



信息论


Allen . Huang

信息量


信息量是对信息的度量，就跟时间的度量是秒一样，当我们考虑一个离散的随机变量 x 的时候，当我们观察到的这个变量的一个具体值的时候，我们接收到了多少信息呢？

多少信息用信息量来衡量，我们接受到的信息量跟具体发生的事件有关。

信息的大小跟随机事件的概率有关。越小概率的事情发生了产生的信息量越大，如湖南产生的地震了；越大概率的事情发生了产生的信息量越小，如太阳从东边升起来了-其实代表着没有任何信息量




一个具体事件的信息量应该是随着其发生概率而递减的，且不能为负。



如果我们有俩个不相关的事件 x 和 y ，那么我们观察到的俩个事件同时发生时获得的信息应该等于观察到的事件各自发生时获得的信息之和，即：

$$h(x,y) = h(x) + h(y)$$



由于 x ， y 是俩个不相关的事件，那么满足 $p(x,y) = p(x)*p(y)$.

根据上面推导，可以看出：

$h(x)$ 一定与 $p(x)$ 的对数有关，因此我们有信息量公式如下：

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



(1) 为什么有一个负号

(2) 为什么底数为2



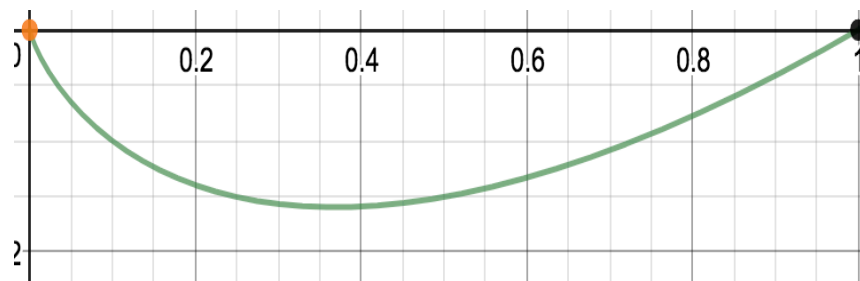
我们具体发生的事情总的来讲的期望信息量有多少呢？


熵

定义：用来度量信息的不确定程度。

解释：熵越大，信息量越大。不确定程度越低，熵越小，比如“明天太阳从东方升起”这句话的熵为0，因为这个句话没有带有任何信息，它描述的的是一个确定无疑的事情。

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





例子：假设有随机变量X，用来表达明天天气的情况。X可能出现三种状态 1) 晴天2) 雨天 3)阴天 每种状态的出现概率均为 $P(i) = 1/3$ ，那么根据熵的公式：

可以计算得到

$$H(X) = - 1/3 * \log(1/3) - 1/3 * \log(1/3) + 1/3 * \log(1/3) = \log 3 = 0.47712 \text{ (10为底)}$$

$$1.5849 \text{ (2为底)}$$

关于“熵”不同的教材会有所区别，例如2、10、e为底，事实上影响的是“单位”，为了统一单位，使用换底公式进行调换



如果这三种状态出现的概率为(0.1, 0.1, 0.8), 那么

$$H(X) = -0.1 * \log(0.1) * 2 - 0.8 * \log(0.8)$$

可以发现前面一种分布X的不确定程度很高，每种状态都很有可能。后面一种分布，X的不确定程度较低，第三种状态有很大概率会出现。所以对应前面一种分布，熵值很高，后面一种分布，熵值较低（2为底）。

$$\text{Log}(0.1) = -3.321$$

$$\text{Log}(0.8) = -0.321$$

$$H(X) = 0.3321 * 2 + 0.2568 = 0.921$$

思考

抛一枚有均匀正反面的硬币，和掷一个均匀六面的骰子，哪一种试验的不确定性更强一点呢？

设随机变量 X 为抛一枚均匀硬币的取值，其中正面朝上用 1 表示，反面朝上用 0 表示，于是有：

$$P\{X = 0, 1\} = \frac{1}{2} \quad \text{注：由于 } X = 0, X = 1 \text{ 概率均相等，为了版面整洁故合并表示。}$$
$$H(X) = -\frac{1}{2} \times \log \frac{1}{2} - \frac{1}{2} \times \log \frac{1}{2} = 1$$

设随机变量 Y 为掷一个六面均匀骰子的取值，其中 $Y = 1, 2, \dots, 6$ ，于是有：


$$P\{Y = 1, 2, \dots, 6\} = \frac{1}{6} \quad H(Y) = 6 \times \left(-\frac{1}{6} \log \frac{1}{6}\right) = \log 6$$

