# **Linear Model**

## 1. Linear Regression

### 1.1 Simple Linear Regression

· model formulation

$$egin{aligned} \hat{y_i} &= w x_i + b \ \ J\left(w, b
ight) &= rac{1}{m} \sum_{i=1}^m \left(\hat{y_i} - y_i
ight)^2 \ \ \left(w^*, b^*
ight) &= rg \min_{\left(w, b
ight)} J\left(w, b
ight) \end{aligned}$$

parameter estimation

$$\begin{split} \frac{\partial J}{\partial w} &= \frac{2}{m} \sum_{i} (\hat{y}_{i} - y_{i}) \frac{\partial}{\partial w} (\hat{y}_{i} - y_{i}) = \frac{2}{m} \sum_{i} (\hat{y}_{i} - y_{i}) x_{i} \\ &= \frac{2}{m} \left( w \sum_{i} x_{i}^{2} - \sum_{i} (y_{i} - b) x_{i} \right) \\ \frac{\partial J}{\partial b} &= \frac{2}{m} \sum_{i} (\hat{y}_{i} - y_{i}) \frac{\partial}{\partial b} (\hat{y}_{i} - y_{i}) = \frac{2}{m} \sum_{i} (\hat{y}_{i} - y_{i}) \\ &= \frac{2}{m} \left( mb - \sum_{i} (y_{i} - wx_{i}) \right) \end{split}$$

· closed-form solution

$$w^* = rac{\sum y_i (x_i - x)}{\sum x_i^2 - rac{1}{m} (\sum x_i)^2} \ b^* = rac{1}{m} \sum_{i=1}^m (y_i - wx_i)$$

# 1.2 Multiple Linear Regression

. model formulation

$$egin{aligned} oldsymbol{w} &= (oldsymbol{w}; b) & oldsymbol{X} = \begin{pmatrix} oldsymbol{x}_1 & oldsymbol{x}_2 & \cdots & oldsymbol{x}_m \\ 1 & 1 & \cdots & 1 \end{pmatrix}^T = \begin{pmatrix} oldsymbol{x}_1^T & 1 \\ oldsymbol{x}_2^T & 1 \\ \vdots \\ oldsymbol{x}_m^T & 1 \end{pmatrix} \\ \hat{oldsymbol{y}}^{(i)} &= oldsymbol{X} oldsymbol{w}^T oldsymbol{x}^{(i)} \\ J(oldsymbol{w}) &= oldsymbol{1}_m \sum_{i=1}^m \left( \hat{oldsymbol{y}}^{(i)} - oldsymbol{y}^{(i)} \right)^2 = oldsymbol{1}_m \| \hat{oldsymbol{y}} - oldsymbol{y} \|^2 \\ &= rac{1}{m} (oldsymbol{X} oldsymbol{w} - oldsymbol{y})^T (oldsymbol{X} oldsymbol{w} - oldsymbol{y}) &\mapsto [MSE_{train}] \\ oldsymbol{w}^* &= rg \min J(oldsymbol{w}) \end{aligned}$$

度量模型性能的一种方法是计算在测试集上的均方误差MSE,为了减小MSE,一种直观方式是最小化训练集上的均方误差

#### parameter estimation

Normal Equation

$$abla_{\boldsymbol{w}} J(\boldsymbol{w}) = 0 \Rightarrow \nabla_{\boldsymbol{w}} (\boldsymbol{X} \boldsymbol{w} - \boldsymbol{y})^T (\boldsymbol{X} \boldsymbol{w} - \boldsymbol{y}) = 0$$

$$\Rightarrow \boldsymbol{w}^* = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

Gradient Descent

$$egin{aligned} rac{\partial J}{\partial oldsymbol{w}_j} &= rac{\partial}{\partial w_j} \Bigg(rac{1}{m} \sum_{i=1}^m \left(\hat{y}_i - y_i
ight)^2\Bigg) \ &= rac{1}{m} \sum_{i=1}^m \left(\hat{oldsymbol{y}}^{(i)} - oldsymbol{y}^{(i)}
ight) oldsymbol{x}_j^{(i)} \end{aligned}$$

### 1.3 Probabilistic interpretation for cost function

 $m{y}^{(i)} = m{ heta}^T m{x}^{(i)} + m{\epsilon}^{(i)}$  ,assume  $m{\epsilon}^{(i)} \sim \mathcal{N}(0, \sigma^2)$  and are distributed IID

$$p(oldsymbol{\epsilon}^{(i)}) = rac{1}{\sqrt{2\pi}\sigma}exp\left(-rac{(oldsymbol{\epsilon}^{(i)})^2}{2\sigma^2}
ight)$$

for given  $x^{(i)}$  and heta

$$p(y_i|x_i; heta) = rac{1}{\sqrt{2\pi}\sigma}exp\left(-rac{\left(oldsymbol{y}^{(i)}-oldsymbol{ heta}^Toldsymbol{x}^{(i)}
ight)^2}{2\sigma^2}
ight)$$

$$L(oldsymbol{ heta}) = L(oldsymbol{ heta}; oldsymbol{X}, oldsymbol{y}) = p(oldsymbol{y} | oldsymbol{X}; oldsymbol{ heta}) = = \prod_{i=1}^m p(oldsymbol{y}^{(i)} | oldsymbol{x}^{(i)}; oldsymbol{ heta})$$

