根据贝叶斯法则，可得

$$\begin{aligned} likelihood(logistic \quad h) &= L(\boldsymbol{w}) \\ &\propto \prod_{n=1}^{N} P(y_n|\boldsymbol{x}_n, \boldsymbol{w}) \\ &= \prod_{i=1}^{N} h(\boldsymbol{x}_n)^{y_n}(1-h(\boldsymbol{x}_n))^{1-y_n} \end{aligned} \tag{2.7}$$

故我们只需要让此似然函数取得最大值即可

$$\max_{\boldsymbol{w}} \quad L(\boldsymbol{w}) \tag{2.8}$$

为了后续求解方便，我们将上述问题做一些变换，上面最大化的问题等价于如下的最小化的问题。

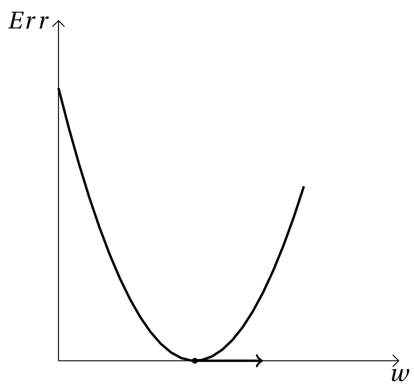
$$\min_{\boldsymbol{w}} \quad -\log L(\boldsymbol{w}) \tag{2.9}$$

主要做了两个处理，取对数不改变函数的极值点和最优解，添加了一个负号，将最大化问题转换为了一个最小化问题。这个新的误差函数在统计学叫做为交叉熵，交叉熵误差为

$$\begin{aligned} \min_{\boldsymbol{w}} \quad Err(\boldsymbol{w}) &= -\log \prod_{i=1}^{N} h(\boldsymbol{x}_n)^{y_n}(1-h(\boldsymbol{x}_n))^{1-y_n} \\ &= -\sum_{n=1}^{N} y_n \log(h(\boldsymbol{x}_n)) + (1-y_n)\log(1-h(\boldsymbol{x}_n)) \end{aligned} \tag{2.10}$$

log是以$e$为底。

由于误差函数$Err(\boldsymbol{w})$是一个连续可导，并且二阶可微的凸函数。根据凸优化理论，存在全局最优解，即为$\nabla Err(\boldsymbol{w}) = 0$



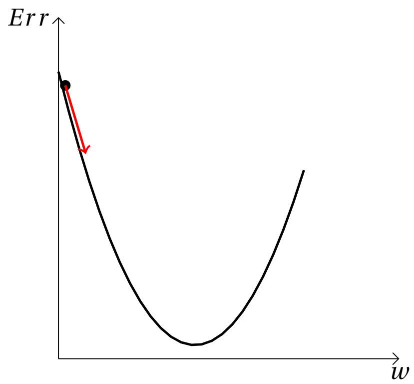
故我们首先要推导得到$\nabla Err(w_i)$，令$u = 1 + e^{-\boldsymbol{w}^T \boldsymbol{x}_n}$和$v = -\boldsymbol{w}^T \boldsymbol{x}_n$，则

