1.机器学习中的熵

1.信息量

1.1自信息

首先考虑一个离散的随机变量x,当我们观察到这个变量的一个具体值的时候,我们接收到多少信息呢?

我们暂时把信息看做在学习x的值时候的”惊讶程度”(这样非常便于理解且有意义)。

当我们知道一件**必然会发生的事情发生了**,比如往下掉的苹果.我们并不惊讶,因为反正这件事情会发生,因此可以认为我们**没有接收到信息**。

但是要是一件平时觉得**不可能发生的事情发生**了,那么我们**接收到的信息要大得多.**

因此,我们对于信息内容的度量就将依赖于概率分布p(x).

因此,我们想要寻找一个函数*h*(x)来表示信息的多少且是关于概率分布的单调函数.我们定义:

$$I(x) = -\log_2 p(x)$$

我们把这个公式叫做**信息量**的公式，前面的负号确保了信息一定是正数或者是0.(**低概率事件带来高的信息量**).有时候有人也叫做**自信息**（self-information）。

2.熵（entropy）

***信息量：*某个概率分布之下,某个概率值对应的信息量的公式。**

熵：整个概率分布对应的信息量的平均值.

$$\begin{aligned} H[x] &= E_{x\sim p}[I(x)] = -E_{x\sim p}[\log p(x)] \\ &= -\sum_x p(x)\log_2 p(x) \\ &= -\int p(x)\log_2 p(x)dx \end{aligned}$$

信息熵的本质可以看做是某个分布的**自信息**的**期望。**

熵越大,随机变量的不确定性就越大

3.相对熵（KL散度）

相对熵又称Kullback-Leible散度（即**KL散度**）。

设p(x)和q(x)是取值的两个概率概率分布，则p对q的相对熵为:

$$D(p\|q)=\sum_{x} p(x)\log\frac{p(x)}{q(x)}=E_{p(x)}\left(\log\frac{p(x)}{q(x)}\right)$$

在一定程度上面，相对熵可以度量两个随机变量的相似程度。当两个随机分布相同的时候，他们的相对熵为0，当两个随机分布的差别增大的时候，他们之间的相对熵也会增大。

4.交叉熵

衡量两个变量之间的差异程度。

$$H(p,q)=-\sum PlogQ$$

交叉熵与KL散度的关系：

$$H(p,q)=-\sum PlogQ==-\sum PlogP+\sum PlogP-\sum PlogQ$$

$$=H(P)+\sum PlogP/Q=H(P)+D_{KL}(P\|Q)$$

交叉熵就是 信息熵 与KL散度的和。

而信息熵是确定的,与模型的参数θ 无关,所以梯度下降求导时，**优化交叉熵和优化kl散度**（相对熵）是一样的；

5.互信息（Mutual Information）

是***一个随机变量中包含的关于另一个随机变量的信息量。衡量两个变量之间的相似程度.***

定义为：**联合分布** 和 **独立分布乘积** 的相对熵。

$$
\begin{aligned}
I(X,Y) &= D(P(X,Y) \,\|\, P(X)P(Y)) \\
&= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}
\end{aligned}
$$