Linear Regression

Linear model

$$h(x) = \sum_{i=0}^{d} \theta_i x_i = \theta^T x$$

数据集 $D = \{(x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)})\}$. $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \cdots, x_d^{(i)})$

$\therefore \theta \in R^d$, $X \in R^{n \times d}$, $h(x) \in R^n$. $x_0 = 1$ (intercept term).

Cost function

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

求解方程其实有2种方法.

1° 解析方程, 多元方程组的直接求解方法.

2° 数值分析/优化的方法 ⇒ 先随机指一个初始值, 然后不断迭代优化逼近正确值.

LMS algorithm (least mean square) (Gradient Descent)

根据梯度下降, 可以使用如下公式在每一步更新参数

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta), \quad j \in (0, 1, \cdots, d)$$

1° $\frac{\partial}{\partial \theta_j} J(\theta)$: $J(\theta)$ 在 θ_j 处的梯度, 也对 $J(\theta)$ 的梯度方向变化最快, 增长最快.

2° $-\frac{\partial}{\partial \theta_j} J(\theta)$: 负梯度方向, 也对 $J(\theta)$ 下降最快, 从而得到最小的误差.

3° α: learning rate. 在梯度负方向上的步长.

假设只有一个训练样本，对θ_j求导：

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{\partial}{\partial \theta_j} \frac{1}{2}(h_\theta(x)-y)^2$$

$$= \frac{1}{2}\cdot 2\,(h_\theta(x)-y)\,\frac{\partial}{\partial \theta_j}(h_\theta(x)-y)$$

$$= (h_\theta(x)-y)\,\frac{\partial}{\partial \theta_j}\left(\sum_{j=0}^{d} \theta_j x_j - y\right)$$

$$= (h_\theta(x)-y)\,x_j \quad \Rightarrow \quad \theta_j := \theta_j + \alpha \sum_{i=1}^{n} \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}.$$

△ 对于全部的训练样本和各个维度的参数：

$$\theta := \theta + \alpha \sum_{i=1}^{n} \left[y^{(i)} - h_\theta(x^{(i)})\right] x^{(i)}$$

1° 每一次迭代，更新的量与误差大小成正比.

2° 上式称为 batch gradient descent ⇒ 每次更新考虑全部的样本.

3° 由于$J(\theta)$是一个凸函数，所以由梯度下降得到的解是全局最优解.

stochastic gradient descent

```
Loop {
    for i=1 to n, {
        θj := θj + α(y^(i) - hθ(x^(i))) xj^(i), (for every j)
    }
}
```

1° SGD 可以尽可能逼近全局最优，但无法保证收敛到全局最优

2° 通过逐步降低α至0，可以达到全局最优.

normal equations

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

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

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

$$= \frac{1}{2}\nabla_\theta\left[(X\theta)^T X\theta - (X\theta)^T y - y^T X\theta + y^T y\right]$$

$$= \frac{1}{2}\nabla_\theta\left[\theta^T X^T X\theta - \theta^T X^T y - y^T X\theta\right]$$

$$= \frac{1}{2}\nabla_\theta\left[\theta^T X^T X\theta - 2y^T X\theta\right]$$

$$= \frac{1}{2}\nabla_\theta\left[\theta^T X^T X\theta - 2(X^T y)^T\theta\right]$$

$\frac{\partial \theta^T(X^TX\theta)}{\partial\theta}$ ↙ ↘ $\frac{(\theta^T X^T X)\theta \to \text{实数}}{\partial\theta}$

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

$$= X^T X\theta - X^T y = 0$$

$$\Rightarrow X^T X\theta = X^T y \quad \Rightarrow \quad \theta = (X^T X)^{-1} X^T y$$

Probabilistic Interpretation (maximum likelihood)

假设 $y^{(i)}$ 和 $x^{(i)}$ 之间存在线性关系：

$$y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}$$

若误差 $\varepsilon^{(i)}$ 服从均值为0的正态分布，$\varepsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$，则：

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

则：$y^{(i)} \mid x^{(i)}; \theta \sim \mathcal{N}(\theta^T x^{(i)}, \sigma^2)$：

$$p(y^{(i)} \mid x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi}\sigma^3} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

则有似然函数：

$$L(\theta) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

对数似然：

$$l(\theta) = \log L(\theta) = \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$= n \log \frac{1}{\sqrt{2\pi}\sigma^2} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^{n} (y^{(i)} - \theta^T x^{(i)})^2$$

⇒最大化对数似然，等价于最小化这个量，且最优的 θ 值与 σ^2 的取值无关.

