知乎地址: https://zhuanlan.zhihu.com/c_1101089619118026752
作者: 小哲
github: https://github.com/lxztju/notes
微信公众号: 小哲AI

各种熵的计算公式及基本思想

1. 信息量

如果一个事件的概率很低,那么其信息量很大:

$$I(x) = -1 * log(p(x))$$

2. 信息熵

对于一个离散性随机变量X的熵H(X),信息熵就是信息量的数学期望, (熵越小越纯净,说明术语同一个类(决策树中), 熵越大,信息量越大,不确定性越高), 定义为:

$$H(X) = -1 * \sum p(x) * log(x)$$

3. 联合熵

对于服从联合概率分布p(x, y)的两个变量x, y,,那么联合熵:

$$H(X,Y) = -1 * \sum_{x \in X} \sum_{y \in Y} p(x,y) * log(p(x,y))$$

4. 条件熵

在X给定的条件下，Y的条件概率分布的熵对X 的数学期望(度量在定情况下,随机变量的不确定性):

$$
\begin{aligned}
H(Y|X) &= \sum_{x \in X} p(x) * H(Y|X=x) \\
&= -1 * \sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) * log(p(y|x)) \\
&= -1 * \sum_{x \in X} \sum_{y \in Y} p(y,x) * log(p(y|x))
\end{aligned}
$$

5. 互信息(好像和信息增益一样)

两个随机变量X和Y，他们的联合概率密度函数为p(x,y)，其边际概率密度函数分别为p(x)和p(y)。互信息I(X;Y)为联合分布p(x,y)和p(x)p(y)之间的相对熵

$$
I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) * log(\frac{p(x,y)}{p(x) * p(y)})
$$

互信息其实就是信息熵与条件熵之差(也就是知道其中一个,另一个不确定度减少的程度):

$$
I(x,y) = H(Y) - H(Y|X) = H(X) - H(X|Y)
$$

6. 相对熵

相对熵也叫做KL散度,表示对于同一个随机变量有两个概率分布P(X) 和Q(X), 衡量这两个分布的相似程度.

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

7. 交叉熵

主要用于度量两个概率分布间的差异性信息, 在分类任务中常用做目标函数(loss function),这里是不是有点疑惑,为什么KL散度用来衡量两个分布的相似程度,交叉熵也用来衡量,请往后看,交叉熵的公式为:

$$
\begin{aligned}
H(p,q) &= \sum_{i=1}^{n} p(x_i) * log(\frac{1}{q(x_i)}) \\
&= -1 * \sum_{i=1}^{n} p(x_i) * log(q(x_i))
\end{aligned}
$$

一般p为label,即真实标签ground truth, q为预测分布.

交叉熵与KL散度(相对熵的关系)

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

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

$$= H(X) - H(p, q)$$

KL散度就是随机变量X的信息熵减去交叉熵, 由于H(X)为常量,因此交叉熵与KL散度一样都是用来评估predict与label之间的差别.(一般采用交叉熵)