逻辑回归

2021-09-05

1. 广义线性模型

1.1. 指数族分布

指数族（Exponential Family）分布是一类分布的总称，该类分布的分布律（或概率密度函数）的一般形式为：

$$p(y;\eta)=b(y)\cdot e^{\left(\eta^T T(y)-a(\eta)\right)} \tag{21}$$

各符号含义如下：

1. η为该分布的自然参数，可为向量
2. $T(y)$为充分统计量，视具体的分布而定，通常等于随机变量y本身
3. $a(\eta)$为配分函数
4. $b(y)$为关于随机变量y的函数

常见的伯努利分布和正态分布均属于指数族分布。以下证明伯努利分布属于指数族分布：

已知伯努利分布的分布律为：

$$\begin{gathered}p(y)=\phi^y(1-\phi)^{1-y}\\ \text{其中，}y\in\{0,1\}\text{，}\phi\text{为}y=1\text{的概率，即}p(y=1)=\phi\end{gathered} \tag{22}$$

对上式恒等变形得：

$$\begin{aligned}p(y)&=\phi^y(1-\phi)^{1-y}\\ &=\exp\left(\ln\left(\phi^y(1-\phi)^{1-y}\right)\right)\\ &=\exp\left(\ln\phi^y+\ln(1-\phi)^{1-y}\right)\\ &=\exp\left(y\ln\phi+(1-y)\ln(1-\phi)\right)\end{aligned} \tag{23}$$

$$
\begin{aligned}
&= \exp\left(y \ln \phi + (1-y) \ln (1-\phi)\right) \\
&= \exp\left(y \ln \phi + \ln (1-\phi) - y \ln (1-\phi)\right) \\
&= \exp\left(y (\ln \phi - \ln (1-\phi)) + \ln (1-\phi)\right) \\
&= \exp\left(y \ln \left(\frac{\phi}{1-\phi}\right) + \ln (1-\phi)\right)
\end{aligned} \tag{23}
$$

对比指数族分布的一般形式 $p(y;\eta) = b(y) \cdot e^{\left(\eta^T T(y) - a(\eta)\right)}$，可知：

$$
\begin{aligned}
b(y) &= 1 \\
\eta &= \ln \left(\frac{\phi}{1-\phi}\right) \\
T(y) &= y \\
a(\eta) &= -\ln (1-\phi) = \ln (1+e^{\eta})
\end{aligned} \tag{24}
$$

由此说明，伯努利分布为指数族分布。

1.2. 广义线性模型的假设

1. 在给定x的条件下，假设随机变量y服从某个指数族分布
2. 在给定x的条件下，我们的目标是得到一个模型$h(x)$能预测出$T(y)$的期望值
3. 假设该指数族分布中的自然参数η和x呈线性关系，即$\eta = w^T x$

2. 逻辑回归

2.1. 模型推导

逻辑回归是对二分类问题进行建模，并且假设被建模的随机变量y取值为0或1，因此，可以很自然得**假设y服从伯努利分布**。此时，如果希望构建一个线性模型来预测给定x的条件下y取值的话，可以考虑使用广义线性模型来进行建模。

已知y服从伯努利分布，而伯努利分布属于指数族分布，所以满足广义线性模型的三条假设。根据第二条假设，可以推出模型$h(x)$的表达式为：

$$
h(x) = E[T(y \mid x)] \tag{25}
$$

注意，$y \mid x$只是表示形式，并不影响$T(y)$的计算，即$T(y \mid x) = T(y)$。根据伯努利分布的$T(y \mid x) = y \mid x$，所以：

$$
h(x) = E(y \mid x) \tag{26}
$$

又因为$E[y \mid \boldsymbol{x}] = 1 \times p(y=1 \mid \boldsymbol{x}) + 0 \times p(y=0 \mid \boldsymbol{x}) = p(y=1 \mid \boldsymbol{x}) = \phi$，所以：

$$
h(x) = \phi \tag{27}
$$

根据公式(4) $\eta = \ln \left(\frac{\phi}{1-\phi}\right)$可知，对伯努利分布：

$$
\frac{1}{1+e^{-\eta}} = \phi \tag{28}
$$

将ϕ带入$h(x)$得：

$$
h(\boldsymbol{x}) = \phi = \frac{1}{1+e^{-\eta}} \tag{29}
$$

根据广义线性模型的第三条假设 $\eta = w^T x$，$h(x)$最终可化简为：

$$
h(\boldsymbol{x}) = \phi = \frac{1}{1+e^{-\boldsymbol{w}^T \boldsymbol{x}}} = p(y=1 \mid \boldsymbol{x}) \tag{30}
$$