$$\begin{aligned}
\frac{\partial Err(w_i)}{\partial w_i} &= -\sum_{n=1}^{N}\left[(y_n)\left(\frac{\partial \log(h(\boldsymbol{x}_n))}{\partial h(\boldsymbol{x}_n)}\right)\left(\frac{\partial h(\boldsymbol{x}_n)}{\partial u}\right)\left(\frac{\partial u}{\partial v}\right)\left(\frac{\partial v}{\partial w_i}\right) + (1-y_n)\left(\frac{\partial \log(1-h(\boldsymbol{x}_n))}{\partial h(\boldsymbol{x}_n)}\right)\left(\frac{\partial h(\boldsymbol{x}_n)}{\partial u}\right)\left(\frac{\partial u}{\partial v}\right)\left(\frac{\partial v}{\partial w_i}\right)\right] \\
&= -\sum_{n=1}^{N}\left[(y_n)\left(\frac{1}{h(\boldsymbol{x}_n)}\right) + (1-y_n)\left(\frac{-1}{1-h(\boldsymbol{x}_n)}\right)\right]\left[\left(\frac{-1}{u^2}\right)(e^v)(-x_{n,i})\right] \\
&= -\sum_{n=1}^{N}\left[(y_n)\left(\frac{1}{h(\boldsymbol{x}_n)}\right) - (1-y_n)\left(\frac{1}{1-h(\boldsymbol{x}_n)}\right)\right][h(\boldsymbol{x}_n)(1-h(\boldsymbol{x}_n))](x_{n,i}) \\
&= -\sum_{n=1}^{N}\left[(y_n)(1-h(\boldsymbol{x}_n)) - (1-y_n)h(\boldsymbol{x}_n))\right](x_{n,i}) \\
&= -\sum_{n=1}^{N}(y_n - h(\boldsymbol{x}_n))(x_{n,i})
\end{aligned} \tag{2.11}$$

故$Err(w_i)$的梯度如下：

$$\nabla Err(w_i) = \sum_{n=1}^{N}\left(\frac{1}{1+e^{-\boldsymbol{w}^T \boldsymbol{x}}} - y_n\right)(x_{n,i}) \tag{2.12}$$

不幸的是，这个梯度表达式为一个非线性函数，故要求解函数零点非常困难，故我们采用了迭代最优化的方式去求解。由于$Err(\boldsymbol{w})$是一个凸函数，故我们只要沿着梯度下降的方向去更新求解$\boldsymbol{w}$，就一定能较迅速找到最优解，因为梯度是函数变化最快的方向。




假设第$t$步我们已经得到了权重$\boldsymbol{w}_t$，那么，根据梯度下降法，我们可以得到如下更新公式：

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \nabla Err(\boldsymbol{w}_t) \tag{2.13}$$

其中，$\eta > 0$表示梯度下降的步长，为人工设置参数。

于是逻辑回归算法去求解分类问题如下：

**Algorithm 1** 逻辑回归训练算法

**输入**: 特征向量集合$\{\boldsymbol{x}\}$和标签集合$\{y\}$

**输出**: 最优解$\boldsymbol{w}_{t+1}$

**初始化**: 随机初始化$\boldsymbol{w}_0$

$for \quad t = 0, 1, ...$

1. 通过公式2.12 计算每一个维度的梯度

   $for \quad i = 0, 1, ..., d$

$$\nabla Err(w_{t,i}) = \sum_{n=1}^{N}\left(\frac{1}{1+e^{-\boldsymbol{w}_t^T\boldsymbol{x}_n}} - y_n\right)(x_{n,i})$$

2. 通过公式2.13迭代更新权重的每一个维度

   $for \quad i = 0, 1, ..., d$

$$w_{t+1,i} = w_{t,i} - \eta\nabla Err(w_{t,i})$$

直到$\nabla Err(\boldsymbol{w}) = 0$或者迭代足够多次

由于我们模型的目标函数为一个光滑的凸函数，故我们有很多可以优化数值计算的方法，保证全局最优解，常用的改进方式可以为牛顿法，拟牛顿法，共轭梯度等。除此之外由于每一次更新都需要重新计算整个训练集的梯度，如果是大数据这种方式就很慢了，相应的办法为随机梯度下降（SGD），上述提到的改进就不在这里详细展开说明了。

故我们最终可以得到特征权重$w_{t+1}$，它是基于已有数据集产生的加权参数。接下来只需要用这个参数去进行预测分类。对于一个测试样本$\boldsymbol{x}_{test}$，计算它属于类别1（positive）的概率$\frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}_{test}}}$。如果该值大于0.5即为类别1（positive），否则就是类别2（positive）。