梯度下降法、正则化与逻辑回归

1.梯度下降法

在介绍梯度下降法之前，先介绍下泰勒公式，泰勒公式的基本形式如下：

$$f(x)=f(x_0)+f^{'}(x_0)(x-x_0)+\frac{1}{2}f^{''}(x_0)(x-x_0)^2+\cdots$$

令 $x=\boldsymbol{w}^{t+1}$，$\boldsymbol{w}^{t+1}$ 代表第 t+1 次参数向量的值；令 $x_0=\boldsymbol{w}^t$，代表第 t 次参数向量的值；其中 $\boldsymbol{w}$ 共有 k 个参数，$\boldsymbol{w}=[w_1,w_2,\dots,w_k]$；令 x-x_0=△$\boldsymbol{w}$，取一阶泰勒公式，则：

$$f(\boldsymbol{w}^{t+1})\approx f(\boldsymbol{w}^t)+f^{'}(\boldsymbol{w}^t)\cdot\Delta\boldsymbol{w}$$

由于是梯度下降，所以 $f(\boldsymbol{w}^{t+1})\le f(\boldsymbol{w}^t)$，所以

$$\Delta\boldsymbol{w}=-\alpha\cdot f^{'}(\boldsymbol{w}^t)$$

令函数 f 为损失函数 J，则

$$\boldsymbol{w}^{t+1}=\boldsymbol{w}^t-\alpha\cdot J^{'}(\boldsymbol{w}^t)$$

故第 t+1 次参数向量的值等于第 t 次参数向量的值减去损失函数偏导乘以学习率 α。

2.正则化

为了防止过拟合，一般采用正则化，正则化一般分为 L1 正则化和 L2 正则化，分别为：

$$J_1(\boldsymbol{w})=J(\boldsymbol{w})+\lambda\sum_{i=1}^{k}|w_i|$$

$$J_2(\boldsymbol{w})=J(\boldsymbol{w})+\frac{\lambda}{2}\sum_{i=1}^{k}w_i^2$$

分别对 w_i 求偏导，得

$$w_i^{t+1}=w_i^t-\alpha\frac{\partial J(\boldsymbol{w})}{\partial w_i^t}-\lambda\alpha\,\mathrm{sgn}(w_i^t)$$

$$w_i^{t+1}=w_i^t-\alpha\frac{\partial J(\boldsymbol{w})}{\partial w_i^t}-\lambda\alpha w_i^t$$

最后，

L1 正则化： $$w_i^{t+1}=w_i^t-\alpha\left(\frac{\partial J(\boldsymbol{w})}{\partial w_i^t}+\lambda\,\mathrm{sgn}(w_i^t)\right)$$

L2 正则化： $$w_i^{t+1}=(1-\lambda\alpha)w_i^t-\alpha\frac{\partial J(\boldsymbol{w})}{\partial w_i^t}$$

从以上公式可以发现两者都可以防止过拟合，但是有时候我们能听到 L1 正则化相对于 L2 正则化更容易产生数据稀疏性，这是为什么呢？下面分两种情况讨论：

(1)令 $w_i^t>0$，假设 L1 正则化比 L2 正则化更容易产生数据稀疏性等价于 L1 正则化比 L2 正则化更接近于 0，等价于：

$$
\begin{aligned}
& w_i^t - \alpha(\frac{\partial J(\boldsymbol{w})}{\partial w_i^t} + \lambda \operatorname{sgn}(w_i^t)) < (1-\lambda\alpha) w_i^t - \alpha \frac{\partial J(\boldsymbol{w})}{\partial w_i^t} \\
& \Leftrightarrow -\lambda\alpha \operatorname{sgn}(w_i^t) < -\lambda\alpha w_i^t \\
& \Leftrightarrow -\lambda\alpha < -\lambda\alpha w_i^t \\
& \Leftrightarrow w_i^t < 1
\end{aligned}
$$

