信息熵、条件熵、联合熵、互信息、相对熵、交叉熵

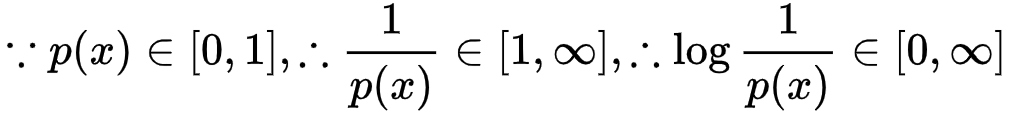
1. 信息量

信息量是通过概率来定义的：如果一件事情的概率很低，那么它的信息量就很大；反之，如果一件事情的概率很高，它的信息量就很低。简而言之，概率小的事件信息量大，因此信息量可以定义如下：

下面解释下为什么要取倒数再取对数。

* 1：先取倒数：表示，信息量和概率呈反比；
* 2：再取对数：log取对数是为了将区间[1,]映射到[0,]。

再总结一下：



1. 信息熵

信息熵是信息量的数学期望。理解了信息量，信息熵的定义式便不难理解。定义如下：

* 熵越小表示越纯，数据越有序；熵越大混乱程度越大，熵越大信息量越大，数据越无序。比如决策树算法在进行特征选择时的一个标准就是使得通过该特征分类以后的类的熵最小。
* 再说一个对信息熵的理解：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，那么它的信息熵也就越大。如果一个系统越简单，出现情况种类越少，那么它的信息熵越小。

1. 条件熵

条件熵的定义为：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。

这个还是比较抽象，下面我们具体解释一下，设有随机变量(X,Y)，其联合概率分布为：

条件熵H(Y|X)表示在已知随机变量X的条件下随机变量Y的不确定性。

下面推导一下条件熵的公式：

注意，这个条件熵，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，变量的不确定性是多少？

因为条件熵中X也是一个变量，意思就是在一个变量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。

下面通过一个例子来说明。



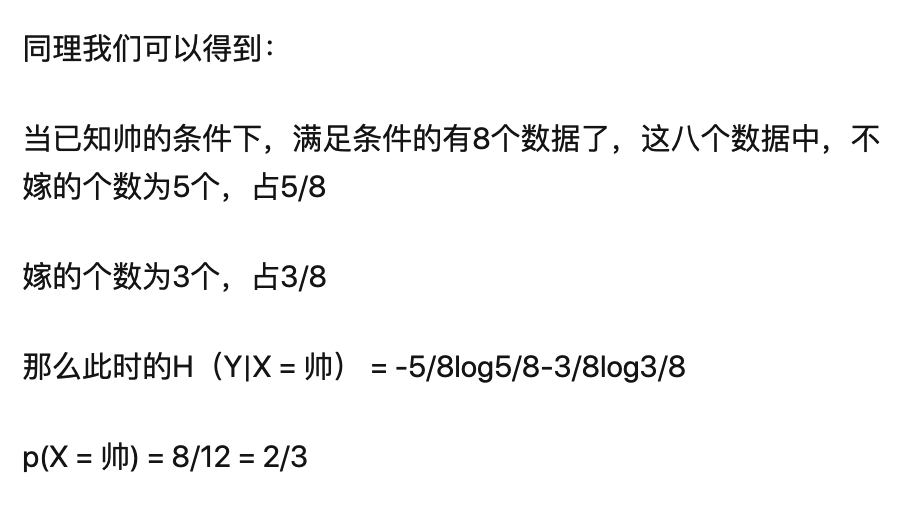
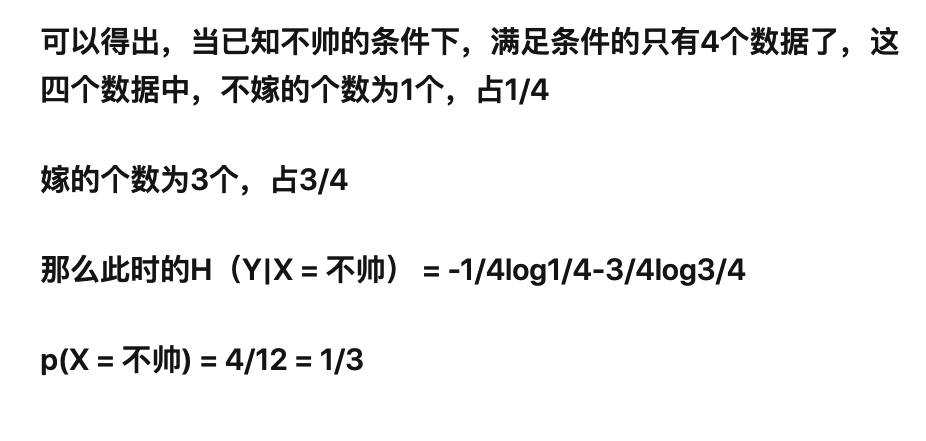
假设有上面的数据，设随机变量Y={嫁，不嫁}