$$\frac{\partial l(\theta)}{\partial \theta} = 0 - \frac{1}{\sigma^2} \cdot \frac{1}{2} \cdot 2 \sum_{i=1}^{n} (y^{(i)} - \theta^T x^{(i)}) \cdot (0 - x^{(i)}) = 0$$

$$\Rightarrow \sum_{i=1}^{n} (y^{(i)} - \theta^T x^{(i)}) \cdot x^{(i)} = 0 \quad \Rightarrow X^T y = (X^T X) \cdot \theta$$

$$\Rightarrow \theta = (X^T X)^{-1} X^T y$$

Locally weighted linear regression

损失函数调整为：

$$J(\theta)=\sum_{i=1}^{n} w^{(i)}\left(y^{(i)}-\theta^T x^{(i)}\right)^2$$

（$w^{(i)}$：赋给权重）

$w^{(i)}$ 通常取值为：

$$w^{(i)}=\exp\left[-\frac{(x^{(i)}-x)^2}{2\tau^2}\right]$$

（x：预测值；τ：带宽（bandwidth））

⇒ 高斯核函数

1° 当 $x^{(i)}$ 与 x 的距离 $|x^{(i)}-x|$ 很大时，$w^{(i)}$ 很小，此时 $J(\theta)$ 几乎忽略 $y^{(i)}-\theta^T x^{(i)}$.

2° 当 $x^{(i)}$ 与 x 的距离 $|x^{(i)}-x|$ 很小时，$w^{(i)}=1$，此时 $J(\theta)$ 几乎是 $y^{(i)}-\theta^T x^{(i)}$ 的最小化误差.

方程求解可得：

$$J(\theta)=w(y-x\theta)^T(y-x\theta)$$

$$\frac{\partial J(\theta)}{\partial\theta}=\frac{\partial}{\partial\theta}w(y-x\theta)^T(y-x\theta)$$

$$=\frac{\partial}{\partial\theta}\left[w(y^T-\theta^T x^T)(y-x\theta)\right]$$

$$=\frac{\partial}{\partial\theta}\left[(wy^T-w\theta^T x^T)(y-x\theta)\right]$$

$$=\frac{\partial}{\partial\theta}\left(wy^Ty-wy^Tx\theta-w\theta^Tx^Ty+w\theta^Tx^Tx\theta\right)$$

$$=(0-wy^Tx-wy^Tx+w\theta^Tx^Tx+w\theta^Tx^Tx)$$

$$=-2wy^Tx+2w\theta^Tx^Tx=0$$

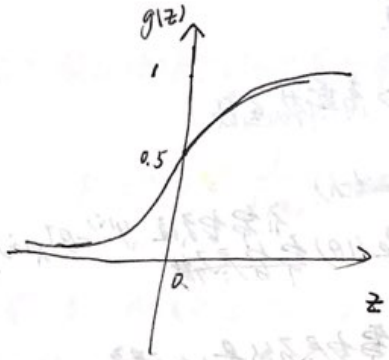
$$\Rightarrow w\theta^Tx^Tx=wy^Tx \Rightarrow w\theta^T=wy^Tx(x^Tx)^{-1} \Rightarrow \theta w^T=wy^Tx(x^Tx)^{-1}$$

$$\Rightarrow \theta=wy^Tx(x^Tx)^{-1}(w^T)^{-1}=wy^Tx(w^Tx^Tx)^{-1}$$

Logistic Regression ⇒ binary classification $\{0,1\}$

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}, \quad g(z) = \frac{1}{1+e^{-z}}.$$

↘ sigmoid function.



Sigmoid 函数在 0 附近的斜率很大，可以将一些临界值映射成 0，避免分类错误.

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

$$= \frac{-1}{(1+e^{-z})^2}\cdot(e^{-z})(-1)$$

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

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

$$= g(z)\cdot(1-g(z)).$$

Probabilistic interpret.

假设 $p(y=1|x;\theta) = h_\theta(x)$.

$p(y=0|x;\theta) = 1-h_\theta(x)$

$\therefore p(y|x;\theta) = (h_\theta(x))^y (1-h_\theta(x))^{1-y}$ ⇒ 伯努利分布.

$\therefore L(\theta) = \prod_{i=1}^{n} p(y^{(i)}|x^{(i)};\theta) = \prod_{i=1}^{n} (h_\theta(x^{(i)}))^{y^{(i)}} (1-h_\theta(x^{(i)}))^{1-y^{(i)}}$.

