金

知识、存储

Traing set

↓

learn algorithm

↓

$x \rightarrow \boxed{h} \rightarrow y$

假设

$h(x) = \theta_0 + \theta_1 x$.

回归：多个特征

x_1, x_2

$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$

目标 $h(x) = h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x_2$. 一组

更加简单 $x_0 = 1$

$$h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T \cdot x.$$

n = 特征数目, $h_\theta(x) = \theta^T x$.

①:⇒ 定义问题.(目标. 假设函数).

$h_\theta(x) = \theta^T x$.

$\theta? \longrightarrow$ learn.

θ: parameters.

$$h_\theta(x) = \theta^T \cdot x$$

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

$$\Rightarrow \begin{cases} J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 \\ \underset{\theta}{\text{minimize}}\ J(\theta). \end{cases}$$

如何最小?

①: set: $\vec{\theta} = \vec{0}$

keep changing $\theta \longrightarrow$ reduce $J(\theta)$

Gradient Descent.

梯度下降最快的方向；

梯度下降数学部分：

Gradient Decisent.

固定 $\Rightarrow \theta_i := \theta_i - \alpha \frac{\partial}{\partial \theta_i} J(\theta)$

$$\frac{\partial}{\partial \theta_i} J(\theta) = \frac{\partial}{\partial \theta_i} \frac{1}{2} \left(h_\theta(x) - y\right)^2.$$

$$= 2 \cdot \frac{1}{2} \cdot \left(h_\theta(x) - y\right) \cdot \frac{\partial}{\partial \theta_i} \left(h_\theta(x) - y\right)$$

$$= \left(h_\theta(x) - y\right) \frac{\partial}{\partial \theta_i} \left(\theta_0 x_0 + \cdots \theta_n x_n - y\right)$$

$$= \left(h_\theta(x) - y\right) \cdot x_i$$

$$\theta_i := \theta_i - \alpha \left(h_\theta(x) - y\right) \cdot x_i$$

第i个特征的值；

一个样本→ m个样本

$$\theta_i := \theta_i - \alpha \underbrace{\sum_{j=1}^{m} (h_\theta(x^{(j)}) - y^{(j)}) \cdot x_i^{(j)}}_{\frac{\partial}{\partial \theta_i} J(\theta)}$$

检验收敛：$J(\theta) \Rightarrow$ 达到一个小值；

求导的结果就是梯度最小值；

上面的：batch GD；→遍历每个集合；

：$\boxed{\text{SGD}}$：：计表特征

与维样本

Repeat：{

For $j=1$ to m {

$\theta_i := \theta_i - \alpha (h_\theta(x^{(j)}) - y^{(j)}) \cdot x_i^{(j)}$

(For all i)

}}.

$j=1$：更新一次，整体，只对这个样本；

⋮

$j=m$ ⟶ 一共迭代 m次

而 batch GD：一个参数迭代m次，

但不会收敛到全局最小值。

最小化:

$$\nabla_\theta J = \begin{bmatrix} \frac{\partial J}{\partial \theta_0} \\ \vdots \\ \frac{\partial J}{\partial \theta_n} \end{bmatrix} \in R^{n+1}$$

GD: $\theta := \theta - \alpha \nabla_\theta J$

θ → $(n+1)$ = n个特征,

$\nabla_\theta J$ → 梯度值.

$f(A)$: $A: m \times n$ 矩阵;

$$\nabla_A f(A) = \begin{bmatrix} \frac{\partial f}{\partial A_{11}} & \cdots & \frac{\partial f}{\partial A_{1n}} \\ \vdots & & \\ \frac{\partial f}{\partial A_{m1}} & & \frac{\partial f}{\partial A_{mn}} \end{bmatrix}.$$

对 $A \in R^{n \times n}$

$trA = \sum A_{ii}$ (迹)

$trAB = trB \cdot A$;

$tr\,ABC = tr\,CAB = tr\,BCA$.

$$\begin{cases} f(A) = trAB \\ \nabla_A f(A) = \nabla_A trAB = B^T; \end{cases}$$

$trA = trA^T$;

对 $a \in R$: $tr\,a = a$;

$\nabla_A tr\,ABA^TC = CAB + C^TAB^T$;

利用公式: 简化

$$X = \begin{bmatrix} \cdots & (x^{(1)})^T & \cdots \\ \cdots & (x^{(2)})^T & \cdots \\ & \vdots & \\ & (x^{(m)})^T & \end{bmatrix} X\theta;$$

$$X\theta = \begin{bmatrix} x^{(1)T}\theta \\ \vdots \\ x^{(m)T}\cdot\theta \end{bmatrix} = \begin{bmatrix} h_\theta(x^{(1)}) \\ \vdots \\ h_\theta(x^{(m)}) \end{bmatrix}$$

$$y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}.$$

$$X\theta - y = \begin{bmatrix} h(x^{(1)}) - y^{(1)} \\ \\ h(x^{(m)}) - y^{(m)} \end{bmatrix}$$

$$\Rightarrow \frac{1}{2}(X\theta - y)^T \cdot (X\theta - y) = \frac{1}{2}\sum_{i=1}^{m} h(x^{(i)} - y^{(i)})^2.$$

$= J(\theta)$

利用矩阵推导

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

$$= \frac{1}{2}(X\theta - y)^T \cdot (X\theta - y)$$

$$(X\theta - y) = \begin{bmatrix} h(x^{(1)} - y^{(1)}) \\ \vdots \\ h(x^{(m)} - y^{(m)}) \end{bmatrix}$$

$J(\theta)$：同样的表达，但后一种用矩阵。

$\nabla_\theta J(\theta) =$

$\nabla_\theta \frac{1}{2}(X\theta-y)^T \cdot (X\theta-y)$

$= \frac{1}{2} \cdot \underset{tr}{\nabla_\theta} (\theta^T X^T \cdot X\theta - \theta^T X^T y - y^T X\theta + y^T y)$

$= \frac{1}{2}\left(\underline{\nabla_\theta \ tr\, \theta\theta^T X^T \cdot X} - \cancel{\nabla_\theta \ tr\, y^T} X\theta \right.$

$\left. \nabla_\theta tr\, y^T X\theta - \nabla_\theta y^T X\theta \right]$

$= \frac{1}{2}[X^T \cdot X\theta + X^T \cdot X\theta - X^T y - X^T y]$

$= X^T \cdot X\theta - X^T y \overset{set}{=} 0$

$\boxed{X^T \cdot X\theta = X^T y} \rightarrow$ 正规方程；

$\theta = (X^T X)^{-1} X^T y$：之前书上忽略的步骤