https://github.com/StevenLzq

一、逻辑回归的一些基本理解

1、关于逻辑回归和对数几率回归：
1）线性回归是做回归预测，逻辑回归是做分类预测；
2）线性回归使用平方损失函数，逻辑回归则是最大似然函数

2、为什么用sigmoid函数？
优点：
1.数据压缩能力，将数据规约在[0，1]之间
2.导数形式优秀，方便计算
缺点：
1.容易梯度消失，x稍大的情况下就趋近一条水平线
2.非0中心化，在神经网络算法等情况下，造成反向传播时权重的全正全负的情况。

3、如果label={-1, +1}，给出LR的损失函数？

设label={-1,+1},则

$$p(y=1|x)=h_\omega(x) \tag{1}$$
$$p(y=-1|x)=1-h_\omega(x) \tag{2}$$

对于sigmoid函数，有以下特性，

$$h(-x)=1-h(x)$$

所以（1）（2）式子可表示为

$$p(y|x)=h_\omega(yx)$$

同样，我们使用MLE作估计，

$$L(\omega)=\prod_{i=1}^{m}p(y_i|x_i;\omega) \qquad =\prod_{i=1}^{m}h_\omega(y_ix_i) \ =\prod_{i=1}^{m}\frac{1}{1+e^{-y_iwx_i}}$$

对上式取对数及负值，得到损失为：

$$-\log L(\omega)=-\log\prod_{i=1}^{m}p(y_i|x_i;\omega) \qquad =-\sum_{i=1}^{m}\log p(y_i|x_i;\omega) \ =-\sum_{i=1}^{m}\log\frac{1}{1+e^{-y_iwx_i}} \qquad =\sum_{i=1}^{m}\log(1+e^{-y_iwx_i})$$

即对于每一个样本，损失函数为：

$$L(\omega)=\log(1+e^{-y_iwx_i})$$

4、关于逻辑回归特征离散化
离散特征的增加和减少都很容易，易于模型的快速迭代；稀疏向量内积乘法运算速度快，计算结果方便存储容易扩展；离散化后的特征对异常数据有很强的鲁棒性。

5、当用LR时的考虑，特征中的某些值很大，是否意味着这个特征重要程度高？
不对，在线性模型中（特征归一化之后保持同一量纲的情况下）我们认为特征对应的参数值越大，其特征重要性越高。我们最终关注的参数值大小。