此即为逻辑回归模型。

2.2. 极大似然估计

已知随机变量y取1和0的概率分别为（考虑偏置项）：

$$p(y=1 \mid \boldsymbol{x})=\frac{e^{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}+b}}{1+e^{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}+b}} p(y=0 \mid \boldsymbol{x})=\frac{1}{1+e^{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}+b}} \tag{31}$$

令$\boldsymbol{\beta}=(\boldsymbol{w} ; b), \hat{\boldsymbol{x}}=(\boldsymbol{x} ; 1)$，则$\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}+b$可简写为$\boldsymbol{\beta}^{\mathbf{T}} \hat{\boldsymbol{x}}$，于是上式可化简为：

$$p(y=1 \mid \boldsymbol{x})=\frac{e^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}}}{1+e^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}}}=p_1(\hat{\boldsymbol{x}} ; \boldsymbol{\beta}) p(y=0 \mid \boldsymbol{x})=\frac{1}{1+e^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}}}=p_0(\hat{\boldsymbol{x}} ; \boldsymbol{\beta}) \tag{32}$$

将上式合并得：

$$p(y \mid \boldsymbol{x} ; \boldsymbol{w}, b)=y \cdot p_1(\hat{\boldsymbol{x}} ; \boldsymbol{\beta})+(1-y) \cdot p_0(\hat{\boldsymbol{x}} ; \boldsymbol{\beta}) \tag{33}$$

或者：

$$p(y \mid \boldsymbol{x} ; \boldsymbol{w}, b)=\left[p_1(\hat{\boldsymbol{x}} ; \boldsymbol{\beta})\right]^y\left[p_0(\hat{\boldsymbol{x}} ; \boldsymbol{\beta})\right]^{1-y} \tag{34}$$

根据对数似然函数的定义可知：

$$\ln L(\boldsymbol{w})=\sum_{i=1}^m \ln f\left(y_i, w_1, w_2, \ldots, w_k\right) \tag{35}$$

因此，逻辑回归的对数似然函数可以表示为：

$$\ell(\boldsymbol{w}, b):=\ln L(\boldsymbol{w}, b)=\sum_{i=1}^m \ln p\left(y_i \mid \boldsymbol{x}_i ; \boldsymbol{w}, b\right) \tag{36}$$

根据$p(y \mid \boldsymbol{x} ; \boldsymbol{w}, b)$的两种形式（公式(13)和公式(14)），可以得到两种对数似然函数，以下将分别推导。

2.2.1. 似然函数形式一

将 $p(y \mid \boldsymbol{x} ; \boldsymbol{w}, b)=y \cdot p_1(\hat{\boldsymbol{x}} ; \boldsymbol{\beta})+(1-y) \cdot p_0(\hat{\boldsymbol{x}} ; \boldsymbol{\beta})$带入似然函数（公式(16)）可得：

$$\ell(\boldsymbol{\beta})=\sum_{i=1}^m \ln \left(y_i p_1\left(\hat{\boldsymbol{x}}_i ; \boldsymbol{\beta}\right)+\left(1-y_i\right) p_0\left(\hat{\boldsymbol{x}}_i ; \boldsymbol{\beta}\right)\right) \tag{37}$$

由于$p_1\left(\hat{\boldsymbol{x}}_i ; \boldsymbol{\beta}\right)=\frac{e^{\boldsymbol{\beta}^T \hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^T \hat{\boldsymbol{x}}_i}}$ ， $p_0\left(\hat{\boldsymbol{x}}_i ; \boldsymbol{\beta}\right)=\frac{1}{1+e^{\boldsymbol{\beta}^T \hat{\boldsymbol{x}}_i}}$ ，上式可化简为：

$$\begin{aligned}
\ell(\boldsymbol{\beta}) & =\sum_{i=1}^m \ln \left(\frac{y_i e^{\boldsymbol{\beta}^T \hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^T \hat{\boldsymbol{x}}_i}}+\frac{1-y_i}{1+e^{\boldsymbol{\beta}^T} \hat{\boldsymbol{x}}_i}\right) \\
& =\sum_{i=1}^m \ln \left(\frac{y_i e^{\boldsymbol{\beta}^T \hat{\boldsymbol{x}}_i}+1-y_i}{1+e^{\boldsymbol{\beta}^T \hat{\boldsymbol{x}}_i}}\right) \\
& =\sum_{i=1}^m\left(\ln \left(y_i e^{\boldsymbol{\beta}^T \hat{\boldsymbol{x}}_i}+1-y_i\right)-\ln \left(1+e^{\boldsymbol{\beta}^T \hat{\boldsymbol{x}}_i}\right)\right)
\end{aligned} \tag{38}$$

