目 录

CONTENTS

交叉熵损失函数概览

交叉熵损失函数

二分类任务中，交叉熵损失函数定义如下：

$$L=-\frac{1}{N}\sum_{i=1}^{N}y^{(i)}\log\hat{y}^{(i)}+\left(1-y^{(i)}\right)\log\left(1-\hat{y}^{(i)}\right)$$

其中，N表示训练样本的数量；$y^{(i)}$表示标签(0或1)；$\hat{y}^{(i)}$表示模型预测属于该类的概率

多分类任务中，交叉熵损失函数定义如下：

$$L=-\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}y_{(i,k)}\log\hat{y}_{(i,k)}$$

其中，N表示训练样本的数量；K表示类别；$y_{(i,k)}$表示第i个样本的类别为k；$\hat{y}_{(i,k)}$表示模型预测第i个样本属于该类的概率

$$L=-\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}y_{(i,k)}\log\hat{y}_{(i,k)}$$

问题

- 交叉熵损失函数公式如何理解?
- 为什么交叉熵损失函数越小，训练得到的模型参数就越好?

两个方面理解

- KL散度与交叉熵
- 极大似然估计与最小化交叉熵

KL散度与交叉熵

信息量与熵

信息量

一条信息的信息量大小和它的不确定性有很大的关系。一句话如果需要很多外部信息才能确定，我们就称这句话的信息量比较大。比如你听到“云南西双版纳下雪了”，那你需要去看天气预报、问当地人等等查证（因为云南西双版纳从没下过雪）。相反，如果和你说“人一天要吃三顿饭”，那这条信息的信息量就很小，因为这条信息的确定性很高。

将事件 x_0 的信息量定义如下（其中 $p(x_0)$ 表示事件 x_0 发生的概率）：

$$I(x_0) = -\log(p(x_0))$$

熵

信息量是对于单个事件来说的，但是实际情况一件事有很多种发生的可能，比如掷骰子有可能出现6种情况，明天的天气可能晴、多云或者下雨等等。

熵是表示随机变量不确定的度量，是对所有可能发生的事件产生的信息量的期望。公式如下：

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log(p(x_i))$$

其中一种比较特殊的情况就是掷硬币，只有正、反两种情况，该种情况（二项分布或者0-1分布）熵的计算可以简化如下：

$$\begin{aligned} H(X) &= -\sum_{i=1}^{n} p(x_i)\log(p(x_i)) \\ &= -p(x)\log(p(x)) - (1-p(x))\log(1-p(x)) \end{aligned}$$

相对熵（KL散度）与交叉熵

相对熵（KL散度）

相对熵又称KL散度，用于衡量对于同一个随机变量x的两个分布p(x)和q(x)之间的差异，KL散度的值越小表示两个分布越接近。在机器学习中，**p(x)常用于描述样本的真实分布**，例如[1,0,0,0]表示样本属于第一类，**而q(x)则常常用于表示预测的分布**，例如[0.7,0.1,0.1,0.1]。q(x)需要不断地学习来拟合准确的分布p(x)

$$D_{KL}(p\|q)=\sum_{i=1}^{n} p\left(x_i\right) \log \left(\frac{p\left(x_i\right)}{q\left(x_i\right)}\right)$$

交叉熵

将KL散度的公式进行log除法变形，得到如下：

$$\begin{aligned} D_{KL}(p\|q) &= \sum_{i=1}^{n} p\left(x_i\right) \log \left(p\left(x_i\right)\right)-\sum_{i=1}^{n} p\left(x_i\right) \log \left(q\left(x_i\right)\right) \\ &= -H(p(x))+\left[-\sum_{i=1}^{n} p\left(x_i\right) \log \left(q\left(x_i\right)\right)\right] \end{aligned}$$

前半部分就是p(x)的熵，后半部分就是交叉熵：

$$H(p, q)=-\sum_{i=1}^{n} p\left(x_i\right) \log \left(q\left(x_i\right)\right)$$

由上可知，**KL散度的值越小 等价于 交叉熵越小，表示两个分布越接近**

极大似然估计与最小化交叉熵损失

二分类交叉熵损失函数

二分类的交叉熵损失函数

$$J(w) = -\frac{1}{N}\sum_{n=1}^{N}[y_n \log \hat{y_n} + (1-y_n)\log(1-\hat{y_n})]$$

