4.1熵

熵是信息的关键度量，熵衡量了预测随机变量的不确定度，不确定性越大熵越大。
针对随机变量XX，其信息熵的定义如下：

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

信息量

信息量是对信息的度量，就跟时间的度量是秒一样，多少信息用信息量来衡量，信息量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。

因此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，我们很容易看出h(x)一定与p(x)的对数有关（因为只有对数形式的真数相乘之后，能够对应对数的相加形式，可以试试）。因此我们有信息量公式如下：

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

两个问题

（1）为什么有一个负号

负号是为了确保信息量一定是正数或者是0

（2）为什么底数为2

这是因为，我们只需要信息量满足低概率事件x对应于高的信息量。那么对数的选择是任意的。我们只是遵循信息论的普遍传统，使用2作为对数的底。

信息量度量的是一个具体事件发生了所带来的信息，而**熵则是在结果出来之前对可能产生的信息量的期望**

考虑该随机变量的所有可能取值，即所有可能发生事件所带来的信息量的期望。即

$$H(x) = -sum(p(x)log_2 p(x))$$

转换一下为：

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

注意：1. 当式中的对数的底为2时，信息熵的单位为比特。它底数为其它时，它对应的单位也不一样。

2. 信息熵是信息论中用于度量信息量的一个概念。一个系统越是有序，信息熵就越低；反之，一个系统越是混乱，信息熵就越高。所以，信息熵也可以说是系统有序化程度的一个度量。

4.2联合熵

联合熵是一集变量之间不确定的衡量手段。

两个随机变量X，Y的联合分布，可以形成联合熵，用H(X,Y)表示。

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{n} P(x_i, y_j) log P(x_i, y_j)$$

4.3条件熵

定义为X给定条件下，Y的条件概率分布的熵对X的数学期望

在随机变量X发生的前提下，随机变量Y发生所新带来的熵定义为Y的条件熵，用H(Y|X)表示，用来衡量在已知随机变量X的条件下随机变量Y的不确定性。

$$H(Y|X)=\sum\sum -p(x,y)\log(p(y|x))$$

联合熵和条件熵的关系是：

$$H(X,Y)=H(X)+H(Y|X)=H(Y)+H(X|Y)=H(Y,X)$$

整个式子表示(X,Y)发生所包含的熵减去X单独发生包含的熵。

注意：1. 这个条件熵，不是指在给定某个数（某个变量为某个值）的情况下，另一个变量的熵是多少，而是**期望**！　因为条件熵中X也是一个变量，意思是在一个变量X的条件下（变量X的每个值都会取），另一个变量Y熵对X的期望。2. 在计算信息增益的时候，经常需要用到条件熵。信息增益（information gain）是指期望信息或者信息熵的有效减少量（通常用“字节”衡量）。通常表示为：信息熵 - 条件熵。