Linear Model

1. Linear Regression

1.1 Simple Linear Regression

- ***model formulation***

$$\hat{y}_i = wx_i + b$$

$$J\left(w, b\right) = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}_i - y_i\right)^2$$

$$\left(w^*, b^*\right) = \underset{(w,b)}{\arg\min}\, J\left(w, b\right)$$

- ***parameter estimation***

$$\begin{aligned}\frac{\partial J}{\partial w} &= \frac{2}{m}\sum_i\left(\hat{y}_i - y_i\right)\frac{\partial}{\partial w}(\hat{y}_i - y_i) = \frac{2}{m}\sum_i\left(\hat{y}_i - y_i\right)x_i \\ &= \frac{2}{m}\left(w\sum_i x_i^2 - \sum_i\left(y_i - b\right)x_i\right) \\ \frac{\partial J}{\partial b} &= \frac{2}{m}\sum_i\left(\hat{y}_i - y_i\right)\frac{\partial}{\partial b}(\hat{y}_i - y_i) = \frac{2}{m}\sum_i\left(\hat{y}_i - y_i\right) \\ &= \frac{2}{m}\left(mb - \sum_i(y_i - wx_i)\right)\end{aligned}$$

- ***closed-form solution***

$$\begin{aligned}w^* &= \frac{\sum y_i(x_i - x)}{\sum x_i^2 - \frac{1}{m}(\sum x_i)^2} \\ b^* &= \frac{1}{m}\sum_{i=1}^{m}\left(y_i - wx_i\right)\end{aligned}$$

1.2 Multiple Linear Regression

- ***model formulation***

$$\boldsymbol{w} = (\boldsymbol{w}; b) \qquad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x_1} & \boldsymbol{x_2} & \dots & \boldsymbol{x_m} \\ 1 & 1 & \dots & 1 \end{pmatrix}^T = \begin{pmatrix} \boldsymbol{x_1^T} & 1 \\ \boldsymbol{x_2^T} & 1 \\ \vdots & \\ \boldsymbol{x_m^T} & 1 \end{pmatrix}$$

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{w} = \begin{pmatrix} \hat{y_1} \\ \hat{y_2} \\ \vdots \\ \hat{y_m} \end{pmatrix} \qquad \hat{\boldsymbol{y}}^{(i)} = \boldsymbol{w}^T \boldsymbol{x}^{(i)}$$

$$\begin{aligned} J(\boldsymbol{w}) &= \frac{1}{m} \sum_{i=1}^{m} \left(\hat{\boldsymbol{y}}^{(i)} - \boldsymbol{y}^{(i)} \right)^2 = \frac{1}{m} \|\hat{\boldsymbol{y}} - \boldsymbol{y}\|^2 \\ &= \frac{1}{m} (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^T (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) \quad \mapsto [MSE_{train}] \end{aligned}$$

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} J(\boldsymbol{w})$$

> 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

$$\begin{aligned} \nabla_{\boldsymbol{w}} J(\boldsymbol{w}) = 0 \Rightarrow \nabla_{\boldsymbol{w}} (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^T (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) = 0 \\ \Rightarrow \boldsymbol{w^* = \left(X^T X\right)^{-1} X^T y} \end{aligned}$$

  - Gradient Descent

$$\begin{aligned} \frac{\partial J}{\partial \boldsymbol{w}_j} &= \frac{\partial}{\partial w_j} \left(\frac{1}{m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 \right) \\ &= \frac{1}{m} \sum_{i=1}^{m} \left(\hat{\boldsymbol{y}}^{(i)} - \boldsymbol{y}^{(i)} \right) \boldsymbol{x}_j^{(i)} \end{aligned}$$

1.3 Probabilistic interpretation for cost function

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

$$p(\boldsymbol{\epsilon}^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(\boldsymbol{\epsilon}^{(i)})^2}{2\sigma^2} \right)$$

for given $x^{(i)}$ and θ

$$p(y_i|x_i;\theta) = \frac{1}{\sqrt{2\pi}\sigma}exp\left(-\frac{\left(\boldsymbol{y}^{(i)} - \boldsymbol{\theta}^T\boldsymbol{x}^{(i)}\right)^2}{2\sigma^2}\right)$$

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

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

$$\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta})$$

$$\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^{m} \log p(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{\theta})$$

$$= \sum\left(\ln\frac{1}{\sqrt{2\pi}\sigma} - \frac{\left(\boldsymbol{y}^{(i)} - \boldsymbol{\theta}^T\boldsymbol{x}^{(i)}\right)^2}{2\sigma^2}\right)$$

$$= m\ln\frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}\sum_{i=1}^{m}\left(\hat{\boldsymbol{y}}^{(i)} - \boldsymbol{y}^{(i)}\right)^2$$

$$\arg\max \ell(\boldsymbol{\theta}) \Rightarrow \arg\min \|\hat{\boldsymbol{y}} - \boldsymbol{y}\|^2$$

```
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***

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

$$\rightarrow \ln\frac{y}{1-y} = \boldsymbol{w}^T\boldsymbol{x} + b$$

```
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)
```

```
LR模型本质上在回归真值的对数几率，即将GLM模型中的联系函数设置为logistic function
```

$$g(z)=\frac{1}{1+e^{-z}}$$

$$Cost(\hat{y},y)=-y\log\hat{y}_i-(1-y)\log(1-\hat{y})$$

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

```
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

$$\frac{\partial J}{\partial \boldsymbol{w}_j}=\frac{1}{m}\sum_{i=1}^{m}\left(\hat{\boldsymbol{y}}^{(i)}-\boldsymbol{y}^{(i)}\right)\boldsymbol{x}_j^{(i)}$$

```
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 $z=\boldsymbol{w}^T\boldsymbol{x}$ in this,$\boldsymbol{w}=(\boldsymbol{w};b)\ and\ \boldsymbol{x}=(\boldsymbol{x};1)$

$$\begin{aligned}\frac{\partial J}{\partial \boldsymbol{w}_j}&=\frac{\partial J}{\partial g(z)}\cdot\frac{\partial g(z)}{\partial z}\cdot\frac{\partial z}{\partial \boldsymbol{w}_j}\\&=-\left(y\frac{1}{g(z)}+(1-y)\frac{1}{1-g(z)}\cdot(-1)\right)\frac{\partial g(z)}{\partial z}\cdot\frac{\partial z}{\partial \boldsymbol{w}_j}\\&=-\left(y\frac{1}{g(z)}-(1-y)\frac{1}{1-g(z)}\right)g(z)(1-g(z))\cdot x_j\\&=-\left(y(1-g(z))-(1-y)g(z)\right)x_j\\&=-(y-g(z))x_j\\&=(\hat{y}-y)x_j\end{aligned}$$

- ***probabilistic interpretation***

Assume y satisfies Bernoulli distribution

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

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

```
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)
```

$$L(\boldsymbol{w}) = p(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{w}) = \prod_{i=1}^{m} p(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{w})$$

$$\ell(\boldsymbol{w}) = \log L(\boldsymbol{w}) = \sum_{i=1}^{m} y^{(i)} \log \hat{y}^{(i)} + (1-y^{(i)}) \log(1-\hat{y}^{(i)})$$
$$\Rightarrow J(\boldsymbol{w}) = -\ell(\boldsymbol{w}) \quad \arg\min J(\boldsymbol{w}) \Rightarrow \arg\max \ell(\boldsymbol{w})$$

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”