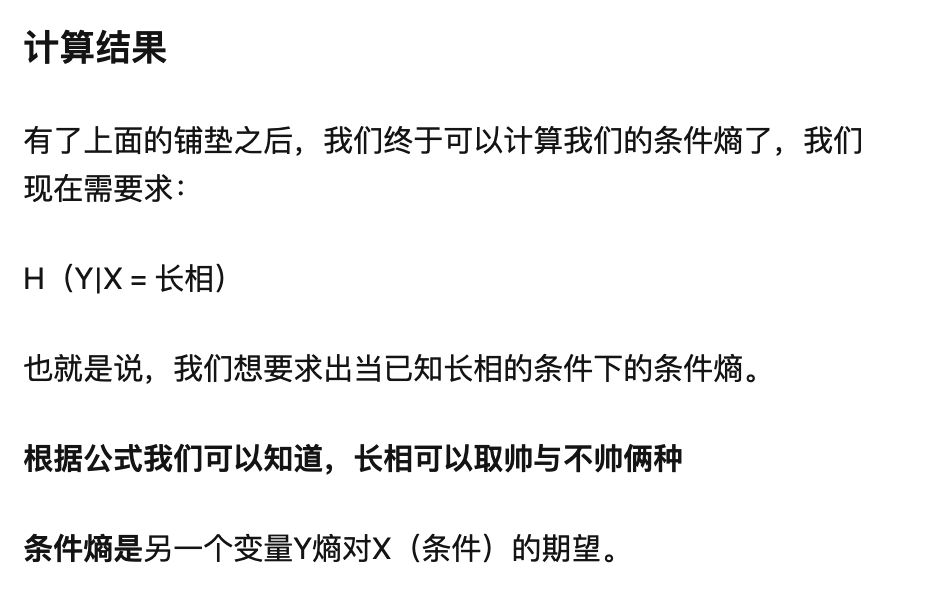
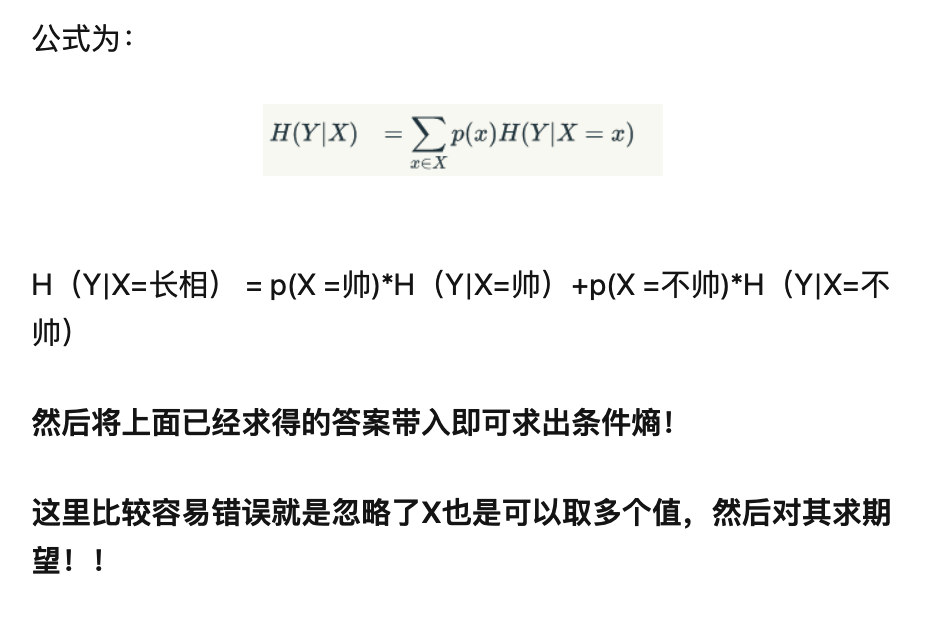
我们可以统计出嫁的个数为6/12=1/2，不嫁的个数为6/12=1/2。