(2)令 $w_i^t < 0$，假设 L1 正则化比 L2 正则化更容易产生数据稀疏性等价于 L1 正则化比 L2 正则化更接近于 0，等价于：

$$
\begin{aligned}
& w_i^t - \alpha(\frac{\partial J(\boldsymbol{w})}{\partial w_i^t} + \lambda \operatorname{sgn}(w_i^t)) > (1-\lambda\alpha) w_i^t - \alpha \frac{\partial J(\boldsymbol{w})}{\partial w_i^t} \\
& \Leftrightarrow -\lambda\alpha \operatorname{sgn}(w_i^t) > -\lambda\alpha w_i^t \\
& \Leftrightarrow \lambda\alpha > -\lambda\alpha w_i^t \\
& \Leftrightarrow w_i^t > -1
\end{aligned}
$$

通过(1)与(2)两种情况可以发现，当 $\left|w_i^t\right| < 1$ 时，L1 正则化比 L2 正则化更容易产生数据稀疏性；当 $\left|w_i^t\right| = 1$ 时，L1 正则化与 L2 正则化产生数据稀疏性的程度相同；当 $\left|w_i^t\right| > 1$ 时，L2 正则化比 L1 正则化更容易产生数据稀疏性。由于在工程当中，我们一般考虑当 $\left|w_i^t\right| \ll 1$ 时，才希望更新的参数更有稀疏性，所以我们才经常听到 L1 正则化比 L2 正则化更容易产生数据稀疏性。

3.逻辑回归

逻辑回归是建立在线性回归的基础上，一般采用 sigmoid 函数来拟合，即

$$
h_w(\boldsymbol{x}) = \frac{1}{1+e^{-\left(\boldsymbol{w}^T \boldsymbol{x} + b\right)}}
$$

其中，$\boldsymbol{w}^T\boldsymbol{x}=w_1x_1+w_2x_2+\ldots+w_nx_n$，$b$ 代表偏置，是常数，$\boldsymbol{x}$ 为样本特征，$\boldsymbol{w}$ 为样本对应的系数，在已知样本特征 $\boldsymbol{x}$ 和最终分类结果 y（1 或者 0）的前提下，求系数 $\boldsymbol{w}$ 使得损失函数最小。

假设有 m 个样本，则相应的极大似然函数为

$$
L(\boldsymbol{w}) = \prod_{i=1}^{m} h_{\boldsymbol{w}}(\boldsymbol{x}_i)^{y_i} (1 - h_{\boldsymbol{w}}(\boldsymbol{x}_i))^{1-y_i}
$$

两边取对数化简得损失函数 $J(\boldsymbol{w})$，求使损失函数最小的参数：

$$
J(\boldsymbol{w}) = -\frac{1}{m} \sum_{i=1}^{m} [y_i \ln h_{\boldsymbol{w}}(\boldsymbol{x}_i) + (1-y_i) \ln(1 - h_{\boldsymbol{w}}(\boldsymbol{x}_i))]
$$

经化简：

$$\frac{\partial J(\boldsymbol{w})}{\partial w_j}=-\frac{1}{m}\sum_{i=1}^{m}[y_i\frac{1}{h_w(\boldsymbol{x}_i)}h_w(\boldsymbol{x}_i)(1-h_w(\boldsymbol{x}_i))x_{ij}+(1-y_i)\frac{-1}{1-h_w(\boldsymbol{x}_i)}h_w(\boldsymbol{x}_i)(1-h_w(\boldsymbol{x}_i))x_{ij}]$$

$$=-\frac{1}{m}\sum_{i=1}^{m}[y_i(1-h_w(\boldsymbol{x}_i))x_{ij}-(1-y_i)h_w(\boldsymbol{x}_i)x_{ij}]$$

$$=-\frac{1}{m}\sum_{i=1}^{m}(y_i x_{ij}-h_w(\boldsymbol{x}_i)x_{ij})$$

$$=-\frac{1}{m}\sum_{i=1}^{m}(y_i-h_w(\boldsymbol{x}_i))x_{ij}$$

$$\frac{\partial J(\boldsymbol{w})}{\partial b}=-\frac{1}{m}\sum_{i=1}^{m}[y_i\frac{1}{h_w(\boldsymbol{x}_i)}h_w(\boldsymbol{x}_i)(1-h_w(\boldsymbol{x}_i))+(1-y_i)\frac{-1}{1-h_w(\boldsymbol{x}_i)}h_w(\boldsymbol{x}_i)(1-h_w(\boldsymbol{x}_i))]$$