去掉 1/N 并不影响函数的单调性，机器学习任务的也可以是最小化下面的交叉熵损失：

$$J(w) = -\sum_{n=1}^{N}[y_n \log \hat{y_n} + (1-y_n)\log(1-\hat{y_n})]$$

去掉负号，等价于最大化下面这个函数：

$$J(w) = \sum_{n=1}^{N}[y_n \log \hat{y_n} + (1-y_n)\log(1-\hat{y_n})]$$

上式就是**对伯努利分布求极大似然中的对数似然函数**(log-likelihood)

伯努利分布

伯努利分布，也可称为二项分布，0-1分布，是一个**离散型概率分布**

- 若伯努利试验成功，则伯努利随机变量取值为1；反之，取值为0
- 记其成功概率为p，失败的概率则为$1-p$，记作q
- 其概率质量函数：

$$f(x) = p^x(1-p)^{1-x} = \begin{cases} p & if\ x = 1, \\ q & if\ x = 0. \end{cases}$$

极大似然估计

极大似然估计，通俗理解来说，**就是利用已知的样本结果信息，反推最具有可能（最大概率）导致这些样本结果出现的模型参数值**

似然函数p(x|θ)的理解：

- 对于这个函数：p(x|θ)输入有两个：x表示某一个具体的数据；θ表示模型的参数

- 如果 θ 是已知确定的，x 是变量，这个函数叫做概率函数(probability function)，它描述对于不同的样本点 x，其出现概率是多少

- 如果 x 是已知确定的，θ 是变量，这个函数叫做似然函数(likelihood function)，它描述对于不同的模型参数，出现 x 这个样本点的概率是多少

求极大似然估计：**取似然函数，整理之后求最大值点**

伯努利分布的极大似然

有二元随机变量 $Y \in \{0,1\}$ (例如：抛硬币实验)，设 $p(Y=1)=\beta$ ，那么它的**概率质量函数**(Probability Mass Function，PMF)为：

$$P(Y \mid \beta)=\beta^{Y}(1-\beta)^{1-Y}$$

现有$D=\{y_1, y_2, \ldots y_n\}$是来自 Y 的、数量为N的一个样本集，元素是0或1，似然函数为：

$$P(D \mid \beta)=\prod_{i=1}^{N} P\left(Y=y_i \mid \beta\right)=\prod_{i=1}^{N} \beta^{y_i}(1-\beta)^{1-y_i}$$

在机器学习模型中，对上述 β 的定义为：

$$\beta=p_\theta\left(Y=1 \mid x_i\right)$$

其中，$X=\{x_1, x_2, \ldots x_n\}$ ， $x_i \in X$, X 是 D 中的**每个样本点对应类别的特征的集合。即给定模型参数 θ 和随机变量的样本点 $Y=1$ 的属性特征 x_i (x_i 可以是一个向量)，让模型估计出事件 $Y=1$ 的概率(同时也是当前伯努利分布的参数)。**

于是，将似然函数的参数 β 替换为θ，所得：

$$P(D \mid \theta, X)=\prod_{i=1}^{N} \beta^{y_i}(1-\beta)^{1-y_i}=\prod_{i=1}^{N} p_\theta\left(Y=1 \mid x_i\right)^{y_i}\left(1-p_\theta\left(Y=1 \mid x_i\right)\right)^{1-y_i}$$

易得**对数似然函数**(log-likelihood)：

$$\begin{aligned}
&\mathcal{L}(\theta ; X, D)=\log \prod_{i=1}^{N} p_\theta\left(Y=1 \mid x_i\right)^{y_i}\left(1-p_\theta\left(Y=1 \mid x_i\right)\right)^{1-y_i} \\
&=\sum_{i=1}^{N} \log p_\theta\left(Y=1 \mid x_i\right)^{y_i}\left(1-p_\theta\left(Y=1 \mid x_i\right)\right)^{1-y_i} \\
&=\sum_{i=1}^{N} \log p_\theta\left(Y=1 \mid x_i\right)^{y_i}+\log \left(1-p_\theta\left(Y=1 \mid x_i\right)\right)^{1-y_i} \\
&=\sum_{i=1}^{N} y_i \log p_\theta\left(Y=1 \mid x_i\right)+\left(1-y_i\right) \log \left(1-p_\theta\left(Y=1 \mid x_i\right)\right)
\end{aligned}$$