那么Y的熵，根据熵的公式来计算，可以得到H(Y) = -1/2log1/2 -1/2log1/2。

为了引出条件熵，我们现在还有一个变量X，代表长相是帅还是不帅，当长相是帅的时候，统计如下红色：







总结：其实条件熵意思是按一个新的变量的每个值对原变量进行分类，比如上面这个题把嫁与不嫁按照帅与不帅分为了两类。

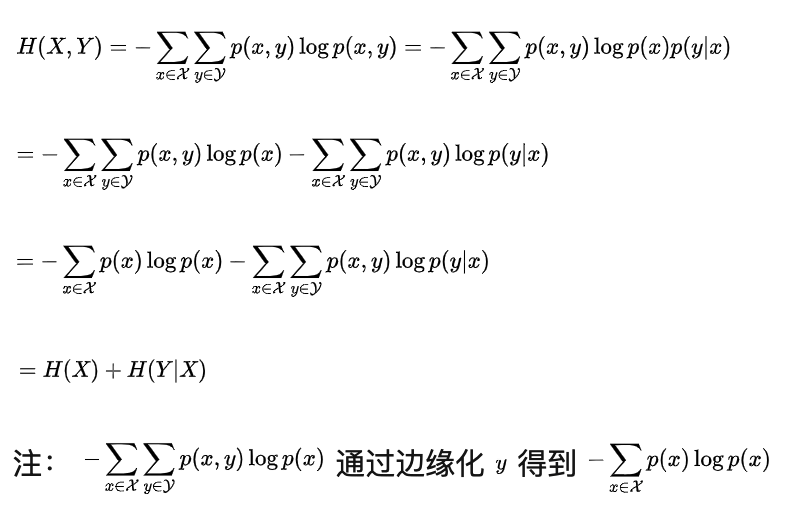
然后在每一个小类里面，都计算一个小熵，然后每一个小熵乘以各个类别的概率，然后求和。

