

# 交叉熵详解 [1]

August 12, 2020

## Contents

1 交叉熵简介 [1]	1
2 信息量	2
3 信息熵	2
4 相对熵 (KL 散度)	2
5 交叉熵	3
6 机器学习中交叉熵的应用 [2]	3
6.1 为什么要用交叉熵做 loss 函数?	3
6.2 交叉熵在单分类问题中的使用	4
6.3 交叉熵在多分类问题中的使用	4
7 总结	4

## 1 交叉熵简介 [1]

在《1-预备知识》中，对信息熵进行了简单描述，并引用了交叉熵的概念，这里专门介绍下交叉熵的概念和原理。

交叉熵是信息论中的一个重要概念，主要用于度量两个概率分布间的差异性。**注意，交叉熵是用于比较两个概率差异性的指标，所以会广泛用于 RankNet 等排序算法中**

## 2 信息量

信息奠基人香农 (Shannon) 认为“信息是用来消除随机不确定性的东西”，也就是说衡量信息量的大小就是看这个信息消除不确定性的程度。信息量的大小与信息发生的概率成反比。**概率越大，信息量越小。概率越小，信息量越大。**

设某一事件发生的概率为  $P(x)$ ，其信息量  $I(x)$  表示为：

$$I(x) = -\log P(x)$$

这里  $\log$  表示以  $e$  为底的自然对数。

## 3 信息熵

信息熵也被称为熵，用来表示**所有**信息量的**期望**。

期望是试验中每次可能结果的概率乘以其结果的总和。

所以给定离散型随机变量  $X$ ，它的熵可表示为：

$$H(X) = -\sum_{i=1}^n P(x_i) \log P(x_i) \quad (X = x_1, x_2, \dots, x_n)$$

## 4 相对熵 (KL 散度)

如果对于同一个随机变量  $X$  有两个单独的概率分布  $P(x)$  和  $Q(x)$ ，则我们可以使用 KL 散度来衡量**这两个概率分布之间的差异**。KL 散度的定义为：

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

$n$  为事件的所有可能性。

KL 散度在信息论中有自己明确的物理意义，它是用来度量使用基于  $Q$  分布的编码来编码来自  $P$  分布的样本平均所需的额外的 Bit 个数。而其在机器学习领域的物理意义则是用来度量两个函数的相似程度或者相近程度。

例如，在机器学习中，常常使用  $P(x)$  来表示样本的真实分布， $Q(x)$  来表示模型所预测的分布。比如在一个三分类任务中，例如一张图片的真实分布  $P(X) = [1, 0, 0]$ （即图片属于第一类），预测分布  $Q(X) = [0.7, 0.2, 0.1]$ ，那么可以计算真实分布  $P(X)$  和预测分布  $Q(X)$  的 KL 散度为：

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

$$D_{KL}(p||q) = p(x_1) \log(\frac{p(x_1)}{q(x_1)}) + p(x_2) \log(\frac{p(x_2)}{q(x_2)}) + p(x_3) \log(\frac{p(x_3)}{q(x_3)}) = 1.0 \times \log(\frac{1}{0.7}) = 0.36$$

KL 散度越小，表示  $P(x)$  和  $Q(x)$  的分布越接近，可以通过反复训练  $Q(x)$  来使  $Q(x)$  的分布逼近  $P(x)$ 。

## 5 交叉熵

首先将 KL 散度公式拆开：

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

$$= \sum_{i=1}^n p(x_i) \log(p(x_i)) - \sum_{i=1}^n p(x_i) \log(q(x_i)) \quad (2)$$

$$= H(p(x)) + \left[ - \sum_{i=1}^n p(x_i) \log(q(x_i)) \right] \quad (3)$$

$$(4)$$

前者  $H(p(x))$  表示信息熵，后者即为**交叉熵**，即**KL 散度 = 信息熵 + 交叉熵**。

交叉熵公式表示为：

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

在机器学习训练网络时，输入数据与标签常常已经确定，那么真实概率分布  $P(x)$  也就确定下来了，所以信息熵在这里就是一个常量。由于 KL 散度的值表示真实概率分布  $P(x)$  与预测概率分布  $Q(x)$  之间的差异，值越小表示预测的结果越好，所以需要**最小化 KL 散度**，而交叉熵等于 KL 散度加上一个常量（信息熵），且公式相比 KL 散度更加容易计算，所以在**机器学习中常常使用交叉熵损失函数来计算 loss 就行了**。

## 6 机器学习中交叉熵的应用 [2]

### 6.1 为什么要用交叉熵做 loss 函数？

在线性回归问题中，常常使用 MSE（Mean Squared Error）作为 loss 函数，比如：

$$loss = \frac{1}{2m} \sum_{i=1}^m (\hat{y}_i - y_i)^2$$

这里的  $m$  表示  $m$  个样本的，loss 为  $m$  个样本的 loss 均值。

MSE 在线性回归问题中比较好用，那么在逻辑分类问题中还是如此么？

## 6.2 交叉熵在单分类问题中的使用

这里的单类别是指，每一张图像样本只能有一个类别，比如只能是狗或只能是猫。

交叉熵在单分类问题上基本是标配的方法：

$$loss = - \sum_{i=1}^n \hat{y}_i \log(y_i)$$

$n$  代表着  $n$  种类别。

## 6.3 交叉熵在多分类问题中的使用

这里的多类别是指，每一张图像样本可以有多个类别，比如同时包含一只猫和一只狗。和单分类问题的标签不同，多分类的标签是 n-hot。

比如，真实值为  $[0, 1, 1]$ （代表同时包含第二类和第三类），预测值为  $[0.1, 0.7, 0.8]$ （这里没有使用 softmax 计算预测值，而是使用 sigmoid 计算，将**每一个节点的输出归一化到  $[0, 1]$  之间**，所以所有预测值的和也不再为 1）。换句话说，每一个 Label 都是独立分布的，相互之间没有影响。所以交叉熵在这里是单独对每一个节点进行计算，每一个节点只有两种可能值，所以是一个二项分布。

对于二分类问题，可以简化一下交叉熵的计算公式为：

$$loss = -\hat{y}_i \log(y_i) - (1 - \hat{y}_i) \log(1 - y_i)$$

所以：

$$loss_{\text{第一类}} = -0 \times \log(0.1) - (1 - 0) \times \log(1 - 0.1) = -\log(0.9)$$

$$loss_{\text{第二类}} = -1 \times \log(0.7) - (1 - 1) \times \log(1 - 0.7) = -\log(0.7)$$

$$loss_{\text{第三类}} = -1 \times \log(0.8) - (1 - 1) \times \log(1 - 0.8) = -\log(0.8)$$

单张样本的 loss 即为： $loss_{\text{第一类}} + loss_{\text{第二类}} + loss_{\text{第三类}}$

## 7 总结

交叉熵能够衡量同一个随机变量中的两个不同概率分布的差异程度，在机器学习中就表示为真实概率分布与预测概率分布之间的差异。交叉熵的值越小，模型预测效果就越好。

交叉熵在分类问题中常常与 softmax 是标配，softmax 将输出的结果进行处理，使其多个分类的预测值和为 1，再通过交叉熵来计算损失。

## References

- [1] “交叉熵损失函数原理详解.” [Online]. Available: <https://blog.csdn.net/b1055077005/article/details/100152102>
- [2] “一文搞懂交叉熵在机器学习中的使用，透彻理解交叉熵背后的直觉.” [Online]. Available: <https://blog.csdn.net/tsyccnh/article/details/79163834>