伯努利分布的极大似然

有二元随机变量 $Y \in \{0,1\}$ (例如：抛硬币实验)，设 $p(Y=1)=\beta$，那么它的**概率质量函数**(Probability Mass Function，PMF)为：

$$P(Y \mid \beta)=\beta^{Y}(1-\beta)^{1-Y}$$

现有$D=\{y_1, y_2, \ldots y_n\}$是来自 Y 的、数量为N的一个样本集，元素是0或1，似然函数为：

$$P(D \mid \beta)=\prod_{i=1}^{N} P\left(Y=y_i \mid \beta\right)=\prod_{i=1}^{N} \beta^{y_i}(1-\beta)^{1-y_i}$$

在机器学习模型中，对上述 β 的定义为：

$$\beta=p_\theta\left(Y=1 \mid x_i\right)$$

其中，$X=\{x_1, x_2, \ldots x_n\}$， $x_i \in X$, X 是 D 中的**每个样本点对应类别的特征的集合。即给定模型参数 θ 和随机变量的样本点 $Y=1$ 的属性特征 x_i (x_i 可以是一个向量)，让模型估计出事件 $Y=1$ 的概率(同时也是当前伯努利分布的参数)。**

于是，将似然函数的参数 β 替换为θ，所得：

$$P(D \mid \theta, X)=\prod_{i=1}^{N} \beta^{y_i}(1-\beta)^{1-y_i}=\prod_{i=1}^{N} p_\theta\left(Y=1 \mid x_i\right)^{y_i}\left(1-p_\theta\left(Y=1 \mid x_i\right)\right)^{1-y_i}$$

易得**对数似然函数**(log-likelihood)：

$$\begin{aligned}
\mathcal{L}(\theta ; X, D) &= \log \prod_{i=1}^{N} p_\theta\left(Y=1 \mid x_i\right)^{y_i}\left(1-p_\theta\left(Y=1 \mid x_i\right)\right)^{1-y_i} \\
&= \sum_{i=1}^{N} \log p_\theta\left(Y=1 \mid x_i\right)^{y_i}\left(1-p_\theta\left(Y=1 \mid x_i\right)\right)^{1-y_i} \\
&= \sum_{i=1}^{N} \log p_\theta\left(Y=1 \mid x_i\right)^{y_i}+\log \left(1-p_\theta\left(Y=1 \mid x_i\right)\right)^{1-y_i} \\
&= \sum_{i=1}^{N} y_i \log p_\theta\left(Y=1 \mid x_i\right)+\left(1-y_i\right) \log \left(1-p_\theta\left(Y=1 \mid x_i\right)\right)
\end{aligned}$$

极大似然估计与最小化交叉熵损失的转换

二分类的交叉熵损失函数

$$J(w)=-\frac{1}{N}\sum_{n=1}^{N}\left[y_n\log\hat{y_n}+(1-y_n)\log(1-\hat{y_n})\right]$$

伯努利分布求极大似然中的对数似然函数(log-likelihood)

$$\begin{aligned}
&\mathcal{L}(\theta;X,D)=\log\textstyle\prod_{i=1}^{N}p_\theta(Y=1\mid x_i)^{y_i}\left(1-p_\theta\left(Y=1\mid x_i\right)\right)^{1-y_i}\\
&=\textstyle\sum_{i=1}^{N}\log p_\theta(Y=1\mid x_i)^{y_i}\left(1-p_\theta\left(Y=1\mid x_i\right)\right)^{1-y_i}\\
&=\textstyle\sum_{i=1}^{N}\log p_\theta(Y=1\mid x_i)^{y_i}+\log\left(1-p_\theta\left(Y=1\mid x_i\right)\right)^{1-y_i}\\
&=\textstyle\sum_{i=1}^{N}y_i\log p_\theta\left(Y=1\mid x_i\right)+(1-y_i)\log(1-p_\theta\left(Y=1\mid x_i\right))
\end{aligned}$$

伯努利分布下，**极大似然估计与最小化交叉熵损失的转换**

$$\begin{aligned}
&\theta_p=\arg\max_\theta\textstyle\sum_{i=1}^{N}y_i\log p_\theta\left(Y=1\mid x_i\right)+(1-y_i)\log(1-p_\theta\left(Y=1\mid x_i\right))\\
&=\arg\max_\theta\textstyle\sum_{i=1}^{N}y_i\log\hat{y}_i+(1-y_i)\log(1-\hat{y}_i)\\
&=\arg\min_\theta-\textstyle\sum_{i=1}^{N}y_i\log\hat{y}_i+(1-y_i)\log(1-\hat{y}_i)\\
&=\arg\min_\theta\textstyle\sum_{i=1}^{N}H\left(y_i,\hat{y}_i\right)
\end{aligned}$$