y^T X\omega - \omega^T X^T y + \omega^T X^T X\omega
\end{aligned} \tag{10}$$

then, we get the derivatives

$$\frac{\partial E_{\hat{\omega}}}{\partial \omega} = \frac{\partial y^T y}{\partial \omega} - \frac{\partial y^T X\omega}{\partial \omega} - \frac{\partial \omega^T X^T y}{\partial \omega} + \frac{\partial \omega^T X^T X\omega}{\partial \omega}. \tag{11}$$

According to wiki's explanation[1] [2]about the derivation of matrix, we can get that:

$$\begin{aligned} \frac{\partial y^T y}{\partial \omega} &= 0, \\ \frac{\partial y^T X\omega}{\partial \omega} = \frac{\partial \omega^T X^T y}{\partial \omega} &= X^T y, \\ \frac{\partial \omega^T X^T X\omega}{\partial \omega} &= 2X^T X\omega. \end{aligned} \tag{12}$$

Therefore,

$$\frac{\partial E_{\hat{\omega}}}{\partial \omega} = -2X^T y + 2X^T X\omega = 2X^T(X\omega - y) \tag{13}$$

when X is non-singular matrix(满秩矩阵), we can get ω by

$$\frac{\partial E_{\hat{\omega}}}{\partial \omega} = 2X^T(X\omega - y) = 0 \Rightarrow \omega = (X^T X)^{-1} X^T y. \tag{14}$$

3 Logistic Regression

3.1 Generalized linear model(广义线性模型)

Support we have g(·) which is monotonicity and differentiability(单调可微函数), and

$$y = g^{-1}(\omega^T x + b), \tag{15}$$

such kinds of models are called generalized linear model. Obviously, logistic regression is a special situation when g(·) = ln(·).

3.2 Logistic regression

In one sentence, logistic regression is using linear reflection to implement the fitting of exponential function(用线性函数映射来间接实现对指数函数的拟合). Here's the derivation for (3.17) to (3.27). For sigmoid function, first we have $y = \frac{1}{1+e^{-z}}$ and $z = \omega^T x + b$, for(3.18) to (3.21),

$$y = \frac{1}{1 + e^{-(\omega^T x + b)}} \Rightarrow \frac{1-y}{y} = e^{-(\omega^T x + b)} \Rightarrow \omega^T x + b = ln\frac{y}{1-y}. \tag{16}$$

If we treated y as the possibility for x to be positive, then 1-y is the the possibility for x to be negative. Then $\frac{y}{1-y}$ is the **odds(**几率**)**.

For (3.22) to (3.27), when we treated y as posterior probability(后验概率), from (3.22), we have $p(y = 1|x) = 1 - p(y = 0|x)$, then from (3.19) we can get

$$\frac{p(y=1|x)}{1 - p(y=0|x)} = e^{\omega^T x + b} \Rightarrow (3.23), (3.24). \tag{17}$$

[1]https://en.wikipedia.org/wiki/Matrix_calculus#Scalar-by-vector_identities
[2]https://blog.csdn.net/nomadlx53/article/details/50849941

(3.25) to (3.27) told us how to calculate ω and b. Since we have the log-likelihood of ω and b as (3.25) we have $\beta = (\omega;\text{b})$ and $\hat{x} = (\text{x};\ 1)$(这里一些矩阵运算的说明，ω , x均为是列向量$\{\omega_1;\omega_2...\omega_d\}$, 因此$\beta$相当于$\omega$ 加了一行b, 而$\hat{x}$是给x加一行1, 所以$\omega^T x + b = \beta^T \hat{x}$)

y具有二值性,可得:

$$(3.26) \Rightarrow p(y|x_i;\omega,b) = \begin{cases} p_1(\hat{x_i};\beta), & y_i = 1 \\ p_0(\hat{x_i};\beta), & y_i = 0 \end{cases} \tag{18}$$