$$=-\frac{1}{m}\sum_{i=1}^{m}[y_i(1-h_w(\boldsymbol{x}_i))-(1-y_i)h_w(\boldsymbol{x}_i)]$$

$$=-\frac{1}{m}\sum_{i=1}^{m}(y_i-h_w(\boldsymbol{x}_i))$$

其中，x_{ij} 是第 i 个样本 $\boldsymbol{x}_i$ 的第 j 个特征，故

$$w_j=w_j+\alpha\sum_{i=1}^{m}(y_i-h_w(\boldsymbol{x}_i))x_{ij}$$

$$b=b+\alpha\sum_{i=1}^{m}(y_i-h_w(\boldsymbol{x}_i))$$

如果 m 是全量样本，则为批量梯度下降法(BGD)，如果 m 是部分样本，则为小批量梯度下降法(MBGD)，如果 m 是一个样本（每次迭代从所有样本中随机选择一个样本代替所有样本），则为随机梯度下降法(SGD)。所以，逻辑回归的 m 个样本对第 j 个特征和参数 b 的梯度分别为：

$$g_j=\sum_{i=1}^{m}(h_w(\boldsymbol{x}_i)-y_i)x_{ij}$$

$$g_b=\sum_{i=1}^{m}(h_w(\boldsymbol{x}_i)-y_i)$$

如果是一个样本，则

$$g_j=(h_w(\boldsymbol{x}_i)-y_i)x_{ij}$$

$$g_b=h_w(\boldsymbol{x}_i)-y_i$$

注：

1.sigmoid 函数： $f(x)=\frac{1}{1+e^{-x}}$ 有如下性质：

(1). $f^{'}(x)=f(x)[1-f(x)]$

(2). $f(-x)=1-f(x)$

2.指数损失函数：

$J(y,h(x))=e^{-y\cdot h(x)}$

3.tanh 函数： $f(x)=\frac{e^x-e^{-x}}{e^x+e^{-x}}$ 有如下性质：

$f^{'}(x)=1-f^2(x)$

4.交叉熵与 logit loss 的关系：

2、Logistic loss

$$L(y,f(x))=log(1+e^{-yf(x)})$$

logistic Loss为Logistic Regression中使用的损失函数，下面做一下简单证明：

Logistic Regression中使用了Sigmoid函数表示预测概率：

$$g(f(x))=P(y=1|x)=\frac{1}{1+e^{-f(x)}}$$

而 $P(y=-1|x)=1-P(y=1|x)=1-\frac{1}{1+e^{-f(x)}}=\frac{1}{1+e^{f(x)}}=g(-f(x))$

因此利用 $y\in\{-1,+1\}$ ，可写为 $P(y|x)=\frac{1}{1+e^{-yf(x)}}$ ，此为一个概率模型，利用极大似然的思想：

$$max\left(\prod_{i=1}^{m}P(y_i|x_i)\right)=max\left(\prod_{i=1}^{m}\frac{1}{1+e^{-y_if(x_i)}}\right)$$

两边取对数，又因为是求损失函数，则将极大转为极小：

$$max\left(\sum_{i=1}^{m}logP(y_i|x_i)\right)=-min\left(\sum_{i=1}^{m}log(\frac{1}{1+e^{-y_if(x_i)}})\right)=min\left(\sum_{i=1}^{m}log(1+e^{-y_if(x_i)}\right)$$

这样就得到了logistic loss。

如果定义 $t=\frac{y+1}{2}\in\{0,1\}$ ，则极大似然法可写为：

$$\prod_{i=1}^{m}(P(t_i=1|x_i))^{t_i}((1-P(t_i=1|x))^{1-t_i}$$

取对数并转为极小得：

$$\sum_{i=1}^{m}\{-t_i\log P(t_i=1|x_i)-(1-t_i)\log(1-P(t_i=1|x_i))\}$$

上式被称为交叉熵损失 (cross entropy loss)，可以看到在二分类问题中logistic loss和交叉熵损失是等价的，二者区别只是标签y的定义不同。