链滴

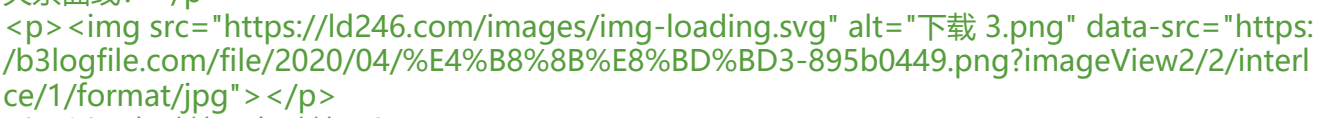
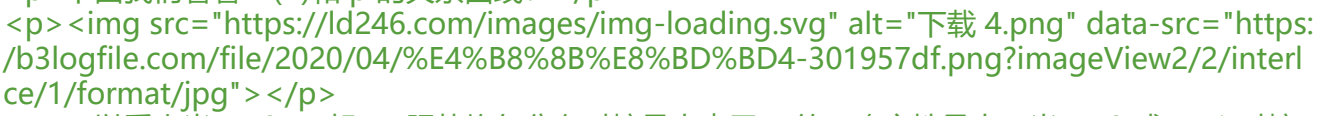
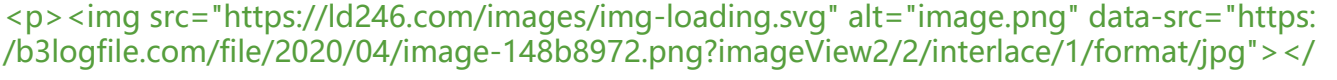
熵、联和熵、条件熵、相对熵、交叉熵、互信息

作者：hailangjiang

原文链接：https://ld246.com/article/1587918032202

来源网站：链滴

许可协议：署名-相同方式共享 4.0 国际 (CC BY-SA 4.0)

<h2 id="自信息">自信息</h2>
熵是信息论中的概念，在讲熵之前先引入自信息的概念。设离散型随机变量 X，p(x)=P(X=x)，自信息的定义为:
i(x)=-logp(x)
概率是对随机变量确定性的度量，而自信息则是对随机变量不确定性的度量。通过定义也可以看出自信息 i(x)与概率 p(x)负相关。
举个栗子，高富帅追到白富美的概率很大，也可以说这个事件没有什么信息量或者不确定性很小因为这本来就是一件大概率、'显而易见'的事情。而矮矬穷追到白富美那信息量或者说不确定性就很了，也就是概率很小。
其中对数 log 的底为 2，自信息的单位是比特（bit）；当对数 log 的底为 e 时，自信息的单位是奈特(nat)，一般情况默认为比特。
对于概率 p(x),有 0<=p(x)<=1,而对于自信息 i(x)有 i(x)>=0,下面我们看看自信息和概率关系曲线：

<h2 id="各种熵">各种熵</h2>
<h3 id="熵">熵</h3>
熵也是对随机变量不确定性的度量。设 X 是一个离散型随机变量，X 的取值空间为\chi,p(x)=P(X=x) ，离散型随机变量 X 的熵 H(X)定义为：
H(X)=\sum_{x\in \chi}-p(x)logp(x)
可以看出熵其实就是自信息的期望。既然说熵是对随机变量不确定性的度量，那么什么时候不确性最大呢？当 X 服从均匀分布的时候不确定性最大，也就是 X 的所有取值的概相等时熵是最大的。当 x\rightarrow0时，xlogx=0，所约定 0log0=0。对于熵 H(X)有 0=<H(x)<=logn，n 为 X 可能的取值个数。
举个栗子，设
X=\begin{cases}
1,概率为p\\
0,概率为(1-p)
\end{cases}
那么
H(X)=-plogp-(1-p)log(1-p)
下面我们看看 H(X)和 p 的关系曲线：

可以看出当 p=0.5 时即 X 服从均匀分布时熵最大表示 X 的不确定性最大，当 p=0 或 p=1 时熵 0，这也符合熵的定义，表明 p=0 或 p=1 时 X 不具有不确定性。
<h3 id="联和熵">联和熵</h3>
将定义推广至两个随机变量的情况，类似熵的定义。对于服从联合分布 p(x,y)的离散型随机变量其联和熵 H(X,Y)定义为：
H(X,Y)=-\sum_{x\in \chi,y\in \gamma}p(x,y)logp(x,y)
<h3 id="条件熵">条件熵</h3>
若（X,Y）~P（x,y），即离散型随机变量 X，Y 服从概率分布 p(x,y)，条件熵 H(Y|X)定义为：

<h3 id="熵-条件熵-联和熵的关系">熵、条件熵、联和熵的关系</h3>
熵、条件熵、联和熵满足如下关系：
H(X,Y)=H(X)+H(Y|X)\\
H(X,Y)=H(Y)+H(X|Y)
证明： 

">
<h3 id="相对熵">相对熵</h3>
相对熵又称为 KL 散度（Kullback-Leibler divergence）或信息度，是两个随机分布之间的不对称距离度量，定义两个概率密度函为 p(x)和 q(x)之间的相对熵或 KL 散度为：

E_p为 p(x)的期望，在上述定义中约定{}0log\frac{0}{0}=0,plog\frac{p}{0}=\infty,0log\frac{0}{q}=0 。
<h3 id="交叉熵-Cross-Entropy-">交叉熵(Cross-Entropy)</h3>
交叉熵和相对熵非常相似，常用作机器学习中代价函数，定义两个概率密度函数为 p(x)和 q(x)之的交叉熵 CE(p,q)为：
CE(p,q)=-\sum_{x\in\chi}p(x)logq(x)=-\sum_{x\in\chi}p(x)log\fra{p(x)q(x)}{p(x)}=H(p)+D(p||q)
可以看出交叉熵于相对熵仅相差一个 H(p)，当交叉熵用作代价函数时 p(x)为真实分布，q(x)为测分布，此时 H(p)可以看作一个常数，最小化交叉熵就等价于最小化相对熵。
<h3 id="互信息">互信息</h3>
互信息是一个随机变量包含另一个随机变量信息量的度量。在决策树算法，互信息也叫信息增益，这时候互信息/信息增益可以理解为在给定一随机变量知识的情况下，原随机变量不确定性的缩减量。设随机变量 X 和 Y，他们的联概率密度函数为 p(x,y)，边际概率密度函数为 p(x),p(y)。互信息 I(X;Y)定义为联合分布 p(x,y)和乘积布 p(x)p(y)之间的相对熵：

<h3 id="熵与互信息的关系">熵与互信息的关系</h3>

由这个式子看出 I(X;Y)是在给定 Y 知识条件下 X 的不确定度的缩减量。
互信息是对称的，同样可以得到I(X;Y)=H(Y)-H(Y|X)由上面 H(X,Y)=H(X)+H(Y|X),互信息还可以写成I(X;Y)=H(X)+H(Y)H(X,Y)。
综上，互信息与熵有以下关系：
最后，关于熵、联和熵、条件熵、互信息的概念可以通过以下维恩图记忆：
<h2 id="注-">注：</h2>
本文中截图来自 Thomas M.cover 著 Elements of Information Theory
<h2 id="Reference">Reference</h2>
Elements of Information Theory