Logistic Regression

用于分类，主要是得到预测结果的概率值：

$$f(x) = P(y|x) \in [0,1]$$

给定数据集$D = \left\{ \left(\overrightarrow{x_1}, y_1\right), \left(\overrightarrow{x_2}, y_2\right), \ldots, \left(\overrightarrow{x_n}, y_n\right) \right\}$，$y_i \in \{-1, +1\}$

预测函数：

$$s(z) = \frac{1}{1 + exp(-z)}$$

$$z = \sum w_i \overrightarrow{x_i} = \overrightarrow{w}^T \overrightarrow{x}$$

$$h(x) = \frac{1}{1 + exp\left(-\overrightarrow{w}^T \overrightarrow{x}\right)}$$

Error function 通过极大似然估计得到：

目标函数$f(x)$下，产生D的概率：

$$P_f(D) = P\left(\overrightarrow{x_1}\right) f\left(\overrightarrow{x_1}\right) P\left(\overrightarrow{x_2}\right) f\left(\overrightarrow{x_2}\right) \ldots P\left(\overrightarrow{x_n}\right) f\left(\overrightarrow{x_n}\right)$$

预测函数$h(x)$下，产生D的概率：

$$P_h(D) = P\left(\overrightarrow{x_1}\right) h\left(\overrightarrow{x_1}\right) P\left(\overrightarrow{x_2}\right) h\left(\overrightarrow{x_2}\right) \ldots P\left(\overrightarrow{x_n}\right) h\left(\overrightarrow{x_n}\right)$$

若h 与f 很接近，则 $P_h(D)$ 与$P_f(D)$ 很接近，称$P_h(D)$ 为似然。

f既然产生了D，可以想成“f产生D 的概率极大”（所以f产生了D）；

因此，最佳的h，与f最接近的h，其产生D的概率也该是极大的。

所以，要求最大似然$max\ likelihood(h)$，从而得到最佳的预测函数$arg\max_h P_h(D)$.

$$P_h(D) = P\left(\overrightarrow{x_1}\right) P\left(y_1 | \overrightarrow{x_1}\right) P\left(\overrightarrow{x_2}\right) P\left(y_2 | \overrightarrow{x_2}\right) \ldots P\left(\overrightarrow{x_n}\right) P\left(y_n | \overrightarrow{x_n}\right)$$

$$\because P\left(y_i \mid \overrightarrow{x_i}\right) = \begin{cases} h\left(\overrightarrow{x_i}\right), & for\ \ y_i = +1 \\ 1 - h\left(\overrightarrow{x_i}\right), & for\ \ y_i = -1 \end{cases}$$

又 $\because$ $logistic$函数有性质：$1 - h(x) = h(-x)$

$$\therefore P\left(y_i \mid \overrightarrow{x_i}\right) = \begin{cases} h(x), & for\ y_i = +1 \\ h(-x), & for\ y_i = -1 \end{cases} = h(y_i x_i)$$

$$\therefore P_h(D) = P\left(\overrightarrow{x_1}\right)h\left(y_1\overrightarrow{x_1}\right) P\left(\overrightarrow{x_2}\right)h\left(y_2\overrightarrow{x_2}\right) \ldots P\left(\overrightarrow{x_n}\right)h\left(y_n \mid \overrightarrow{x_n}\right)$$

$\because P\left(\overrightarrow{x_i}\right)$固定，所以上式 $\propto \prod h\left(y_i\ \overrightarrow{x_i}\right)$

$$\therefore \max_h P_h(D) \ \propto \prod h\left(y_i\ \overrightarrow{x_i}\right)$$

$\because$ 每个h 对应一个参数$\overrightarrow{w}$， $\therefore \max_{\overrightarrow{w}} likelihood\left(\overrightarrow{w}\right) \propto \prod s\left(y_i\ \overrightarrow{w}^T \overrightarrow{x_i}\right)$

$\because$ 连乘不容易求解，所以取对数变成连加：$\max_{\overrightarrow{w}} ln \prod s\left(y_I \overrightarrow{w}^T\ \overrightarrow{x_i}\right) = \max_h \sum ln\ s\left(y_i \overrightarrow{w}\ \overrightarrow{x_i}\right)$

$\because$ 似然最大，即误差最小，目的也是求误差函数，所以max变为min 加一个负号：

$$\min_{\overrightarrow{w}} \frac{1}{N} \sum -ln\ s\left(y_i \overrightarrow{w}\ \overrightarrow{x_i}\right)$$

$\therefore s(z) = \dfrac{1}{1+exp(-z)}$ 代入上式得到：$\min_{\overrightarrow{w}} \dfrac{1}{N} \sum ln\left(1 + exp\left(-y_i\ \overrightarrow{w}^T\ \overrightarrow{x_i}\right)\right)$

其中，定义交叉熵错误$cross\ entropy\ error : err\left(\overrightarrow{w}, \overrightarrow{x_i}, y_i\right) = ln(1 + exp\left(-y_i \overrightarrow{w}^T\ \overrightarrow{x_i}\right)$

综上，$error\ function$: $\dfrac{1}{N} \sum ln\left(1 + exp\left(-y_i \overrightarrow{w}^T \overrightarrow{x_i}\right)\right) = \dfrac{1}{N} \sum err\left(\overrightarrow{w}, \overrightarrow{x_i}, y_i\right)$

如何解min error function得到最佳$\overrightarrow{w}$：梯度及梯度下降

$$E\left(\overrightarrow{w}\right) = \frac{1}{N} \sum ln\left(1 + exp\left(-y_i \overrightarrow{w}^T \overrightarrow{x_i}\right)\right)$$

$$\begin{aligned} \nabla E\left(\overrightarrow{w}\right) = \frac{\partial E\left(\overrightarrow{w}\right)}{\partial w_i} &= \frac{1}{N} \sum \left(\frac{1}{\square}\right)(exp(\triangle))(-y_i\ \overrightarrow{x_i}) \\ &= \frac{1}{N} \sum \left(\frac{exp(\triangle)}{1 + exp(\triangle)}\right)\left(-y_i\ \overrightarrow{x_i}\right) \\ &= \frac{1}{N} \sum s(\triangle)\left(-y_i\ \overrightarrow{x_i}\right) \\ &= \frac{1}{N} \sum s\left(-y_i \overrightarrow{w}^T \overrightarrow{x_i}\right)\left(-y_i\ \overrightarrow{x_i}\right) \end{aligned}$$

因为$E\left(\overrightarrow{w}\right)$是连续，可微的凸函数，所以令$\nabla E\left(\overrightarrow{w}\right) = 0$，可解的最佳$\overrightarrow{w}$。

另外有迭代优化方法来得到最佳$\vec{w}$:

1. 设置权值向量$\vec{w}$初始值为$\overrightarrow{w_0}$，设迭代次数为t，$t = 0, 1, 2, \ldots$;

2. 计算梯度$\nabla E\left(\overrightarrow{w_t}\right) = \frac{1}{N}\sum s\left(-y_i \overrightarrow{w_t}^T \overrightarrow{x_i}\right)\left(-y_i \overrightarrow{x_i}\right)$;

3. 更新$\vec{w}$：$\overrightarrow{w_{t+1}} = \overrightarrow{w_t} - \eta \nabla E\left(\overrightarrow{w_t}\right)$

4. 直到$\nabla E\left(\overrightarrow{w_t}\right) \approx 0$或迭代次数足够多