信息论----信息学的熵

2016年4月8日 星期五　　下午10:52

信息熵、自信息、互信息、条件熵、联合熵、KL散度（相对熵）、交叉熵

一、信息熵

名称：熵、信息熵、信源熵、平均自信息量、香农熵。

作用：信息的混乱程度、信息量、不确定性。

物理意义：平均最短字节长度

公式：H（P）=-ΣP（x）logP（x） 、自信息的期望、-log（P）的期望

解释：

（维基百科中熵的介绍很不错）

熵，来源于热力学，在热力学中，代表热力学系统无序的程度。

举一个特殊情况的例子：

随机变量中，只有一个事件A的概率是1，其他n-1个事件都是0。这

强、平均信息量很小、平均编码长度很短（试想使用最优的编码----

编码是0（长度是1），其他编码无所谓，可以是111、100、1010这

编码长度），熵为1.

再举一个特殊情况的例子：

均匀分布n个变量概率都是1/n，导致很不好预测、完全没有规律、

码长度很长（试想使用最优的编码----huffman编码，很难做到平均

大-log（1/n）。

实际遇到的问题常常介于上述两种情况之间，概率分布越均匀、规律性越

大。对于一个事件而言，我们衡量它的自信息，也就是这个事件的不确定

定性越大，发生这个事件所包含的信息量越大，我们越能通过接收到这个

例子：

“一个每天八点上班的人，今天十一点还没上班”可能说明他有某些事情

样很好预测、规律性很
huffman编码，事件A的
样的不过不影响最终平均

平均信息量很大、平均编
编码长度小），熵是最

越小、约不好预测、熵越
性。事件的概率越小不确
个事件来发现一些问题。

情，“一个每天十一点上班

的人，今天十一点还没上班”，仅仅可能是他偶然晚了几分钟，并不能推
“一个外地人出现在凶杀现场”这个人很可能就是凶手，我们能得到很多
在这里的几率很低，“一个住在附近的老大爷出现在凶杀现场”并不能说
来的几率很高。

计算

如何定理的计算熵。

熵是接收的每条消息中包含的信息的平均量。

所包含的信息量如何衡量？对于一个随机变量而言，我们能掌握的信息是
越大，

（以下是我的理解）

我们可以根据概率对随机变量做编码，编码的长度代表信息量。这里我们
以编码必须是最优的，没有冗余的----huffman编码。

最优编码下，这个随机变量在各个时间下的平均字节长度，就是熵。

这个平均字节长度，不是直接平均，而是用概率进行加权的。其实我们发
节长度的期望。

（from维基百科）熵实际是对随机变量的比特量和顺次发生概率相乘再总

例子：

假设一个随机变量X，取三种可能值x1，x2，x3，概率分别为，1/2、1/4、
分别是0、10、11。比特长度/编码长度分别是1、2、2。平均字节长度是

定义：

依据Boltzmann's H-theorem，香农把随机变量X的熵值 H（希腊字母E
{x1, ..., xn}：

$$\mathrm{H}(X) = \mathrm{E}[\mathrm{I}(X)] = \mathrm{E}[-\ln(\mathrm{P}(X))]$$

这里ln如果以2为底的话，熵的单位是bit，数值就等于平均字节长度

性质：

推断出更多的信息。
的信息，因为他平时出现
说明什么，因为他经常来，

是其概率，概率越小信息量

们要用编码表示信息量，所

发现也就是随机变量平均字

总和的数学期望。

、1/4，那么huffman编码
3/2，其熵为3/2。

ta）定义如下，其值域为

连续性

该量度应连续，概率值小幅变化只能引起熵的微小变化。

对称性

符号xi重新排序后，该量度应不变。

$$\mathrm{H}_n\left(p_1, p_2, \ldots\right) = \mathrm{H}_n\left(p_2, p_1, \ldots\right)$$

极值性

当所有符号等可能出现的情况下，熵达到最大值（所有可能的事件等

$$\mathrm{H}_n(p_1, \ldots, p_n) \leq \mathrm{H}_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = \log_b(n)$$

