Logistic Regression

Friday, April 10, 2020 5:09 PM

$$D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)})\}$$

$$x^{(i)} = \begin{bmatrix} 1 \\ x_1^{(i)} \\ \vdots \\ x_d^{(i)} \end{bmatrix} \in R^{d+1} \qquad y^{(i)} \in \{0, 1\}$$

$$\begin{aligned} \sigma(z)' &= [(1+e^{-z})^{-1}]' \\ &= -1(1+e^{-z})^{-2} \cdot (1+e^{-z})' \\ &= -(1+e^{-z})^{-2} \cdot e^{-z} \cdot (-z)' \\ &= -(1+e^{-z})^{-2} \cdot e^{-z} \cdot (-1) \\ &= -\underbrace{\frac{1}{1+e^{-z}}}_{} \cdot \underbrace{\frac{e^{-z}}{1+e^{-z}}}_{} \cdot (-1) \\ &= \sigma(z) \cdot (1-\sigma(z)) \end{aligned}$$

Logistc Function: $\sigma(z) = \dfrac{1}{1+e^{-z}}$

导数 $\sigma'(z) = \sigma(z) \cdot (1 - \sigma(z))$

模型参数 $w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}$ $w_0 \rightarrow$ bias

定义概率

$$p(y=1 \mid x; w) = \sigma(w^T \cdot x)$$

$$p(y=0 \mid x; w) = 1 - \sigma(w^T \cdot x)$$

单个样本的统一概率表示:

$$p(y \mid x; w) = \sigma(w^T \cdot x)^{y} \cdot (1 - \sigma(w^T \cdot x))^{(1-y)}$$

数据集似然，（IID假设）

$$L(D) = \prod_{i=1}^{N} p(y^{(i)} \mid x^{(i)}; w) = \prod_{i=1}^{N} \sigma(w^T x^{(i)})^{y^{(i)}} \cdot (1 - \sigma(w^T x^{(i)}))^{(1-y^{(i)})}$$

Log 似然

$$\begin{aligned} \mathcal{L}(D) &= \sum_{i=1}^{N} \log p(y^{(i)} \mid x^{(i)}; w) \\ &= \sum_{i=1}^{N} \log \left\{ \sigma(w^T \cdot x^{(i)})^{y^{(i)}} \cdot (1 - \sigma(w^T \cdot x^{(i)}))^{(1-y^{(i)})} \right\} \\ &= \sum_{i=1}^{N} \left\{ y^{(i)} \log \sigma(w^T \cdot x^{(i)}) + (1-y^{(i)}) \log(1 - \sigma(w^T x^{(i)})) \right\} \end{aligned}$$

损失函数 Negative log likelihood (NLL)

$$\mathcal{L}(w) = -\sum_{i=1}^{N} \left\{ y^{(i)} \log \sigma(w^T \cdot x^{(i)}) + (1-y^{(i)}) \log(1 - \sigma(w^T \cdot x^{(i)})) \right\}$$

通过判断Hessian矩阵是否半正定验证Logistic Regression的损失函数为凸函数。

为简化Notation，我们用偷懒的方式：

可以先验证单个数据的损失函数是凸函数，然后根据凸函数加法的保凸性，可知完整的损失函数也是凸函数。

单个样本损失函数：

$$l(w) = -y\log\sigma(w^T\cdot x) - (1-y)\log(1-\sigma(w^T\cdot x))$$

一阶导数：

$$\frac{\partial l(w)}{\partial w} = -y\frac{\partial\log\sigma(w^T\cdot x)}{\partial w} - (1-y)\cdot\frac{\partial\log(1-\sigma(w^T\cdot x))}{\partial w}$$

$$= -y\frac{1}{\sigma(w^T\cdot x)}\cdot\frac{\partial\sigma(w^T\cdot x)}{\partial w} - (1-y)\frac{1}{1-\sigma(w^T\cdot x)}\cdot\frac{\partial(1-\sigma(w^T\cdot x))}{\partial w}$$

$$= -y\frac{1}{\sigma(w^T\cdot x)}\cdot\sigma(w^T\cdot x)\cdot(1-\sigma(w^T\cdot x))\cdot x$$
$$- (1-y)\cdot\frac{1}{1-\sigma(w^T\cdot x)}\cdot[-\sigma(w^T x)(1-\sigma(w^T\cdot x))]\cdot x$$

$$= -y(1-\sigma(w^T\cdot x))x + (1-y)\sigma(w^T\cdot x)x$$

$$= -yx + \sigma(w^T\cdot x)y\cdot x + \sigma(w^T x)\cdot x - y\sigma(w^T\cdot x)\cdot x$$

$$= (\sigma(w^T\cdot x) - y)\cdot x$$

$$\frac{\partial l(w)}{\partial w} = \underbrace{(\sigma(w^T\cdot x) - y)}_{\in R}\cdot\underbrace{x}_{\in R^{d+1}}$$