$\therefore l(\theta) = \log L(\theta) = \sum_{i=1}^{n} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)}))$

假设有

对于一个样本，对θ_j求导：

$$\frac{\partial}{\partial\theta_j} l(\theta) = \frac{\partial}{\partial\theta_j} y\log h_\theta(x) + (1-y)\log(1-h_\theta(x))$$

$$= \left[y\frac{1}{h_\theta(x)} \cdot h_\theta(x)\cdot(1-h_\theta(x)) + (1-y)\frac{1}{1-h_\theta(x)}\cdot(-1)\cdot h_\theta(x)(1-h_\theta(x))\right]\cdot\frac{\partial\theta^T x}{\partial\theta_j}$$

$$= \left[y(1-h_\theta(x)) + (y-1)h_\theta(x)\right]\cdot\frac{\partial\theta^T x}{\partial\theta_j}$$

$$= (y - h_\theta(x))\cdot x_j$$

所以，对于随机梯度下降有：

$$\theta_j := \theta_j + \alpha\left(y^{(i)} - h_\theta(x^{(i)})\right)x_j^{(i)}$$

Newton's method:

$$\theta := \theta - \frac{f(\theta)}{f'(\theta)}, \qquad f: \mathbb{R}\to\mathbb{R}.$$

~~maximum $f(\theta)$ $\Rightarrow$ find θ which makes $f(\theta)=0$.~~

maximum $l(\theta)$ equals find θ^* makes $l'(\theta^*)=0$.

$$\therefore\quad \theta := \theta - \frac{l'(\theta)}{l''(\theta)}$$

考虑向量表示：

$$\theta := \theta - H^{-1}\nabla_\theta l(\theta), \text{ 其中 } H_{ij} = \frac{\partial^2 l(\theta)}{\partial\theta_i\partial\theta_j}.$$

Hessian Matrix ↑

↙ Newton-Raphson method.

当牛顿法被应用于最大化逻辑回归的似然函数时，得到的函数是 Fisher Scoring.

Generalized Linear Models

exponential family

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

- η: natural parameter.
- $T(y)$: efficient statistic.
- $a(\eta)$: log partition function.

Bernoulli distribution

$$p(y;\phi) = \phi^y (1-\phi)^{1-y}$$

$$= \exp(y\log\phi + (1-y)\log(1-\phi))$$

$$= \exp\left(y\log\frac{\phi}{1-\phi} + \log(1-\phi)\right)$$

- $b(y)$
- $T(y)$
- $\eta = \log\frac{\phi}{1-\phi}$
- $a(\eta)$

$$\Rightarrow \begin{cases} b(y) = 1 \\ T(y) = y \\ a(\eta) = -\log(1-\phi) = \log(1+e^{\eta}) \end{cases}$$

$$\eta = \log\frac{\phi}{1-\phi}$$

$$\therefore \frac{\phi}{1-\phi} = e^{\eta}$$

$$\therefore \phi = (1-\phi)e^{\eta}$$

$$\therefore \phi(1+e^{\eta}) = e^{\eta}$$

$$\therefore \phi = \frac{1}{1+e^{-\eta}}$$

$$1-\phi = \frac{1+e^{\eta}-e^{\eta}}{1+e^{\eta}} \quad \log\frac{\phi}{e^{\eta}}$$

Gaussian distribution.

假设 $\sigma^2 = 1$, 则有:

$$p(y;\eta) = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(y-u)^2\right]$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)\exp\left(uy - \frac{1}{2}u^2\right)$$

- $b(y)$
- η
- $T(y)$
- $a(\eta)$

$$\therefore \begin{cases} b(y) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right) \\ \eta = u \\ T(y) = y \\ a(\eta) = \frac{1}{2}u^2 = \frac{1}{2}\eta^2 \end{cases}$$

构建GLMs

广义线性模型基于以下3个假设：

1. 给定 x 和 θ，y 的分布服从指数家族分布：$y|x;\theta \sim ExponentialFamily(\eta)$

2. 给定 x，GLMs 的目标是使得学习到的假设 $h(x) = E[y|x]$

3. η 和 x 是线性相关：$\eta = \theta^T x$

线性回归（Least Square）（Gaussian Distribution.）

$h_\theta(x) = E[y|x;\theta]$ ⇒上面的假设(2).

$= \mu$ ⇒ $y|x;\theta$ 服从高斯分布.

$= \eta$ ⇒ 高斯分布属于指数家族（$\eta = \mu$）

$= \theta^T x$ ⇒ ~~$y|x;\theta \sim \mathcal{N}(\theta^T x, \sigma^2)$~~. /上面的假设(3)

逻辑回归（Logistic Regression）（Bernoulli Distribution）

$h_\theta(x) = E[y|x;\theta]$ ⇒ 上面的假设(2).

