Machine Learning Course （Li Hongyi 2023)

1.ChatGPT Introduction

Possible Method：

Pre-Train 预训练 => Self-Supervised 自督导式学习 => Supervised learning 督导式学习 => Intensive training 强化训练

Foreground:

Prompting 工程，Neural Editing , AI 检测，Machine Unlearning

2.Regression

Step1 Linear Model

$$y = b + \omega * x_{cp} => y = b + \sum w_i * x_i$$

$$x_i : feature, w_i : weight, b : bias$$

Step2 Goodness of function

Loss Function L

$$L(f) = L(\omega, b)$$

Normal

$$L(f) = \sum_{n=1}^{10} (\hat{y}^n - (b + \omega * x_{cp}^n))^2$$

Step3 Gradient Descent

For one parameter

$$\omega^* = argmin_\omega L(\omega)$$

1.pick initial value ω^0

2.$\omega^1 \leftarrow \omega^0 - \eta * \frac{dL}{d\omega}|_{\omega=\omega^0}$

3.$\omega^2 \leftarrow \omega^1 - \eta * \frac{dL}{d\omega}|_{\omega=\omega^1}$

.... => Local optimal (not global)

For two parameters

1.$\omega^1 \leftarrow \omega^0 - \eta * \frac{dL}{d\omega}|_{\omega=\omega^0, b=b^0}$

$b^1 \leftarrow b^0 - \eta * \frac{dL}{db}|_{\omega=\omega^0, b=b^0}$

2.$\omega^2 \leftarrow \omega^1 - \eta * \frac{dL}{d\omega}|_{\omega=\omega^1, b=b^1}$

$b^2 \leftarrow b^1 - \eta * \frac{dL}{db}|_{\omega=\omega^1, b=b^1}$

......

For Many Types

back to design model

$y = b_1 * \delta(x_s = pidgey) + \omega_1 * \delta(x_s = pidgey) * x_{cp} + ... + b_4 * \delta(x_s = Eevee) + \omega_4 * \delta(x_s = Eevee)$

$$\delta = \begin{cases} 1, & x_s = type \\ 0, & x_s \neq type \end{cases}$$

Regularization

$$L = \sum_n (\hat{y} - (b + \sum \omega_i x_i))^2 + \lambda \sum w_i^2$$

λ越大，找到的越平滑

3.Classification

$$x \rightarrow Function \rightarrow ClassN$$

Ideal Alternatives

Function(Model)

$$x \Rightarrow \begin{cases} g(x) > 0 | Output = class1 \\ else | Output = class2 \end{cases}$$

Loss Function

$$L(f) = \sum_n \delta(f(x^n) \neq \hat{y}^n)$$

Number of times f get incorrect results on training data.

Find best function

Perceptron,SVM

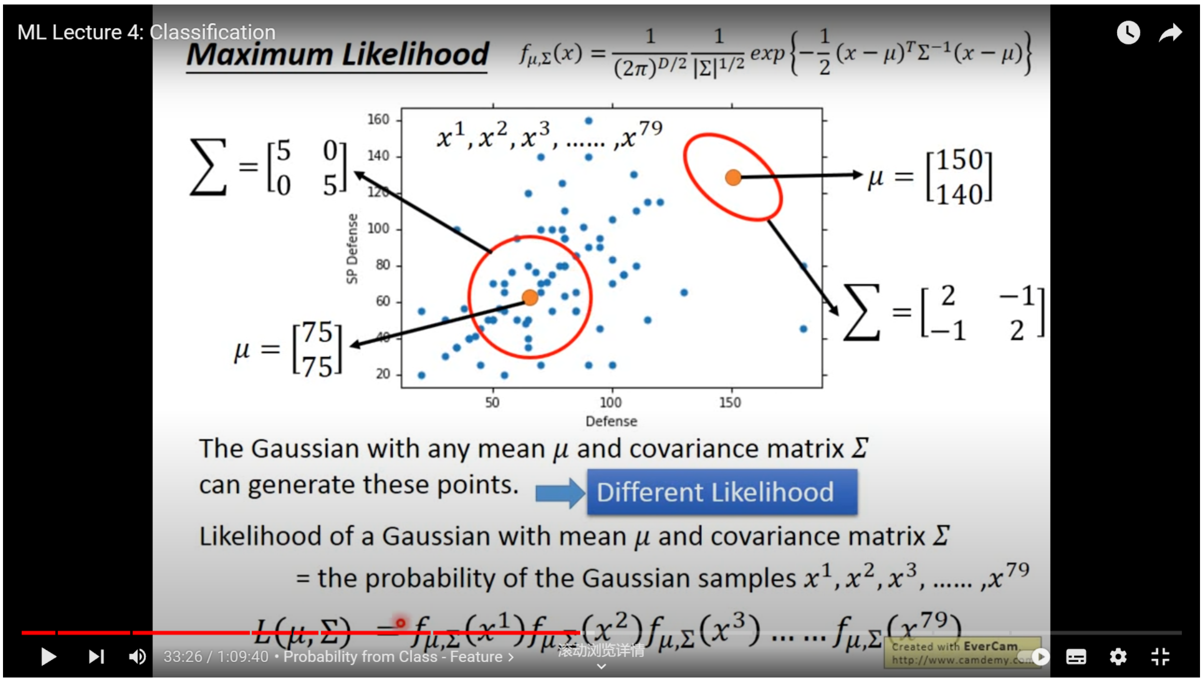
Generative Model(Gaussian Distribution)