综上所述我们有如下结论：

$$H(X) = 1 = \log 2 < H(Y) = \log 6$$




必然事件的熵是多少呢？



信息熵还可以作为一个系统复杂程度的度量，如果系统越复杂，出现不同情况的种类越多，那么他的信息熵是比较大的。

如果一个系统越简单，出现情况种类很少（极端情况为1种情况，那么对应概率为1，那么对应的信息熵为0），此时的信息熵较小。



熵表明了单个随机变量的不确定程度，那么熵的值是确定不变的吗？我们有办法缩减这个不确定性吗？如果能缩减那缩减多少可以量化吗？

条件熵

定义：在一个条件下，随机变量的不确定性。

条件熵的公式：

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

举例说明：

假设随机变量X表示明天的天气情况，随机变量Y表示今天的湿度，Y 有两种状态

1) 潮湿 2) 干燥

假设基于以往的18个样本，X 的三种状态，概率均为 0.33，Y的两种状态，概率为0.5

Y \ X	晴天0	雨天1	阴天2
潮湿0	1	5	3
干燥1	5	1	3



条件概率可以通过朴素贝叶斯公式进行计算:

$$P(X=0|Y=0) = P(X=0, Y=0)/P(Y=0) = (1/18)/(9/18) = 1/9$$

$$P(X=1|Y=0) = P(X=1, Y=0)/P(Y=0) = (5/18)/(9/18) = 5/9$$

$$P(X=2|Y=0) = P(X=2, Y=0)/P(Y=0) = (3/18)/(9/18) = 3/9$$

$$P(X=0|Y=1) = P(X=0, Y=1)/P(Y=1) = (1/18)/(9/18) = 1/9$$

$$P(X=1|Y=1) = P(X=1, Y=1)/P(Y=1) = (5/18)/(9/18) = 5/9$$

$$P(X=2|Y=1) = P(X=2, Y=1)/P(Y=1) = (3/18)/(9/18) = 3/9$$



根据这个公式： 10为底


$$H(X|Y) = (1/18) * \log(1/9) + (5/18) * \log(5/9) + (3/18) * \log(3/9) + (1/18) * \log(1/9) + (5/18) * \log(5/9) + (3/18) * \log(3/9) = 0.406885$$

$$\text{信息增益} = \text{熵} - \frac{\text{条件熵}}{\text{熵}}$$

信息增益的应用：我们在利用进行分类的时候，常常选用信息增益更大的特征，信息增益大的特征对分类来说更加重要。决策树就是通过信息增益来构造的，信息增益大的特征往往被构造成底层的节点。

思考


假设现在我给你一枚硬币，告诉你这是均匀的，请你抛100次然后告诉我结果，结果你抛了100次后，记录的结果是：正面朝上90次，反面朝上10次，你就会开始怀疑“这真是一枚均匀的硬币吗？”



由第一部分熵中，我们知道，这一枚硬币的熵应该是1 bit，但是这样的试验之后，这枚硬币的熵还是1 bit吗？我们可以假设正面朝上的概率为0.9，反面朝上的概率为0.1，计算一下这个熵：

$$H(X|\hat{X}) = -0.9\log 0.9 - 0.1\log 0.1 \approx 0.469$$

其中， $H(X|\hat{X})$ 表示为知道90次正面朝上的事实后，原硬币的熵。




经过抛掷100次后，我们知道这枚硬币可能是不均匀的，且新的熵为0.469 bit，也就是说我们在知道90次正面朝上，10次反面朝下的事实之后，这个硬币的熵缩小了0.531 bit，这个0.531的信息量，我们就称为互信息。

互信息

定义：指的是两个随机变量之间的相关程度。

理解：确定随机变量X的值后，另一个随机变量Y不确定性的削弱程度，因而互信息取值最小为0，意味着给定一个随机变量对确定另一个随机变量没有关系，最大取值为随机变量的熵，意味着给定一个随机变量，能完全消除另一个随机变量的不确定性。这个概念和条件熵相对。

$$\begin{aligned} I(X; Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(y)p(x|y)}{p(x)p(y)} \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x|y)}{p(x)} = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x|y) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x) \\ &= - \sum_{x \in \mathcal{X}} p(x) \log p(x) - \left[- \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x|y) \right] = H(X) - H(X|Y) \end{aligned}$$