似然函数(likelihood fuction),  $L(\theta)$ 表示在概率密度函数的参数是 $\theta$ 时,得到这组样本的概率

$$egin{aligned} m{ heta}_{ML} &= rg \max_{m{ heta}} L(m{ heta}) = rg \max_{m{ heta}} \log L(m{ heta}) \ \ell(m{ heta}) &= \log L(m{ heta}) = \sum_{i=1}^m \log p(m{y}^{(i)} | m{x}^{(i)}; m{ heta}) \ &= \sum_{i=1}^m \left( rac{m{1}}{\sqrt{2\pi}\sigma} - rac{m{y}^{(i)} - m{ heta}^T m{x}^{(i)}}{2\sigma^2} 
ight) \ &= m \ln rac{1}{\sqrt{2\pi}\sigma} - rac{1}{2\sigma^2} \sum_{i=1}^m \left( \hat{m{y}}^{(i)} - m{y}^{(i)} 
ight)^2 \ & rg \max \ell(m{ heta}) \Rightarrow rg \min \| \hat{m{y}} - m{y} \|^2 \end{aligned}$$

最大化关于w的对数似然和最小化均方误差会得到相同的参数估计

#### 1.4 Generalized Linear Models

For a monotonic and differentiable function  $g(\cdot)$ 

$$y = g^{-1}(\boldsymbol{w}^T\boldsymbol{x} + b)$$

These broader family of models Generalized Linear Models,其中函数 $g(\cdot)$ 称为"联系函数"(link function)

# 2. Logistic Regression

#### · model formulation

$$\hat{y} = rac{1}{1+e^{-(oldsymbol{w}^Toldsymbol{x}+b)}}$$

$$ightarrow \ln rac{y}{1-y} = oldsymbol{w}^T oldsymbol{x} + b$$

y/(1-y):几率(odds)反映了x为正例的相对可能性、 $ln\ y/(1-y):$ 对数几率( $log\ odds$ 亦称logit)

$$g(z) = rac{1}{1+e^{-z}}$$
  $Cost(\hat{y},y) = -y\log\hat{y_i} - (1-y)\log(1-\hat{y})$ 

$$J(w) = -rac{1}{m} \sum_{i=1}^m \left[ m{y}^{(i)} \log \hat{m{y}}^{(i)} + (1-m{y}^{(i)}) \log (1-\hat{m{y}}^{(i)}) 
ight]$$

在这里不使用均方误差代价函数是由于均方误差代价函数在这里不是凸函数,存在多个局部极小值点

### · parameter estimation

Gradient Descent

$$rac{\partial J}{\partial oldsymbol{w}_j} = rac{1}{m} \sum_{i=1}^m \left( \hat{oldsymbol{y}}^{(i)} - oldsymbol{y}^{(i)} 
ight) oldsymbol{x}_j^{(i)}$$

对logistic function求导得 g'=g(1-g)

$$\frac{\mathrm{d}}{\mathrm{d}z}g(z) = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1}{1+e^{-z}} \cdot \frac{e^{-1}+1-1}{1+e^{-z}} = \frac{1}{1+e^{-z}} \cdot \left(1 - \frac{1}{1+e^{-z}}\right) = g(z)(1-g(z))$$

for one training example,assume  $z={m w}^T{m x}$  in this, ${m w}=({m w};b)~and~{m x}=({m x};1)$ 

$$\frac{\partial J}{\partial \boldsymbol{w}_{j}} = \frac{\partial J}{\partial g(z)} \cdot \frac{\partial g(z)}{\partial z} \cdot \frac{\partial z}{\partial \boldsymbol{w}_{j}}$$

$$= -\left(y\frac{1}{g(z)} + (1-y)\frac{1}{1-g(z)} \cdot (-1)\right) \frac{\partial g(z)}{\partial z} \cdot \frac{\partial z}{\partial \boldsymbol{w}_{j}}$$

$$= -\left(y\frac{1}{g(z)} - (1-y)\frac{1}{1-g(z)}\right) g(z)(1-g(z)) \cdot x_{j}$$

$$= -(y(1-g(z)) - (1-y)g(z)) x_{j}$$

$$= -(y-g(z))x_{j}$$

$$= (\hat{y}-y)x_{j}$$

### probabilistic interpretation

Assume y satisfies Bernoulli distribution

$$P(y = 1|\boldsymbol{x}) = \hat{y}$$
  
 $P(y = 0|\boldsymbol{x}) = 1 - \hat{y}$ 

$$\Rightarrow p(y|x;w) = \hat{y}^y (1 - \hat{y})^{1-y} \tag{1}$$

将y=1带入(1)式得p(y|x)=h(x),将y=0带入(1)式得p(y|x)=1-h(x)

$$egin{aligned} L(oldsymbol{w}) &= p(oldsymbol{y} | oldsymbol{X}; oldsymbol{w}) = \prod_{i=1}^m p(oldsymbol{y}^{(i)} | oldsymbol{x}^{(i)}; oldsymbol{w}) \ \ell(oldsymbol{w}) &= \log L(oldsymbol{w}) = \sum_{i=1}^m y^{(i)} \log \hat{y}^{(i)} + (1-y^{(i)}) \log (1-\hat{y}^{(i)}) \ &\Rightarrow J(oldsymbol{w}) = -\ell(oldsymbol{w}) \quad ext{arg min } J(oldsymbol{w}) \Rightarrow ext{arg max } \ell(oldsymbol{w}) \end{aligned}$$

### 3. LDA

### PLA 与样本线性组合

#### 4. Reference

- [1] Andrew Ng, Machine Learning, https://www.coursera.org/learn/machine-learning/
- [2] 周志华, 《机器学习》, 清华大学出版社
- [3] Ian Goodfellow and Yoshua Bengio and Aaron Courville, "Deep Learning"