可加性

熵的量与该过程如何被划分无关。一个系统可以划分为多个子系统

作用是已知的，则可以通过子系统的熵来计算一个系统的熵。

这一点很重要，后面的条件熵、互信息、联合熵都是利用这一性质

义。

二、自信息

名称：自信息、self-information

定义：是与概率空间中的单一事件或离散随机变量的值相关的信息量的量度。

作用：某个随机事件/变量 的混乱程度、信息量、不确定性。

物理意义：单个事件如果放在编码系统中，最优编码系统的字节长度，提供的

公式：-logP（x）

三、条件熵

名称：条件熵、conditional entropy

作用：在某个随机变量X情况下，随机变量Y的信息熵，反映了在给定X情况下，

说，在已知X的情况下，Y的熵还剩多少。

物理意义：是H（Y|X = x）对随机变量x的期望（yes or no？），是H（Y|X

权求和

公式：

等概率时不确定性最高）

，如果子系统之间的相互

，并且有自己的物理意

信息量

Y的不确定性。换句话

X = x）对随机变量x的加

公式：

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

延伸：

H（Y|X）=0 代表 X确定Y就确定了，H（Y|X）=H（Y）代表Y与X独立。

与联合熵的关系

$$\mathrm{H}(Y|X)\ =\ \mathrm{H}(X,Y)-\mathrm{H}(X)\,.$$

四、联合熵

作用：是多个随机变量不确定性的衡量手段。比如两个随机变量x，y的联合熵

不确定性

物理意义：就是随机变量（x，y）的熵，使用（x，y）的联合概率分布P（x，

公式：

$$\mathrm{H}(X,Y) = -\sum_{x}\sum_{y} P(x,y)\log_2[P(x,y)]$$

$$\mathrm{H}(X_1,...,X_n) = -\sum_{x1}...\sum_{x_n} P(x_1,...,x_n)\log_2[P(x_1,...,x_n)]$$

五、互信息

名称：互信息、Mutual information、转移信息、transinformation

含义：两组随机变量之间的依赖程度、互依赖性的度量。反映了随机变量的联

缘概率“P（x），P（y）的相似性程度。

物理意义：它度量知道这两个变量其中一个，对另一个不确定度（熵）减少的

（PMI）的期望值。

物理意义的公式：

$$I(X;Y) = H(X) - H(X|Y)$$

公式：

$$\left(\frac{p(x,y)}{\ }\right)$$

表示x，y事件同时发生的

y）

合概率P（x，y）与”边

程度。是点互信息

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right),$$

解释：

可以理解为知道一组随机变量Y之后，对估计另一组随机变量X的帮助（其
小的程度）。

这里可以先看一下六、中的图

例如：

在分类任务中，这里X是待预测的label，Y是样本的一维特征。

在不知道任何样本信息时，对X的估计仅能通过X的先验概率，这时X的不
加入一维特征Y之后，我们相当于有了观测数据的信息，知道这一维特征Y
到纠正，我们得到X的后验概率，X的不确定性也随之减小，成为H（X|Y
述过程也就是贝叶斯决策过程。

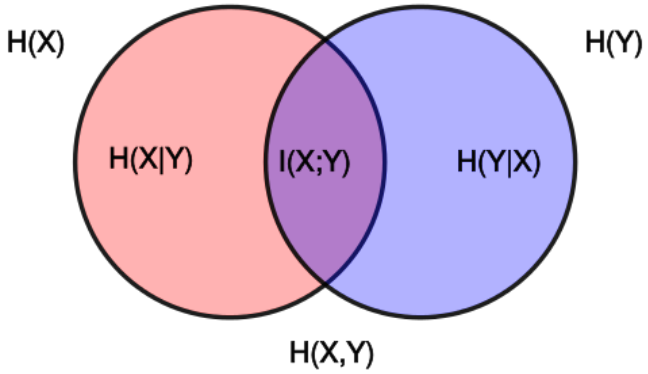
上述过程中，Y这一维特征到底对估计X起了多大作用？我们用互信息表示
定性减少的程度，"，也就是对熵减小的程度。即，I（X；Y）=H（X）-

延伸：

互信息具有对称性，即I（X；Y）=I（Y;X）

互信息可以衡量两组随机变量之间的依赖程度，但不完全同于相关性系数
随机变量。它更加一般且决定着联合分布 p(X,Y) 和分解的边缘分布的乘积

六、信息熵、条件熵、联合熵、互信息之间关系：



从图上可以轻易的看出：

$$I(X;Y) = H(X) - H(X|Y)$$

其实也就是对X不确定性减

确定性为H（X）。当我们
Y情况下，X的先验概率得
）（这就是条件熵）。上

示，互信息就是"Y对X不确
H（X|Y）

数，互信息并不局限于实值
积 p(X)p(Y) 的相似程度。

$$= H(Y) - H(Y|X)$$
$$= H(X) + H(Y) - H(X,Y)$$
$$= H(X,Y) - H(X|Y) - H(Y|X)$$
$$\mathrm{H}(X|Y) = \mathrm{H}(X,Y) - \mathrm{H}(Y)$$
$$H(Y|X) = H(X|Y) - H(X) + H(Y).$$

等等

七、相对熵（KL散度）

名称：相对熵（relative entropy）、KL散度（Kullback–Leibler divergence（information divergence）、信息增益（information gain）。