$$f_{\mu,\sum(x)} = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\sum|^{1/2}} exp\{\frac{-1}{2}(x-\mu)^T * \sum^{-1}(x-\mu)\}$$

Determined by Mean μ ,convariance matrix $\sum$

Maximum Likelihood

![](



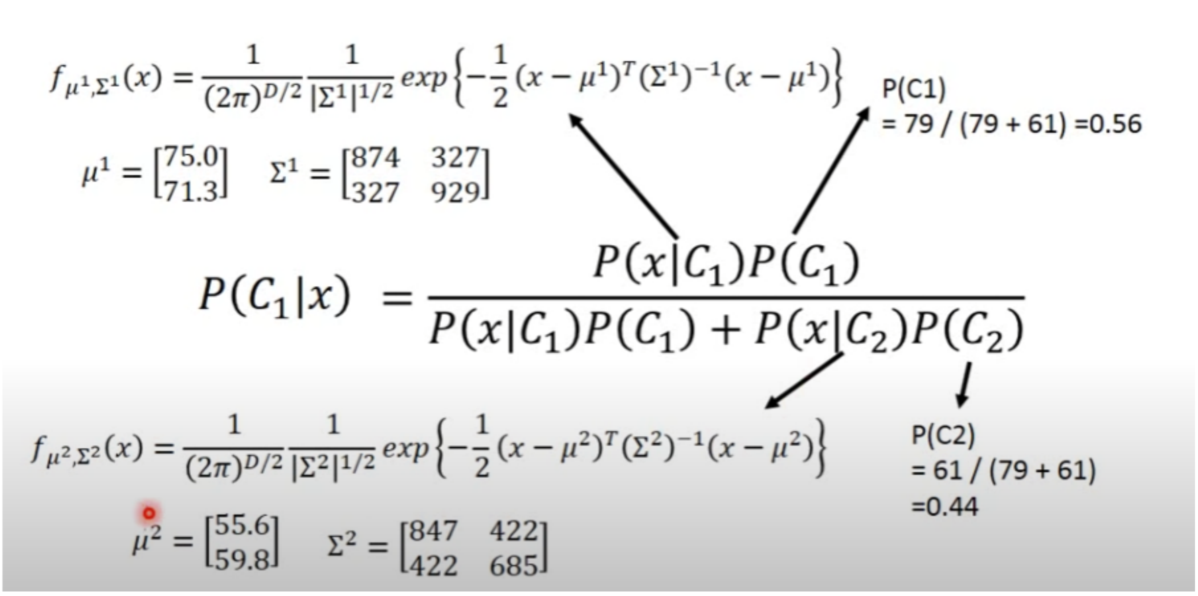
$$L(\mu,\sum) = f_{\mu,\sum}(x^1) * f_{\mu,\sum}(x^2)\ldots\ldots f_{\mu,\sum}(x^79)$$