1. 联合熵

两个变量X和Y的联合熵的表达式如下：

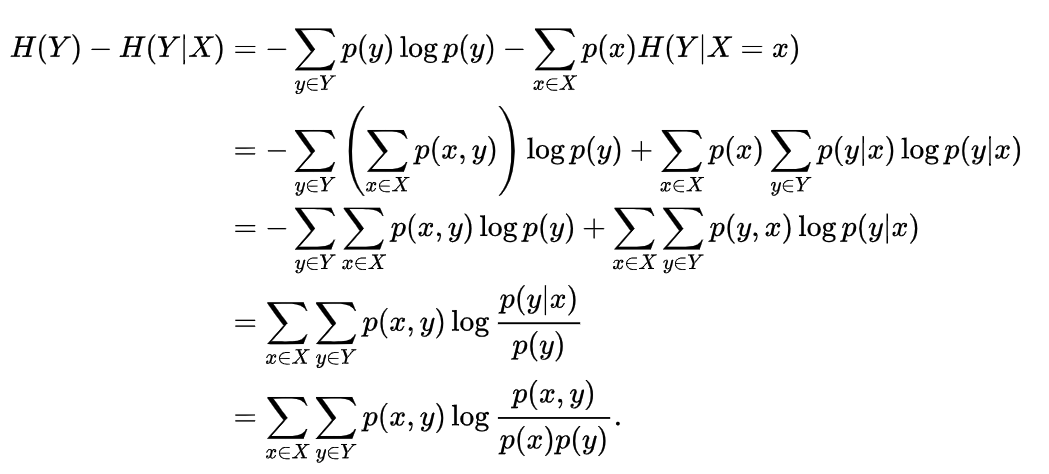
与其他熵的关系，在条件熵的定义中，使用了联合熵：

推导过程如下：

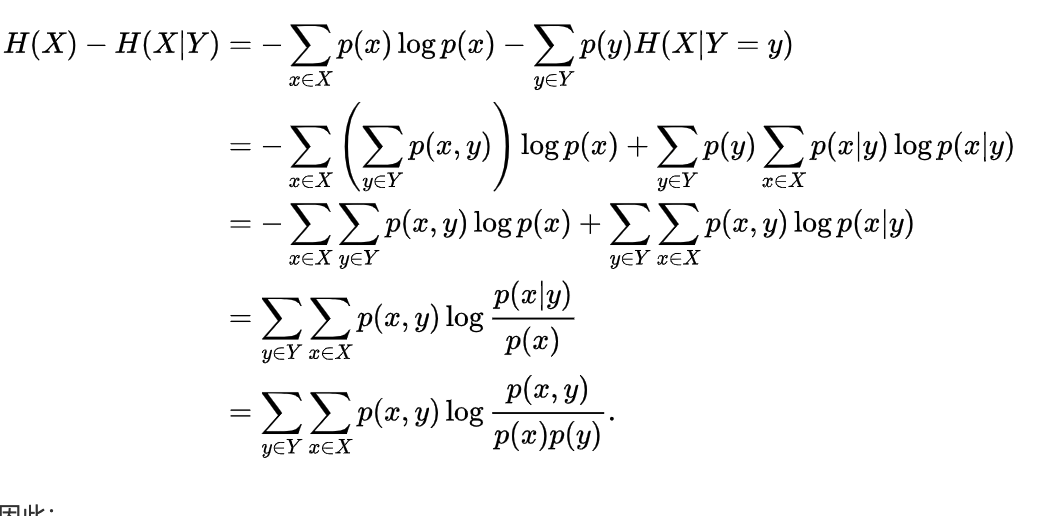


1. 互信息

根据信息熵、条件熵的定义式，可以计算出信息熵与条件熵之差：



同理，可以算出：

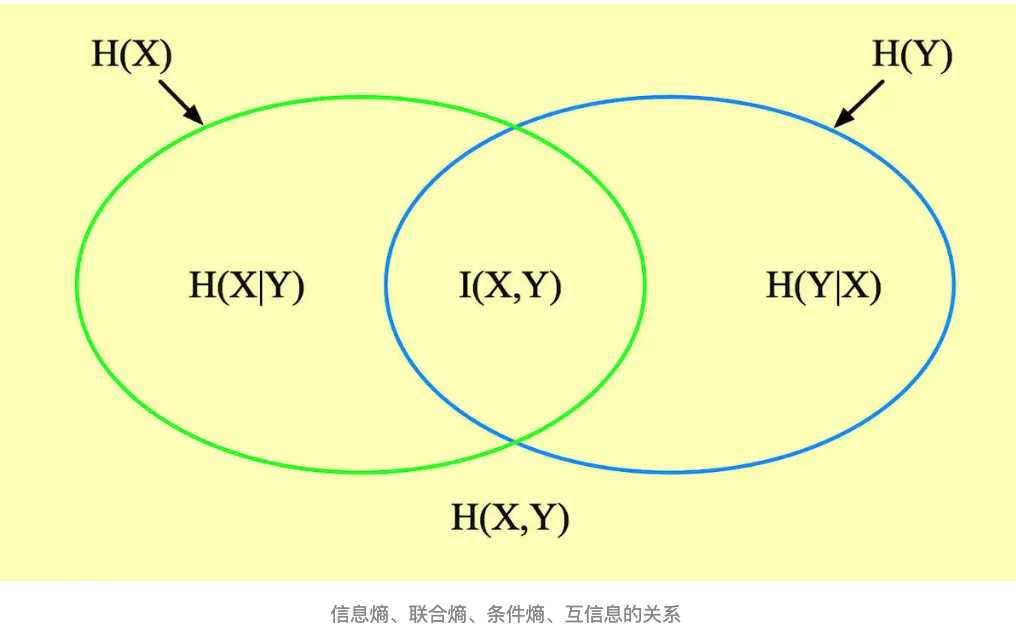


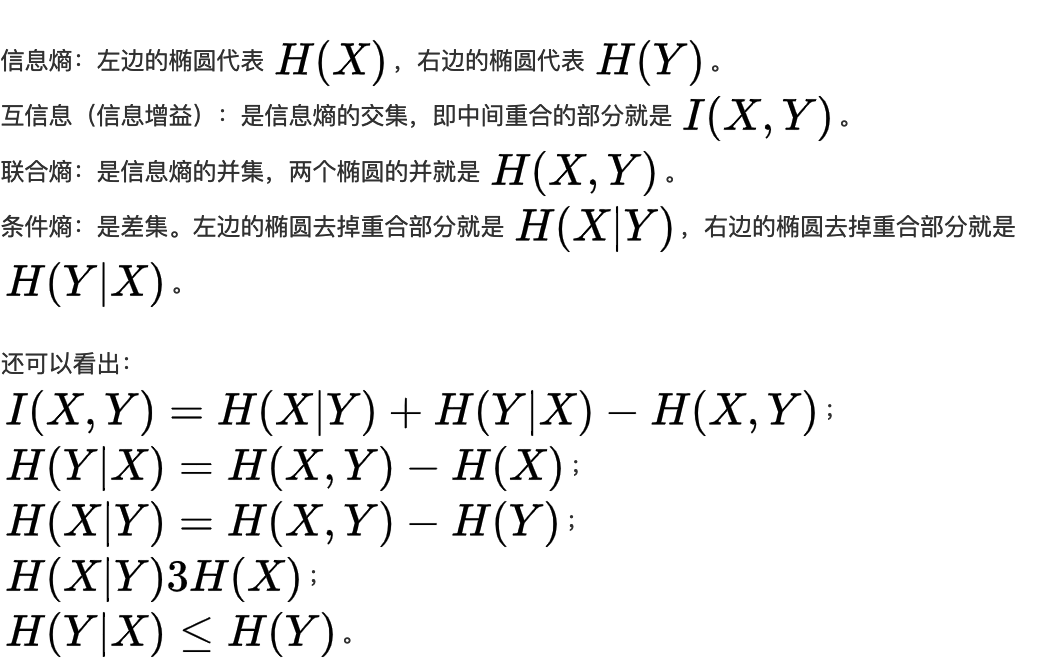
因此，可以得到：

那么，互信息(信息增益)可以定义为：

I(X,Y) = H(Y) – H(Y|X) = H(X) – H(X|Y)

下面用一张图看清他们之间的关系：

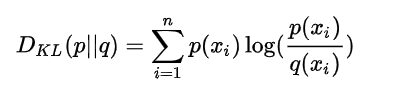




