Math Basis of Logistic Regression

周陆延

1. 给出利用梯度下降法求解逻辑回归的前反向公式推导，在这一部分中，你只需要考虑最朴素的二分类逻辑回归模型

   - 前向传播

     线性回归

     $$f(x)=\boldsymbol{\omega}^{\top}x+b \tag{1}$$

     令 $x'=(1,x)$ 则 $f(x)$ 可统一为

     $$f(x)=\boldsymbol{\omega}'^{\top}x \tag{2}$$

     为处理二分类问题将线性回归 $(-\infty,+\infty)$ 的值域转换为 $(0,1)$ 对概率进行预测，嵌套 $\sigma(x)=\frac{1}{1+e^{-x}}$ 函数

     即逻辑回归的前向传播公式

     $$h_{\boldsymbol{\omega}}(x)=\frac{1}{1+e^{-\boldsymbol{\omega}^{\top}x}} \tag{3}$$

   - 反向传播

     根据逻辑回归的极大似然估计设置交叉熵损失函数

     $$J(\boldsymbol{\omega})=-\frac{1}{m}\sum_{i=1}^{m}y\log\hat{y}+(1-y)\log(1-\hat{y}) \tag{4}$$

     求梯度得

     $$\begin{aligned}\frac{\delta J(\boldsymbol{\omega})}{\delta\boldsymbol{\omega}}&=-\frac{1}{m}\sum_{i=1}^{m}y\frac{1}{\hat{y}}\hat{y}'+(1-y)\frac{1}{1-\hat{y}}\hat{y}'\\&=-\frac{1}{m}\sum_{i=1}^{m}\left(\frac{y}{\hat{y}}+\frac{1-y}{1-\hat{y}}\right)\hat{y}'\\&=-\frac{1}{m}\sum_{i=1}^{m}\left(\sigma\left(\boldsymbol{\omega}^{\top}x^{(i)}\right)-y^{(i)}\right)x^{(i)}\end{aligned} \tag{5}$$

     使用梯度下降法更新 $\boldsymbol{\omega}:=\boldsymbol{\omega}-\alpha\frac{\delta J(\boldsymbol{\omega})}{\delta\boldsymbol{\omega}}$ 其中 α 为学习率

2. 假设标签集不是 $\{0,1\}$ 而是 $\{1,-1\}$，将会有什么变化，请给出推导

   似然函数变为

   $$L(\boldsymbol{\omega})=\sqrt{\Pr(x)^{1+y}(1-\Pr(x))^{1-y}} \tag{6}$$

   取负对数得到损失函数

   $$J(\boldsymbol{\omega})=-\frac{1}{2}((1+y)\log\hat{y}+(1-y)\log(1-\hat{y})) \tag{7}$$

   对 $y=1,y=-1$ 分类讨论得

$$J(\boldsymbol{\omega}) = \log\left(1 + e^{y*\boldsymbol{\omega}^\top x}\right) \tag{8}$$

3. 在问题2.1的基础上，即标签集为 {0,1} 的情况，分别增加L1正则化和L2正则化，公式和模型效果分别会有什么变化，请给出推导

   最基本的正则化方法是在原目标（代价）函数中添加惩罚项，对复杂度高的模型进行“惩罚”，即

$$\tilde{J}(x) = J(x) + \lambda\Omega(\boldsymbol{\omega}) \tag{9}$$

   正则化可理解为对原损失函数最优化过程添加约束

$$\begin{aligned} &\min_{\boldsymbol{\omega}} J(\boldsymbol{\omega}) \\ &s.t. \Omega(\boldsymbol{\omega}) \leq C \end{aligned} \tag{10}$$

   利用拉格朗日算子法，我们可将上述带约束条件的最优化问题转换为不带约束项的优化问题，构造拉格朗日函数

$$L(\boldsymbol{\omega}, \lambda) = J(\boldsymbol{\omega}) + \lambda(\Omega(\boldsymbol{\omega}) - C) \tag{11}$$

   设 λ 最优解为 λ^* 则对拉格朗日函数最小化等价于

$$\min_{\boldsymbol{\omega}} J(\boldsymbol{\omega}) + \lambda^*\Omega(\boldsymbol{\omega}) \tag{12}$$

   - L2 正则化

     即使用L2范数作为惩罚

$$J(\boldsymbol{\omega}) = -\frac{1}{m}\sum_{i=1}^{m} y\log\hat{y} + (1-y)\log(1-\hat{y}) + \frac{\lambda}{2m}\|\boldsymbol{\omega}\|_2^2 \tag{13}$$

     梯度则变为

$$\frac{\delta J(\boldsymbol{\omega})}{\delta \boldsymbol{\omega}} = -\frac{1}{m}\left(\sum_{i=1}^{m}\left(\sigma\left(\boldsymbol{\omega}^\top x^{(i)}\right) - y^{(i)}\right)x^{(i)}\right) + \frac{\lambda}{m}\boldsymbol{\omega} \tag{14}$$

