1 经验熵

假设对随机变量 Y 独立同分布采样 N 次，得到数据集 D。在 D 中，Y 有 K 个不同的取值，每个取值的集合记为 D_k，$k=1,\ldots,K$。记 $|D|$ 为 D 中元素的个数，即 $|D|=N$。同理，记 $|D_k|$ 为集合 D_k 中元素的个数。那么随机变量 Y 在样本集 D 下的经验熵 为

$$H(Y)=-\sum_{k=1}^{K}\frac{|D_k|}{|D|}\log_2\frac{|D_k|}{|D|} \tag{1.1}$$

相当于，Y 的第 k 个取值的概率 $p(y_k)=\frac{|D_k|}{|D|}$。

2 经验条件熵

假设有随机变量 X 和 Y，对他们进行独立同分布采样 N 次，得到数据集 $D=\{(x_1,y_1),\ldots,(x_N,y_N)\}$，$X$ 和 Y 的每个取值对的集合记为 D_{ik}，$i=1,\ldots,I$，$k=1,\ldots,K$。那么随机变量 X 对 Y 在样本集 D 下的经验条件熵 为

$$H(Y|X)=-\sum_{i=1}^{I}\frac{|D_i|}{|D|}\sum_{k=1}^{K}\frac{|D_{ik}|}{|D_i|}\log_2\frac{|D_{ik}|}{|D_i|} \tag{2.1}$$

其中，D_i 为 D_{ik} 在所有 k 下的并集。相当于，$p(y_k|x_i)=\frac{|D_{ik}|}{|D_i|}$。

3 决策树中的经验条件熵

在决策树中，我们希望使用一个指标（也叫特征）去进行判断，当我们选择的那个特征，可以最大限度的降低判断的不确定性，这个特征就是最优的选择。假设我们想要判断的输出是随机变量 Y，我们使用的特征是 X，那不确定性的减少量是 $H(Y)-H(Y|X)$，也就是互信息 $I(X;Y)$。

以我们的数据集为例，首先计算输出 Y 的经验熵，即 $visit\_library\_in\_Sunday$ 的经验熵。在数据集中，一共 15 个样本，$visit\_library\_in\_Sunday=True$ 的样本有 9 个。

$$H(Y)=-\frac{9}{15}\log_2\frac{9}{15}-\frac{6}{15}\log_2\frac{6}{15}=0.971 \tag{3.1}$$

然后计算特征 X_1：*age* 的对输出的条件经验熵，

$$H(Y|X_1) = -\frac{1}{3}*\frac{2}{5}\log_2\frac{2}{5}-\frac{1}{3}*\frac{3}{5}\log_2\frac{3}{5} \tag{3.2}$$

$$-\frac{1}{3}*\frac{3}{5}\log_2\frac{3}{5}-\frac{1}{3}*\frac{2}{5}\log_2\frac{2}{5} \tag{3.3}$$

$$-\frac{1}{3}*\frac{4}{5}\log_2\frac{4}{5}-\frac{1}{3}*\frac{1}{5}\log_2\frac{1}{5} \tag{3.4}$$

$$=0.888 \tag{3.5}$$

不确定性的减少量是 $H(Y)-H(Y|X_1)=0.971-0.888=0.083$

同理，我们可以计算得到 X_2：*male* 和 X_3：*single* 的条件经验熵与不确定性减少量

$$H(Y)-H(Y|X_2) = 0.971-\frac{5}{15}*\frac{5}{5}\log_2\frac{5}{5}-\frac{5}{15}*\frac{0}{5}\log_2\frac{0}{5} \tag{3.6}$$

$$-\frac{10}{15}*\frac{4}{10}\log_2\frac{4}{10}-\frac{10}{15}*\frac{6}{10}\log_2\frac{6}{10} \tag{3.7}$$

$$=0.324 \tag{3.8}$$

$$H(Y)-H(Y|X_3) = 0.971-\frac{6}{15}*\frac{6}{6}\log_2\frac{6}{6}-\frac{6}{15}*\frac{0}{6}\log_2\frac{0}{6} \tag{3.9}$$

$$-\frac{9}{15}*\frac{3}{9}\log_2\frac{3}{9}-\frac{9}{15}*\frac{3}{9}\log_2\frac{3}{9} \tag{3.10}$$

$$=0.420 \tag{3.11}$$

因此，我们选用 *single* 来判断 *visit_library_in_Sunday* 时，不确定性减少的最多，准确率也最高。