$$\Rightarrow (\mu^*,\sum^*) = argMax_{\mu,\sum}L(\mu,\sum)$$

$$\mu^* = \frac{1}{79}\sum_{n=1}^{79} x^n$$

$$\sum^* = \frac{1}{79}\sum_{n=1}^{79}(x^n-\mu^*)(x^n-\mu^*)^T$$

Back To Classification



$$P(C_1|x) > 0.5 \Rightarrow x \in Class1$$

从多维空间来看增加更多参数容易导致overfitting

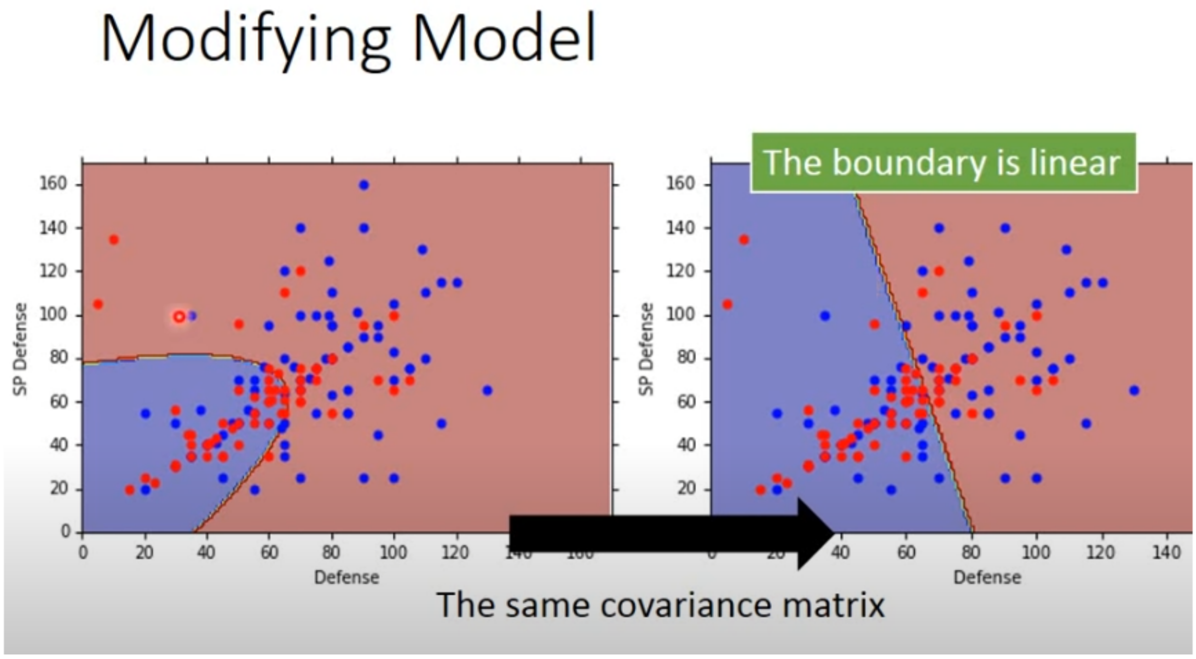
Resolution

给两边分类相同的$\sum$

$$L(\mu_1, \mu_2, \sum) = f_{\mu_1,\sum}(x^1) * f_{\mu_1,\sum}(x^2).....f_{\mu_1,\sum}(x^{79}) f_{\mu_2,\sum}(x^{80})...f_{\mu_2,\sum}(x^{140})$$

$$\Rightarrow \mu_1 = \mu_2$$

$$\Rightarrow \sum = \frac{79}{140}\sum^1 + \frac{61}{140}\sum^2$$



Three Steps

Function Set(Model)

$$x \Rightarrow P(C_1|x) = \frac{P(x|C_1) * P(C_1)}{P(x|C_1) * P(C_1) + P(x|C_2) * P(C_2)} \Rightarrow \begin{cases} P(C_1|x) > 0.5 \rightarrow class1 \\ P(C_1|x) < 0.5 \rightarrow class2 \end{cases}$$