we put (3.23) (3.24) and (3.26) into (3.25), then

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{m} ln(y_i \cdot \frac{e^{\beta^T x}}{1+e^{\beta^T x}} + (1-y_i)\frac{1}{1+e^{\beta^T x}}) \\ &= \sum_{i=1}^{m} (ln(y_i e^{\beta^T x} + 1 - y_i) - ln(1+e^{\beta^T x}) \\ &= \begin{cases} -ln(1+e^{\beta^T x}), & y_i = 0 \\ \beta^T x - ln(1+e^{\beta^T x}), & y_i = 0 \end{cases} \\ &\Rightarrow (3.27) \end{aligned} \tag{19}$$

注意(3.25)的 l(ω,b) = -l(β) (3.27). 所以说最大化(3.25) 即为最小化(3.27), 可以利用梯度下降等方法求得.

4 Softmax Regression

Softmax regression is the extension of logistic regression in multi-classification problem. When we want to estimate the probability $p(y = j|x)$ for input x to be classified as class j, we set

$$p(y = j|x = x_i) = \frac{e^{\omega_j^T x_i}}{\sum_{l=1}^{k} e^{\omega_l^T x_i}} \tag{20}$$

The loss function J(θ) is

$$\begin{aligned} J(\theta) &= -\frac{1}{m}[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y^{(i)} = j\} \log p(y^i = j|x = x_i)], \\ 1\{y^{(i)} = j\} &= \begin{cases} 1, y^{(i)} = j \\ 0, \qquad y^{(i)}! = j \end{cases} \end{aligned} \tag{21}$$

We can see the loss of softmax is the summation of logistic regression for k different classes. Usually we add penalty to the loss to revise the parameter, then(21) will be

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y^{(i)} = j\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}] + \frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=0}^{n}\theta_{ij}^2, \tag{22}$$

and we use gradient descent to calculate each θ_j to get minimum J(θ),

$$\begin{aligned}\frac{\partial J(\theta)}{\partial \theta} &= -\frac{1}{m}\sum_{i=1}^{m}(1\{y^{(i)}=j\}\frac{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}{e^{\theta_j^T x^i}} \cdot \frac{e^{\theta_j^T x_i}\cdot x_i \cdot \sum_{l=1}^{k} e^{\theta_l^T x^i} - e^{\theta_l^T x^i}\cdot x_i \cdot e^{\theta_j^T x^i}}{(\sum_{l=1}^{k} e^{\theta_l^T x^i})^2}) + \lambda\theta_j \\ &= -\frac{1}{m}\sum_{i=1}^{m}(1\{y^{(i)}=j\}\frac{\sum_{l=1}^{k} e^{\theta_l^T x^i} - e^{\theta_j^T x_i}}{\sum_{l=1}^{k} e^{\theta_l^T x^i}}\cdot x_i) + \lambda\theta_j \\ &= -\frac{1}{m}\sum_{i=1}^{m}[x_i \cdot (1\{y^{(i)}=j\} - p(y_i=j|x_i;\theta))] + \lambda\theta_j \end{aligned} \tag{23}$$

Gradient descent update θ_j by $(\theta_j - \alpha\nabla_{\theta_j}J(\theta))$

5 Class Imbalance Problem

In previous classification, we resume the positive and negative samples have basically same number, but sometimes there may have big differences between the number of different samples, which called class-imbalance. In normal linear regression, we treat y as the probability to classified as positive, and 1-y is probability to be negative. Therefore, when y < 1-y, that is, $\frac{y}{1-y} > 1$ will be positive. When we faced class-imbalance problem, suppose we have m^+ positive samples and m^- negative samples, there are three main methods to solve the problem.(when $m^- > m^+$)

1. undersampling(欠采样) means we cam remove some negative samples.
2. oversampling(过采样) means adding some positive samples(BUT NOT just repeat sampling which will cause serious overfitting).
3. rescaling(再缩放) means setting $\frac{y'}{1-y'} = \frac{y}{1-y} \times \frac{m^-}{m^+}$