$= \phi$ ⇒ $y|x;\theta$ 服从伯努力分布.

$= \frac{1}{1+e^{-\eta}}$ ⇒ 伯努力分布属于指数家族.

$= \frac{1}{1+e^{-\theta^T x}}$ ⇒上面的假设(3)

Softmax Regression

k分类问题：$y \in \{1, 2, \cdots, k\}$，使用$\phi_i$表示$y$取$i$类别的概率。

实际上，我们使用$k-1$个自由参数来建模，因为所有概率和为1.

$T(y) \in \mathbb{R}^{k-1}$

$$T_1(y) = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \cdots, \quad T_{k-1}(y) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad T_k(y) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$$

使用多项式建模：$y|x;\phi \sim Mult(\phi_1, \cdots \phi_k)$

$\therefore E[T_i(y)] = P(y=i) = \phi_i$

证明多项式属于指数族：

$$\begin{aligned} \because P(y;\phi) &= \phi_1^{I(y=1)} \phi_2^{I(y=2)} \cdots \phi_k^{T(y=k)} \\ &= \phi_1^{T_1(y)} \phi_2^{T_2(y)} \cdots \phi_k^{T_k(y)} \\ &= \phi_1^{T_1(y)} \phi_2^{T_2(y)} \cdots \phi_{k-1}^{T_{k-1}(y)} \phi_k^{1-\sum_{i=1}^{k-1} T_i(y)} \\ &= \exp\left[T_1(y) \log\phi_1 + T_2(y) \log\phi_2 + \cdots + T_{k-1}(y) \log\phi_{k-1} + (1 - \sum_{i=1}^{k-1} T_i(y)) \log\phi_k \right] \\ &= \exp\left[T_1(y) \log\frac{\phi_1}{\phi_k} + T_2(y) \log\frac{\phi_2}{\phi_k} + \cdots + T_{k-1}(y) \log\frac{\phi_{k-1}}{\phi_k} + \log\phi_k \right] \end{aligned}$$

$$\therefore \begin{cases} b(y) = 1 \\ \eta = \begin{bmatrix} \log(\phi_1/\phi_k) \\ \log(\phi_2/\phi_k) \\ \vdots \\ \log(\phi_{k-1}/\phi_k) \end{bmatrix} \end{cases}$$

~~$T(y) = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^{k-1}$~~

$a(\eta) = -\log\phi_k$

设 η_i 表示 η 中的第 i 个元素，则 $\eta_i = \log \frac{\phi_i}{\phi_k}$，且 $\eta_k = \log \frac{\phi_k}{\phi_k} = 0$

$$\therefore e^{\eta_i} = \frac{\phi_i}{\phi_k}$$

$$\therefore \phi_i = \phi_k e^{\eta_i}$$

$$\because \sum_{i=1}^{k} \phi_i = \sum_{i=1}^{k} \phi_k e^{\eta_i} = 1$$

$$\therefore \phi_k = \frac{1}{\sum_{i=1}^{k} e^{\eta_i}} \Rightarrow \phi_i = \phi_k e^{\eta_i} = \frac{e^{\eta_i}}{\sum_{i=1}^{k} e^{\eta_i}}$$

$$\therefore \cancel{\phi_i = \frac{\phi_i}{\phi_k}} =$$

$$\therefore p(y=i|x;\theta) = \phi_i = \frac{e^{\eta_i}}{\sum_{i=1}^{k} e^{\eta_i}} = \frac{e^{\theta_i^T x}}{\sum_{i=1}^{k} e^{\theta_i^T x}}$$

$h_\theta(x) = E[T(y)|x;\theta]$ ⇒ 假设(2)，对 $T(y)|x;\theta$ 求期望

$$= E\left[\begin{matrix} I(y=1) \\ \vdots \\ I(y=k-1) \end{matrix} \middle| x;\theta\right]$$

⇒ $T(y)|x;\theta$ 服从多项式分布.

$$= \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_{k-1} \end{bmatrix}$$

$$= \begin{bmatrix} e^{\theta_1^T x} / \sum_{i=1}^{k} e^{\theta_i^T x} \\ \vdots \\ e^{\theta_{k-1}^T x} / \sum_{i=1}^{k} e^{\theta_i^T x} \end{bmatrix}$$

↓ 多项式分布属于指数族.

且 假设(3) $\eta_i = \theta_i^T x$

⇒ Softmax regression.