Goodness of a function

mean μ ,convariance $\sum$ Maximizing the likelihood

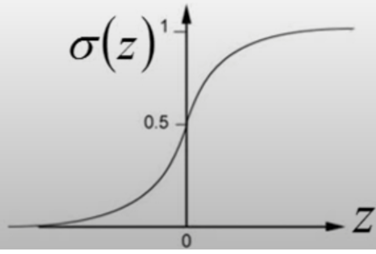
Transformation

Posterior Probability

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

$$= \frac{1}{1 + \frac{P(x|C_2)P(C_2)}{P(x|C_1)P(C_1)}} = \frac{1}{1 + exp(-z)} = \sigma(z)$$

Sigmoid function



$$z = ln\frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)}$$

ssification

$$z = \boxed{ln\frac{P(x|C_1)}{P(x|C_2)}} + \boxed{ln\frac{P(C_1)}{P(C_2)}} = \frac{N_1}{N_2}$$

$$P(x|C_1) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma^1|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1)\right\}$$

$$P(x|C_2) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma^2|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)\right\}$$

$$ln\frac{\cancel{\frac{1}{(2\pi)^{D/2}}}\boxed{\frac{1}{|\Sigma^1|^{1/2}}} exp\left\{-\frac{1}{2}(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1)\right\}}{\cancel{\frac{1}{(2\pi)^{D/2}}}\boxed{\frac{1}{|\Sigma^2|^{1/2}}} exp\left\{-\frac{1}{2}(x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)\right\}}$$

$$\boxed{ln\frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}}} exp\left\{-\frac{1}{2}[(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1) - (x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)]\right\}$$

If $\sum_1 = \sum_2 = \sum$

$$z = (\mu^1 - \mu^2)^T \sum^{-1} x - \frac{1}{2}(\mu^1)^T(\sum^1)^{-1}\mu^1 + \frac{1}{2}(\mu^2)^T(\sum^1)^{-1}\mu^2 + \ln\frac{N1}{N2}$$

$$z = w^T x - b$$

$$\Rightarrow P(C_1|x) = \sigma(w * x + b)$$

4.Logistic Regression

Comparison with Linear Regression

Step1

For Logistic Regression

$$f_{w,b}(x) = \sigma(\sum_i w_i x_i + b)$$

Output: between 0 and 1

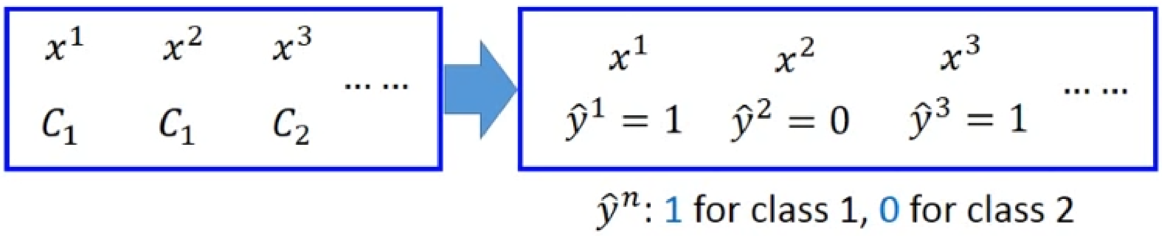
For Linear Regression

$$f_{w,b}(x) = \sum_i w_i x_i + b$$

Output: any value

Step2 Goodness of a Function

Training Data :



