**交叉熵原理：**

信息量：交叉熵是信息论中的概念，首先要知道什么是信息量。假设我们听到了两件事，分别如下：

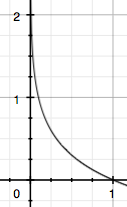
事件A：巴西队进入了2018世界杯决赛圈。

事件B：中国队进入了2018世界杯决赛圈。

仅凭直觉来说，显而易见事件B的信息量比事件A的信息量要大。究其原因，是因为事件A发生的概率很大，事件B发生的概率很小。所以当越不可能的事件发生了，我们获取到的信息量就越大。越可能发生的事件发生了，我们获取到的信息量就越小。那么信息量应该和事件发生的概率有关。

信息量的定义：

假设是离散型随机变量，其取值集合为，概率分布函数 p(X)=Pr(=x), xϵ,则定义事件=x0的信息量为：

，由于是概率所以p(x0)的取值范围是[0,1],绘制图如下：

该图符合我们对信息量的直觉

**熵：**考虑另外一个问题，对于某个事件，有n种可能性，每一种可能性都有一个概率p(x)这样就可以计算出某一种可能性的信息量。举一个例子，假设你拿出了你的电脑，按下开关，会有三种可能性，下表列出了每一种可能的概率及其对应的信息量



用熵来表示所有信息量的期望，即：



其中n代表所有的n种可能性，所以上面问题的结果就是

H(x)=0.7\*036+0.2\*1.61+0.1+2.30=0.804

相对熵(KL散度):

相对熵又称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)，我们可以使用 KL 散度（Kullback-Leibler (KL) divergence）来衡量这两个分布的差异，

在机器学习中，**P**往往用来表示样本的真实分布，比如 [1,0,0] 表示当前样本属于第一类。**Q**用来表示模型所预测的分布，比如 [0.7,0.2,0.1]

直观的理解就是如果用P来描述样本，那么就非常完美。而用Q来描述样本，虽然可以大致描述，但是不是那么的完美，信息量不足，需要额外的一些“信息增量”才能达到和P一样完美的描述。如果我们的Q通过反复训练，也能完美的描述样本，那么就不再需要额外的“信息增量”，Q等价于P。

**KL散度的计算公式：**



n为事件的所有可能性，D的值越小，表示q分布和q分布越接近。

例：



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 事件 | 真实概率p | 预测概率q | 信息量I |
| A | 猫 | 1 | 0.7 | -log(q(A))=0.36 |
| B | 狗 | 0 | 0.2 | -log(q(B))=1.61 |
| C | 兔 | 0 | 0.1 | -log(q(C))=2.30 |



**交叉熵：**对KL散度的计算公式的变形，可得：



等式的前一部分恰好是p的熵，等式的后一部分，就是交叉熵：



在机器学习中，我们需要评估label和predicts之间的差距，使用KL散度刚刚好，即DKL(y||y^)，由于KL散度中的前一部分−H(y)不变，故在优化过程中，只需要关注交叉熵就可以了。所以一般在机器学习中直接用交叉熵做loss，评估模型。

**交叉熵在多分类问题中的使用：**

这里的多类别是指，每一张图像样本可以有多个类别，比如同时包含一只猫和一只狗 和单分类问题的标签不同，多分类的标签是n-hot。

比如下面这张样本图，即有青蛙，又有老鼠，所以是一个多分类问题



对应的标签和预测值：



值得注意的是，这里的Pred不再是通过softmax计算的了，这里采用的是sigmoid。将每一个节点的输出归一化到[0,1]之间。所有Pred值的和也不再为1。换句话说，就是每一个Label都是独立分布的，相互之间没有影响。所以交叉熵在这里是单独对每一个节点进行计算，每一个节点只有两种可能值，所以是一个二项分布。前面说过对于二项分布这种特殊的分布，熵的计算可以进行简化。

同样的，交叉熵的计算也可以简化，即

loss=−ylog(y^)−(1−y)log(1−y^)

注意，上式只是针对一个节点的计算公式。这一点一定要和单分类loss区分开来。

例子中可以计算为：