作用：两个概率分布之间的差别（可以看做是距离）的一种度量。

实际意义：D_KL（P||Q）指用基于随机变量Q的编码来编码随机变量P平均所需数）。经典情况下，P表示数据真实分布，Q表示理论/模型推测的分布。

公式：

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}.$$

特性：是一种非对称的距离衡量D_KL（P||Q）!= D_KL（Q||P）

D_KL（P||Q）>= 0 当 P = Q时 D_KL（P||Q）= 0

D_KL（P||Q）越大说明两个概率分布差别越大

延伸：

自信息（en:self-information）和KL散度

$$I(m) = D_{\mathrm{KL}}(\delta_{im}\|\{p_i\}),$$

互信息（en:Mutual information）和KL散度

$$\begin{aligned} I(X;Y) &= D_{\mathrm{KL}}(P(X,Y)\|P(X)P(Y)) \\ &= \mathbb{E}_X\{D_{\mathrm{KL}}(P(Y|X)\|P(Y))\} \\ &= \mathbb{E}_Y\{D_{\mathrm{KL}}(P(X|Y)\|P(X))\} \end{aligned}$$

信息熵（en: Shannon entropy）和KL散度

$$\begin{aligned} H(X) &= (\mathrm{i})\, \mathbb{E}_x\{I(x)\} \\ &= (\mathrm{ii})\, \log N - D_{\mathrm{KL}}(P(X)\|P_U(X)) \end{aligned}$$

条件熵（en:conditional entropy）和KL散度

，简称KLD）、信息散度

需要的额外位元数（bit

$$
\begin{aligned}
H(X|Y) &= \log N - D_{\mathrm{KL}}(P(X,Y)\|P_U(X)P(Y)) \\
&= \text{(i)}\ \log N - D_{\mathrm{KL}}(P(X,Y)\|P(X)P(Y)) - D_{\mathrm{KL}}(P(X)\|P_U(X)) \\
&= H(X) - I(X;Y) \\
&= \text{(ii)}\ \log N - \mathbb{E}_Y\{D_{\mathrm{KL}}(P(X|Y)\|P_U(X))\}
\end{aligned}
$$

交叉熵（en:cross entropy）和KL散度

$$\mathrm{H}(p,q) = \mathrm{E}_p[-\log q] = \mathrm{H}(p) + D_{\mathrm{KL}}(p\|q).$$

互信息也可以表示为两个随机变量的边缘分布 X 和 Y 的乘积 p(x) × p(y) 相对于
熵 p(x,y) 的相对熵。

八、交叉熵

名称：交叉熵、 cross entropy

作用：衡量模型预测的label与真实label的差距，可以作为机器学习模型中的损

物理意义：真实label p的熵，加上真实label分布p与预测的label分布q的差距
预测的分布q表示出真实的labelp所需要的额外字节数/信息量。

$$H(p,q) = \mathrm{E}_p[-\log q] = H(p) + D_{\mathrm{KL}}(p\|q),$$

也就是

公式：

$$C = -\frac{1}{n}\sum_x [y \ln a + (1-y)\ln(1-a)]$$

化简之后是这样，a表示预测的输出值，y表示标记的真实输出值。

解释：

详见

deep learning in NLP

1、如果使用sigmoid为模型输出前的非线性函数，使用corss entropy比均

2、相比均方误差，交叉熵对初始条件（模型初始参数）的要求不高，即使

3、分类问题中，极大似然法求解得到的损失函数就是corss entropy（我猜
其他分类不确定）

4、主要用于分类问题，而非回归问题。

于随机变量的联合

损失函数。

，这个“差距”也就是从

均方误差好很多

使很差也能收敛

准到的二分类是这样的，

九、参考资料

信息学熵 https://zh.wikipedia.org/wiki/Category:%E4%BF%A1%E6%81%AF%E5%A

信息熵的wiki写的挺好 https://zh.wikipedia.org/wiki/%E7%86%B5_(%E4%BF%A1%

自信息 https://zh.wikipedia.org/wiki/%E8%87%AA%E4%BF%A1%E6%81%AF

条件熵 https://zh.wikipedia.org/wiki/%E6%9D%A1%E4%BB%B6%E7%86%B5

互信息 https://zh.wikipedia.org/wiki/%E4%BA%92%E4%BF%A1%E6%81%AF

联合熵 图好 https://zh.wikipedia.org/wiki/%E8%81%94%E5%90%88%E7%86%B5

相对熵 https://zh.wikipedia.org/wiki/%E7%9B%B8%E5%AF%B9%E7%86%B5

D%B8%E7%86%B5

E6%81%AF%E8%AE%BA)