二、从几率角度理解

几率代表的知识正例和反例的比值；
我们以y代表正例的概率那么1-y即代表反例的概率

$$y=\frac{1}{1+e^{w^Tx+b}} \tag{1}$$

我们对式子1进行形式转化

$$\frac{1}{y}=1+e^{w^Tx+b} \tag{2}$$

将1移到左边后

$$\frac{1-y}{y} = e^{w^Tx+b} \tag{3}$$

两侧同取对数

$$ln\frac{1-y}{y} = -(w^Tx+b) \tag{4}$$

两侧同乘-1,即可得到对数几率

$$ln\frac{y}{1-y} = w^Tx+b \tag{5}$$

为了确定w和b，从后验概率（条件概率）来估计$p(y=1|x)$

关于先验、似然和后验，引用网络上一个类比的说法

1）先验——根据若干年的统计（经验）或者气候（常识），某地方下雨的概率；

2）似然——下雨（果）的时候有乌云（因/证据/观察的数据）的概率，即已经有了果，对证据发生的可能性描述；

3）后验——根据天上有乌云（原因或者证据/观察数据），下雨（结果）的概率；

后验 ~ 先验*似然 ： 存在下雨的可能（先验），下雨之前会有乌云（似然）~ 通过现在有乌云推断下雨概率（后验）；

在此基础上我们对式子（5）对数几率函数进行更新

$$ln\frac{p(y=1|x)}{p(y=0|x)} = w^Tx+b \tag{6}$$

那么有

$$\frac{p(y=1|x)}{p(y=0|x)} = e^{w^T+b} \tag{7}$$

$$p(y=1|x)+p(y=0|x) = 1 \tag{8}$$

根据式子7和8可得到

$$\frac{p(y=1|x)}{1-p(y=1|x)} = e^{w^T+b} \tag{9}$$

继而求得

$$p(y=1|x) = \frac{e^{w^Tx+b}}{1+e^{w^Tx+b}} \tag{10}$$

$$p(y=0|x) = \frac{1}{1+e^{w^Tx+b}} \tag{11}$$

接下来以极大似然法来估计w和b，对于给定的数据集$(x_i,y_i)_{i=1}^m$,对率回归模型最大化对数似然

$$l(w,b) = \sum_{i=1}^m ln(p(y_i|x_i;w,b)) \tag{12}$$

就是说每个样本属于其真实标记概率越大越好

令$\beta=(w;b),\hat{x}=(x;1)$

令$p_1(\hat{x};\beta)=p(y=1|\hat{x};\beta),p_0(\hat{x};\beta)=1-p_1(\hat{x};\beta)$

那么有

$$p(y_i|x_i;w,b) = y_ip_1(\hat{x_i};\beta)+(1-y_i)p_0(\hat{x_i};\beta) \tag{13}$$

我们把13代入12

$$l(\beta) = \sum_{i=1}^m ln(\frac{y_ie^{\beta^T\hat{x_i}}+1-y_i}{1+e^{\beta^T\hat{x_i}}}) = \sum_{i=1}^m (ln(y_ie^{\beta^T\hat{x_i}}+1-y_i)-ln(1+e^{\beta^T\hat{x_i}})) \tag{14}$$

再结合式子10和11可得到

$$l(\beta) = \{ \sum_{i=1}^m (-ln(1+e^{\beta^T\hat{x_i}})), y_i=0 \quad \sum_{i=1}^m (\beta^T\hat{x_i}-ln(1+e^{\beta^T\hat{x_i}})), y_i=1 \tag{15}$$

综合15中的两种情况

$$l(\beta) = \sum_{i=1}^m (y_i\beta^T\hat{x_i}-ln(1+e^{\beta^T\hat{x_i}})) \tag{16}$$

而在机器学习中一般不求最大，更偏爱最小因此等价于

$$l(\beta) = \sum_{i=1}^m (-y_i\beta^T\hat{x_i}+ln(1+e^{\beta^T\hat{x_i}})) \tag{17}$$

在此基础上就可以依据梯度下降、牛顿法和拟牛顿法等方式进行求最优解，

这一部分内容会单独另外作讲解说明。

三、另外一个角度理解

以上部分是按照对数几率的思路来理解逻辑回归的似然函数，下面部分内容从另外一个角度来尝试

设

$$p(Y=1|x)=F(x), p(Y=0|x)=1-F(x) \tag{1}$$

那么似然函数为

$$\prod_{i=1}^m ([f(x_i)]^{y_i}[1-f(x_i)]^{1-y_i}) \tag{2}$$

两侧同取对数把连乘换成加法，进而变成对数似然函数

$$
\begin{aligned}
l（w）&=\sum_{i=1}^{m}[y_i ln(f(x_i))+(1-y_i)ln(1-f(x_i))] \\
&=\sum_{i=1}^{m}[y_i ln(f(x_i))+ln(1-f(x_i)-y_i ln(1-f(x_i))] \\
&=\sum_{i=1}^{m}[y_i ln(\frac{f(x_i)}{1-f(x_i)})+ln(1-f(x_i)] \\
&=\sum_{i=1}^{m}[y_i(wx_i)+ln(\frac{1}{1+e^{wx_i}})]\text{，（此处参考对数几率思路式子9）} \\
&=\sum_{i=1}^{m}[y_i(wx_i)-ln(1+e^{wx_i})]
\end{aligned}
$$

堪培拉的小熊
湖北 武汉
扫一扫上面的二维码图案，加我微信