$$L(w,b) = f_{w,b}(x^1) f_{w,b}(x^2)(1 - f_{w,b}(x^3)) \ldots f_{w,b}(x^N)$$

The most likely w^* and b^* is the one with the largest $L(w,b)$

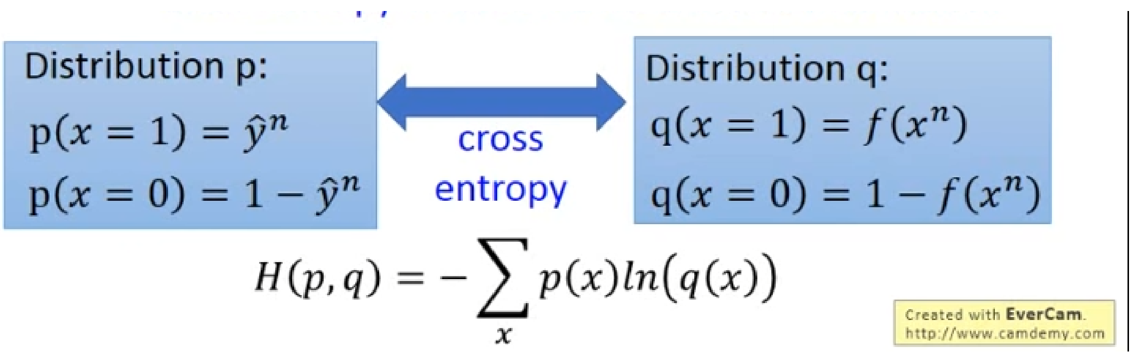
$$w^*, b^* = arg \max_{w,b} L(w,b)$$

$$\Rightarrow w^*, b^* = arg \max_{w,b} - \ln L(w,b)$$

$$\begin{aligned} -ln f_{w,b}(x^1) &\Rightarrow -[\hat{y}^1 ln f(x^1) + (1 - \hat{y}^1 ln(1 - f(x^1))] \\ -ln f_{w,b}(x^2) &\Rightarrow -[\hat{y}^2 ln f(x^2) + (1 - \hat{y}^2 ln(1 - f(x^2))] \\ -ln f_{w,b}(x^3) &\Rightarrow -[\hat{y}^3 ln f(x^3) + (1 - \hat{y}^3 ln(1 - f(x^3))] \\ &\ldots . \end{aligned}$$

$$\Rightarrow -lnL(w,b) = \sum_n -[\hat{y}^n ln f_{w,b}(x^n) + (1 - \hat{y}^n ln(1 - f_{w,b}(x^n))]$$

Cross entropy between two Bernoulli distribution



For Logistic Regression

$\hat{y}^n$:1 for class 1, 0 for class 2

$$L(f) = \sum_n C(f(x^n), \hat{y}^n)$$

For Linear Regression

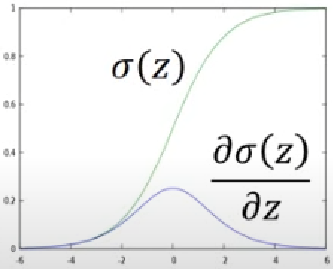
$$L(f) = \frac{1}{2}\sum_n (f(x^n) - \hat{y}^n)^2$$

Step3 Find the best function

$$\frac{-\ln L(w,b)}{\partial w_i} = \sum_n -\left[\hat{y}^n \overset{\left(1-f_{w,b}(x^n)\right)x_i^n}{\boxed{\frac{\ln f_{w,b}(x^n)}{\partial w_i}}} + (1-\hat{y}^n)\frac{\ln\left(1-f_{w,b}(x^n)\right)}{\partial w_i}\right]$$

$$\frac{\partial \ln f_{w,b}(x)}{\partial w_i} = \frac{\partial \ln f_{w,b}(x)}{\partial z}\frac{\partial z}{\partial w_i} \qquad \frac{\partial z}{\partial w_i} = x_i$$

$$\frac{\partial \ln \sigma(z)}{\partial z} = \frac{1}{\sigma(z)}\frac{\partial \sigma(z)}{\partial z} = \frac{1}{\cancel{\sigma(z)}}\cancel{\sigma(z)}(1-\sigma(z))$$



$$\frac{\partial \ln\left(1-f_{w,b}(x)\right)}{\partial w_i} = \frac{\partial \ln\left(1-f_{w,b}(x)\right)}{\partial z}\frac{\partial z}{\partial w_i} \qquad \frac{\partial z}{\partial w_i} = x_i$$

$$\frac{\partial \ln(1-\sigma(z))}{\partial z} = -\frac{1}{1-\sigma(z)}\frac{\partial \sigma(z)}{\partial z} = -\frac{1}{\cancel{1-\sigma(z)}}\sigma(z)\cancel{(1-\sigma(z))}$$

