前面我们总结了信息熵的概念通俗理解信息熵 - 知乎专栏,这次我们来理解一下条件熵。

我们首先知道信息熵是**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。公式如下：**

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

我们的条件熵的定义是：**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**

这个还是比较抽象，下面我们解释一下：

设有随机变量（X,Y），其联合概率分布为

$$p(X = x_i, Y = y_i) = p_{ij}, \ i = 1, 2, \ldots, n; \ j = 1, 2, \ldots, m$$

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。**随机变量X给定的条件下随机变量**Y的条件熵H(Y|X)**

公式

下面推导一下条件熵的公式：

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

注意

注意，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，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。

这是最容易错的!

例子

下面通过例子来解释一下：

帅？	性格好？	身高？	上进？	嫁与否
帅	不好	矮	不上进	不嫁
不帅	好	矮	上进	不嫁
帅	好	矮	上进	嫁
不帅	爆好	高	上进	嫁
帅	不好	矮	上进	不嫁
帅	不好	矮	上进	不嫁
帅	好	高	不上进	嫁
不帅	好	中	上进	嫁
帅	爆好	中	上进	嫁
不帅	不好	高	上进	嫁
帅	好	矮	不上进	不嫁
帅	好	矮	不上进	不嫁

假如我们有上面数据：

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，代表长相是帅还是帅，当长相是不帅的时候，统计如下红色所示：

帅？	性格好？	身高？	上进？	嫁与否
帅	不好	矮	不上进	不嫁
不帅	好	矮	上进	不嫁
帅	好	矮	上进	嫁
不帅	好	高	上进	嫁
帅	不好	矮	上进	不嫁
帅	不好	矮	上进	不嫁
帅	好	高	不上进	嫁
不帅	好	中	上进	嫁
帅	好	中	上进	嫁
不帅	不好	高	上进	嫁
帅	好	矮	不上进	不嫁
帅	好	矮	不上进	不嫁

可以得出，当已知不帅的条件下，满足条件的只有4个数据了，这四个数据中，不嫁的个数为1个，占1/4

嫁的个数为3个，占3/4

那么此时的H（Y|X = 不帅） = -1/4log1/4-3/4log3/4

p(X = 不帅) = 4/12 = 1/3

同理我们可以得到：

当已知帅的条件下，满足条件的有8个数据了，这八个数据中，不嫁的个数为5个，占5/8

嫁的个数为3个，占3/8

那么此时的H（Y|X = 帅） = -5/8log5/8-3/8log3/8

p(X = 帅) = 8/12 = 2/3

计算结果

有了上面的铺垫之后，我们终于可以计算我们的条件熵了，我们现在需要求：

H（Y|X = 长相）

也就是说，我们想要求出当已知长相的条件下的条件熵。

根据公式我们可以知道，长相可以取帅与不帅俩种

条件熵是另一个变量Y熵对X（条件）的期望。

公式为：

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X=x)$$

H（Y|X=长相） = p(X =帅)*H（Y|X=帅）+p(X =不帅)*H（Y|X=不帅）

然后将上面已经求得的答案带入即可求出条件熵！

这里比较容易错误就是忽略了X也是可以取多个值，然后对其求期望！！

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帅，不帅分成了俩类。

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。

我们用另一个变量对原变量分类后，原变量的不确定性就会减小了，因为新增了Y的信息，可以感受一下。不确定程度减少了多少就是信息的增益。

后面会讲信息增益的概念，信息增益也是决策树算法的关键。

致谢：

德川，皓宇，继豪，施琦