     考虑对模型的影响，令 $\boldsymbol{\omega}^*$ 为未正则化的目标函数的最优解，对 $J(\boldsymbol{\omega})$ 作二阶泰勒展开近似（$\boldsymbol{\omega}^*$ 为最优，无一阶导项；略去样本数量 m）

$$\hat{J}(\boldsymbol{\omega}) = J(\boldsymbol{\omega}^*) + \frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\omega}^*)^\top H(\boldsymbol{\omega} - \boldsymbol{\omega}^*) \tag{15}$$

     当 $\hat{J}(\boldsymbol{\omega})$ 最小时，其梯度为

$$\frac{\delta \hat{J}(\boldsymbol{\omega})}{\delta \boldsymbol{\omega}} = H(\boldsymbol{\omega} - \boldsymbol{\omega}^*) = 0 \tag{16}$$

     加入惩罚项，记此时的最优解为 $\boldsymbol{\omega}'$ 得

$$\begin{aligned} &H(\boldsymbol{\omega}' - \boldsymbol{\omega}^*) + \lambda\boldsymbol{\omega}' = 0 \\ &\boldsymbol{\omega}' = (H + \lambda I)^{-1}H\boldsymbol{\omega}^* \end{aligned} \tag{17}$$

     由 H 实对称，将其合同到对角矩阵 $H = Q\Lambda Q^\top$ 带入上式得

$$
\begin{aligned}
\boldsymbol{\omega}' &= (Q\Lambda Q^\top + \lambda I)^{-1} Q\Lambda Q^\top \boldsymbol{\omega}^* \\
&= (Q\Lambda Q^\top + Q(\lambda I)Q^\top)^{-1} Q\Lambda Q^\top \boldsymbol{\omega}^* \\
&= (\Lambda + \lambda I)^{-1} \Lambda \boldsymbol{\omega}^*
\end{aligned} \tag{18}
$$

发现 $\boldsymbol{\omega}'$ 相比 $\boldsymbol{\omega}^*$ 是依据 H 的特征值在对应分量上做了缩放，在 H 特征值较大的方向影响较小，在特征值较小的方向影响较大，使对减少目标函数作用显著的参数被保留，作用微弱的参数被衰减

- L1 正则化

  即使用L1范数作为惩罚

$$
J(\boldsymbol{\omega}) = -\frac{1}{m}\left(\sum_{i=1}^{m} y^{(i)} \log \hat{y}^{(i)} + \big(1 - y^{(i)}\big)\log\big(1 - \hat{y}^{(i)}\big)\right) + \frac{\lambda}{m}\|\boldsymbol{\omega}\|_1 \tag{19}
$$

  梯度则变为

$$
\frac{\delta J(\boldsymbol{\omega})}{\delta \boldsymbol{\omega}} = -\frac{1}{m}\left(\sum_{i=1}^{m} \big(\sigma\big(\boldsymbol{\omega}^\top x^{(i)}\big) - y^{(i)}\big) x^{(i)}\right) + \frac{\lambda}{m}\mathrm{sgn}(w) \tag{20}
$$

  与L2正则化类比，但是由于 $\mathrm{sgn}(x)$ 的特殊性，进一步假设 H 为对角阵（对数据进行主成分分析后成立），则

$$
\begin{aligned}
\hat{\tilde{J}}(\boldsymbol{\omega}) &= J(\boldsymbol{\omega}^*) + \frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\omega}^*)^\top H (\boldsymbol{\omega} - \boldsymbol{\omega}^*) + \lambda\|\boldsymbol{\omega}\|_1 \\
&= J(\boldsymbol{\omega}^*) + \sum_i \left(\frac{1}{2} H_{i,i}(\boldsymbol{\omega}_i - \boldsymbol{\omega}_i^*)^2 + \lambda|\boldsymbol{\omega}_i|\right)
\end{aligned} \tag{21}
$$

  令

$$
\hat{\tilde{J}}'(\boldsymbol{\omega}) = \sum_i H_{i,i}(\boldsymbol{\omega}_i - \boldsymbol{\omega}_i^*) + \lambda\mathrm{sgn}(\boldsymbol{\omega}_i) = 0 \tag{22}
$$

  得

$$
\boldsymbol{\omega}_i = \mathrm{sgn}(\boldsymbol{\omega}_i^*)\max\left\{|\boldsymbol{\omega}^*| - \frac{\lambda}{H_{i,i}}, 0\right\} \tag{23}
$$

  可见当 $|\boldsymbol{\omega}^*| \le \frac{\lambda}{H_{i,i}}$ 时会使得 $\boldsymbol{\omega}_i$ 变成 0，使得参数稀疏化

4. 给出核逻辑回归的对偶形式

   为书写方便给出标签集为 $\{+1, -1\}$ 的情况，其他情况可将标签集 $\{a, b\}$ 进行映射 $y_i \to \frac{2}{b-a}\big(y_i - \frac{b-a}{2}\big)$

$$
\begin{aligned}
&\min_{\boldsymbol{\alpha}} \frac{1}{2\lambda}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j \kappa\big(x_i, x_j\big) + \sum_{i=1}^{m}[\alpha_i \log\alpha_i + (1-\alpha_i)\log(1-\alpha_i)] \\
&\text{s.t. } 0 \le \alpha_i < 1
\end{aligned} \tag{24}
$$