由于$y_i \in\{0,1\}$，所以

当 $y_i=0$时，
$$\begin{aligned}&\ln\left(y_i e^{\boldsymbol{\beta}^T\hat{\boldsymbol{x}}_i}+1-y_i\right)-\ln\left(1+e^{\boldsymbol{\beta}^T\hat{\boldsymbol{x}}_i}\right)\\&=ln\left(0\cdot e^{\beta^T\hat{x}_i}+1-0\right)-\ln\left(1+e^{\Theta^T\hat{x}_i}\right)\\&=\ln 1-\ln\left(1+e^{\beta^T}\hat{x}_i\right)\\&=-\ln\left(1+e^{\beta^T\hat{x}_i}\right)\end{aligned}$$

当$y_i=1$时，
$$\begin{aligned}&\ln\left(y_i e^{\boldsymbol{\beta}^T\hat{\boldsymbol{x}}_i}+1-y_i\right)-\ln\left(1+e^{\boldsymbol{\beta}^T\hat{\boldsymbol{x}}_i}\right)\\&=ln\left(1\cdot e^{\beta^T\hat{x}_i}+1-1\right)-\ln\left(1+e^{\Theta^T\hat{x}_i}\right)\\&=\beta^T\hat{x}_i-\ln\left(1+e^{\beta^T}\hat{x}_i\right)\end{aligned}$$

综合可得：

$$\ell(\boldsymbol{\beta})=\sum_{i=1}^{m}\left(y_i\boldsymbol{\beta}^T\hat{\boldsymbol{x}}_i-\ln\left(1+e^{\boldsymbol{\beta}^T\hat{\boldsymbol{x}}_i}\right)\right) \tag{39}$$

2.2.2. 似然函数形式二

若$p(y\mid\boldsymbol{x};\boldsymbol{w},b)=\left[p_1(\hat{\boldsymbol{x}};\boldsymbol{\beta})\right]^y\left[p_0(\hat{\boldsymbol{x}};\boldsymbol{\beta})\right]^{1-y}$，将其带入对数似然可得：

$$\begin{aligned}\ell(\boldsymbol{\beta})&=\sum_{i=1}^{m}\ln\left(\left[p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\right]^{y_i}\left[p_0\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\right]^{1-y_i}\right)\\&=\sum_{i=1}^{m}\left[\ln\left(\left[p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\right]^{y_i}\right)+\ln\left(\left[p_0\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\right]^{1-y_i}\right)\right]\\&=\sum_{i=1}^{m}\left[y_i\ln\left(p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\right)+(1-y_i)\ln\left(p_0\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\right)\right]\\&=\sum_{i=1}^{m}\left\{y_i\left[\ln\left(p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\right)-\ln\left(p_0\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\right)\right]+\ln\left(p_0\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\right)\right\}\\&=\sum_{i=1}^{m}\left[y_i\ln\left(\frac{p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)}{p_0\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)}\right)+\ln\left(p_0\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)\right)\right]\\&=\sum_{i=1}^{m}\left[y_i\ln\left(e^{\boldsymbol{\beta}^T\hat{\boldsymbol{x}}_i}\right)+\ln\left(\frac{1}{1+e^{\boldsymbol{\beta}^T\hat{\boldsymbol{x}}_i}}\right)\right]\\&=\sum_{i=1}^{m}\left(y_i\boldsymbol{\beta}^T\hat{\boldsymbol{x}}_i-\ln\left(1+e^{\boldsymbol{\beta}^T\hat{\boldsymbol{x}}_i}\right)\right)\end{aligned} \tag{40}$$

2.3. 模型求解

对似然函数$\ell(\boldsymbol{\beta})$求极大，或求损失函数$-\ell(\boldsymbol{\beta})$的极小。

3. 参考文献

1. Andrew Ng. cs229 -notes1