二阶导数（向量对向量求导，结果为Hessian矩阵）

$$\frac{\partial^2 l(w)}{\partial w\partial w^T} = \frac{\partial[(\sigma(w^T x)-y)\cdot x]}{\partial w^T}$$

$$= \begin{bmatrix} \frac{\partial[(\sigma(w^T\cdot x)-y)\cdot x_0]}{\partial w^T} \\ \frac{\partial[(\sigma(w^T\cdot x)-y)\cdot x_1]}{\partial w^T} \\ \vdots \\ \frac{\partial[(\sigma(w^T\cdot x)-y)\cdot x_d]}{\partial w^T} \end{bmatrix} = \begin{bmatrix} \frac{\partial[(\sigma(w^T x)-y)\cdot x_0]}{\partial w_0}, & \frac{\partial[(\sigma(w^T x)-y)\cdot x_0]}{\partial w_1}, & \cdots \\ \frac{\partial[(\sigma(w^T x)-y)\cdot x_1]}{\partial w_0}, & \frac{\partial[(\sigma(w^T x)-y)\cdot x_1]}{\partial w_1}, & \cdots \\ \vdots & \ddots & \vdots \\ \frac{\partial[(\sigma(w^T x)-y)\cdot x_d]}{\partial w_0}, & \frac{\partial[(\sigma(w^T x)-y)\cdot x_d]}{\partial w_1}, & \cdots \end{bmatrix}$$

$$= \begin{bmatrix} \sigma(w^T\cdot x)(1-\sigma(w^T\cdot x))x_0^2, & \sigma(w^T\cdot x)(1-\sigma(w^T\cdot x))x_0\cdot x_1, & \cdots \\ \sigma(w^T\cdot x)(1-\sigma(w^T\cdot x))x_1\cdot x_0, & \sigma(w^T\cdot x)(1-\sigma(w^T\cdot x))x_1^2, & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

$$= \begin{bmatrix} \cdots & & & \\ \vdots & \vdots & \ddots & \\ \sigma(w^T\cdot x)(1-\sigma(w^T\cdot x))x_d x_0, & \sigma(w^T\cdot x)(1-\sigma(w^T\cdot x))\cdot x_d x_1, & \cdots & \end{bmatrix}$$

$$= \underset{(d+1)\times 1}{\begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}} \cdot \underset{1\times 1}{\underbrace{\left[\sigma(w^T\cdot x)(1-\sigma(w^T\cdot x)) \right]}} \cdot \underset{1\times(d+1)}{\left[x_0, x_1, \cdots, x_d \right]}$$

$$= x \cdot \sigma(w^T\cdot x)(1-\sigma(w^T\cdot x)) \cdot x^T$$

$$\Rightarrow \frac{\partial^2 l(w)}{\partial w \cdot \partial w^T} = x\cdot\sigma(w^T\cdot x)\cdot(1-\sigma(w^T\cdot x))\cdot x^T$$

令 $u \in R^{d+1}$ 则

$$u^T\cdot\frac{\partial^2 l(w)}{\partial w \partial w^T} u = u^T\cdot x\cdot\sigma(w^T\cdot x)\cdot(1-\sigma(w^T x))\cdot x^T\cdot u$$

$$= \underbrace{(x^T u)^T\cdot(x^T u)}_{\geqslant 0}\cdot\underbrace{\sigma(w^T\cdot x)\cdot(1-\sigma(w^T\cdot x))}_{>0}$$

$$\geqslant 0$$

$$\Rightarrow \frac{\partial^2 l(w)}{\partial w\cdot\partial w^T} \succcurlyeq 0$$

$\therefore$ $l(w)$ 为凸函数

$\therefore$ $L(w)=\sum_{i=1}^{N} l_i(w)$ 也为凸函数

不使用保凸性，也可由单个样本的损失得到整体损失函数的Hassian矩阵，直接对整体损失函数判断半正定属性。

$$\frac{\partial^2 L(w)}{\partial w \partial w^T} = \sum_{i=1}^{N} x^{(i)} \sigma(w^T \cdot x^{(i)})(1-\sigma(w^T \cdot x^{(i)})) x^{(i)T}$$

$$= \underbrace{\left[x^{(1)}, x^{(2)}, \dots, x^{(N)}\right]}_{(D+1)\times N} \underbrace{\begin{bmatrix} \sigma(w^T x^{(1)})(1-\sigma(w^T x^{(1)})) & & \\ & \sigma(w^T x^{(2)})(1-\sigma(w^T x^{(2)})) & \\ & & \ddots \\ & & & \sigma(w^T x^{(N)})\cdot(1-\sigma(w^T x^{(N)})) \end{bmatrix}}_{N\times N} \underbrace{\begin{bmatrix} x^{(1)T} \\ x^{(2)T} \\ \vdots \\ x^{(N)T} \end{bmatrix}}_{N\times(D+1)}$$

$$= X \; D \; X^T \qquad (D_{ii} = \sigma(w^T x^{(i)})(1-\sigma(w^T x^{(i)})))$$

$\forall u \in R^{(D+1)}$:

$$u^T \cdot X \cdot D \cdot X^T \cdot u = (X^T u)^T \cdot D \, (X^T \cdot u) \geq 0$$

$\because \dfrac{\partial^2 L(w)}{\partial w \partial w^T} \succcurlyeq 0 \quad \therefore L(w)$ 为凸函数