$$\frac{-\ln L(w,b)}{\partial w_i} = \sum_n -\left[\hat{y}^n \overset{\left(1-f_{w,b}(x^n)\right)x_i^n}{\boxed{\frac{\ln f_{w,b}(x^n)}{\partial w_i}}} + (1-\hat{y}^n)\overset{f_{w,b}(x^n)x_i^n}{\boxed{\frac{\ln\left(1-f_{w,b}(x^n)\right)}{\partial w_i}}}\right]$$

$$= \sum_n -\left[\hat{y}^n \left(1-f_{w,b}(x^n)\right)x_i^n - (1-\hat{y}^n)f_{w,b}(x^n)x_i^n\right]$$

$$= \sum_n -\left[\hat{y}^n - \cancel{\hat{y}^n f_{w,b}(x^n)} - f_{w,b}(x^n) + \cancel{\hat{y}^n f_{w,b}(x^n)}\right] x_i^n$$

$$= \sum_n -\left(\hat{y}^n - f_{w,b}(x^n)\right)x_i^n$$

Larger difference, larger update

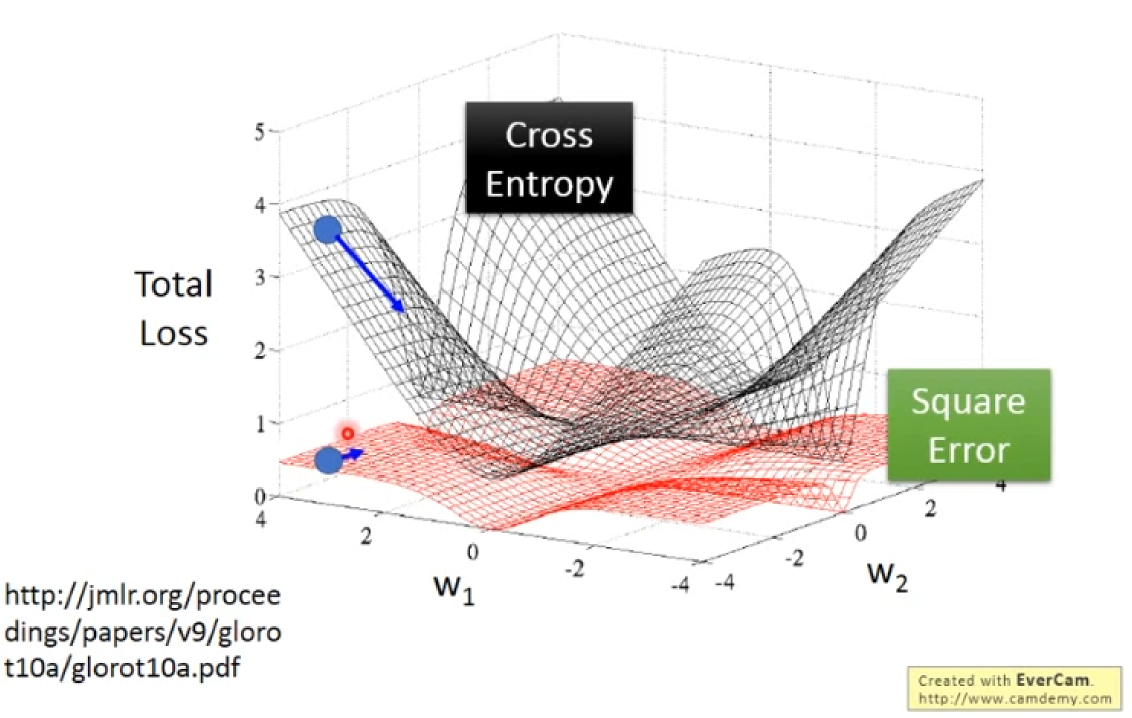
$$w_i \leftarrow w_i - \eta \sum -\left(\hat{y}^n - f_{w,b}(x^n)\right)x_i^n$$

For Logistic Regression and LInear Regression

The same

$$w_i \leftarrow w_i - \eta \sum_n -(\hat{y}^n - f_{w,b}(x^n))x_i^n$$

If Use Logistic Regression with square error



用Square error 离目标很远时趋势也很小

30：42