1. 相对熵

相对熵也叫做KL散度，如果我们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)，使用KL散度来衡量这两个分布的差异。差异越大则相对熵越大，差异越小则相对熵越小，一般的取值范围为(0,1)，相对熵我们希望它越小越好。

计算公式如下：



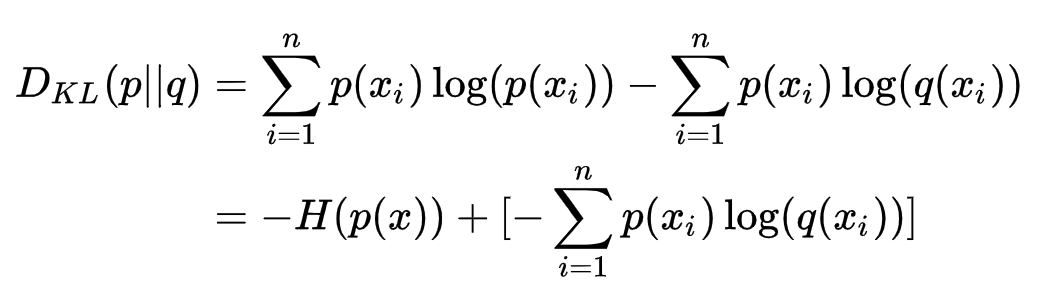
1. 交叉熵

交叉熵容易和相对熵弄混，二者练习紧密，但又有所区别。假设有两个分布p, q，则他们在给定样本集上的交叉熵定义如下：

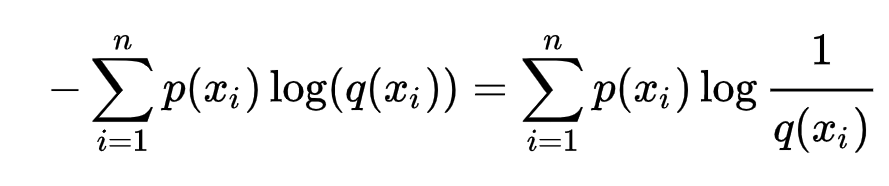
交叉熵我们希望它越小越小。

相对熵和交叉熵的关系：

这个可以从相对熵的定义式推导出来：



这里：



就是交叉熵的定义式。