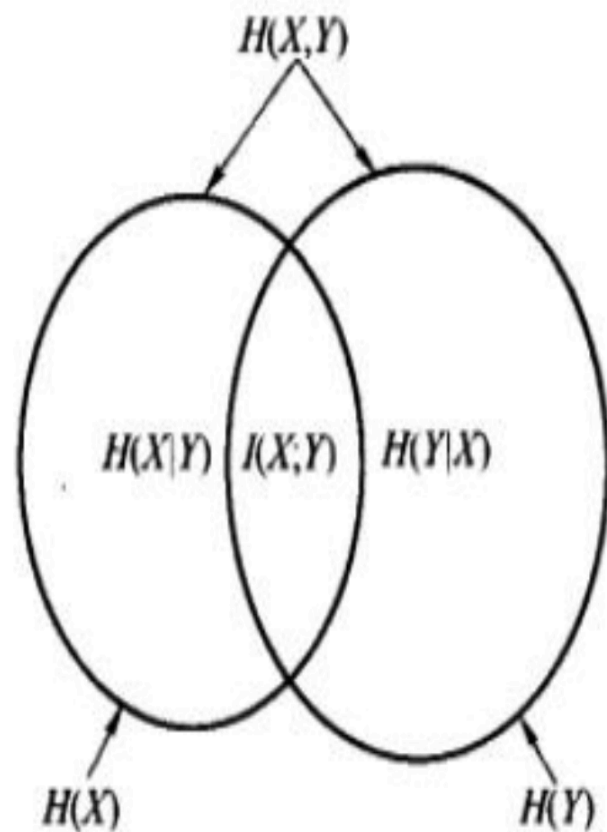
经过推导后，我们可以直观地看到 $H(X)$ 表示为原随机变量 X 的信息量， $H(X|Y)$ 为知道事实 Y 后 X 的信息量，互信息 $I(X;Y)$ 则表示为知道事实 Y 后，原来信息量减少了多少。

假设 X, Y 完全无关， $H(X) = H(X|Y)$ ，那么 $I(X; Y) = 0$

假设 X, Y 完全相关， $H(X|Y) = 0$ ，那么 $I(X; Y) = H(X)$

条件熵越大，互信息越小，条件熵越小，互信息越大。

互信息和信息增益实际是同一个值。



思考

根据上面的叙述，我们了解到：

信息论中，对于孤立的一个随机变量我们可以用熵来量化，对于两个随机变量有依赖关系，我们可以用互信息来量化，那么：

对于两个随机变量之间相差多少？

也就是说，这两个随机变量的分布函数相似吗？

如果不相似，那么它们之间差可以量化吗？

交叉熵

如果使用**估计的分布q**来表示来自**真实分布p**的平均编码长度，则

$$H(p, q) = - \sum_x p(x) \log q(x)$$

注意不要和联合熵混淆 $H(X, Y)$

$$H(p, q) = - \sum_x p(x) \log q(x)$$


假如X为一组已知的输入特征值，Y为一组已知的输出分类。优化的目标是为了找到一个映射模型F, 使得预测值 $Y_ = F(X)$ ，与真值Y最相似。但现实世界的Y和Y_的分布肯定不是完全一致的。

所以：

Y 服从 p分布（即真实分布）

Y_ 服从 q分布

交叉熵cross_entropy 即为描述p,q两个分布差异性的指标。



因为我们编码的样本来自于真实的分布 p ，所以乘的是真实概率。在图像分类的时候，比如softmax分类器，在训练的时候，我们已经给定图像的标签，所以这个时候每幅图片的真实概率就是1，这个时候的损失函数就是：

$$H(p, q) = - \sum_i \log(q_i)$$

交叉熵要大于等于真实分布的信息熵（最优编码）


对于样本服从分布 $P = \{p_1, p_2, \dots, p_n\}$, 对于其他任何概率分布 $Q = \{q_1, q_2, \dots, q_n\}$, 都有:

$$-\sum_{i=1}^n p_i \log(p_i) \leq -\sum_{i=1}^n p_i \log(q_i)$$

当且仅当 $p_i = q_i, i = 1, \dots, n$ 时, 等号成立。

相对熵

根据上面的叙述，我们了解到信息论中，对于孤立的一个随机变量我们可以用熵来量化，对于两个随机变量有依赖关系，我们可以用互信息来量化，那么对于两个随机变量之间相差多少？也就是说，这两个随机变量的分布函数相似吗？如果不相似，那么它们之间差可以量化吗？



由交叉熵可知，用估计的概率分布所需的编码长度，比真实分布的编码长，但是长多少呢？这个就需要另一个度量，相对熵，也称KL散度。

$$D(p||q) = H(p, q) - H(p) = - \sum_{i=1}^n p_i \log(q_i) - (- \sum_{i=1}^n p_i \log(p_i)) = \sum_{i=1}^n p_i \log \frac{p_i}{q_i}$$

总结

- 信息熵是衡量随机变量分布的混乱程度，是随机分布各事件发生的信息量的期望值，随机变量的取值个数越多，状态数也就越多，信息熵就越大，混乱程度就越大。当随机分布为均匀分布时，熵最大；信息熵推广到多维领域，则可得到联合信息熵 $H(X,Y)$ ；条件熵表示的是在 X 给定条件下， Y 的条件概率分布的熵对 X 的期望。
- 相对熵可以用来衡量两个概率分布之间的差异。
- 交叉熵可以用来衡量在给定的真实分布下，使用非真实分布所指定的策略消除系统的不确定性所需要付出的努力的大小。

或者：

- 信息熵是传输一个随机变量状态值所需的比特位下界（最短平均编码长度）。
- 相对熵是指用 q 来表示分布 p 额外需要的编码长度。
- 交叉熵是指用分布 q 来表示本来表示分布 p 的平均编码长度。