数据科学基础
Foundations of Data Science

6.2 信息熵计算

陈振宇
南京大学智能软件工程实验室
www.iselab.cn

联合熵（Joint Entropy）

联合熵是一个集合中的变量之间不确定性的衡量手段。

定义：对于两个离散的随机变量 X 和 Y，联合熵的定义为：

$$H(X,Y) = -\sum_{x}\sum_{y} p(x,y)\log\ [p(x,y)]$$

其中x和y是X和Y的特定值，相应地， $p(x,y)$是这些值一起出现的联合概率。

若$p(x,y)=0$，则定义$p(x,y)\log\ [p(\mathrm{x},\mathrm{y})]$的值为0.

联合熵（Joint Entropy）

一般形式：

对于两个以上的随机变量$X_1, \ldots, X_n$，联合熵的定义为：

$$H(X_1, \ldots, X_n) = -\sum_{x_1} \ldots \sum_{x_n} \mathrm{P}(x_1, \ldots, x_n) \log [P(x_1, \ldots, x_n)]$$

其中$x_1, \ldots, x_n$是$X_1, \ldots, X_n$的特定值，相应地，$\mathrm{P}(x_1, \ldots, x_n)$是这些值一起出现的联合概率。

若$\mathrm{P}(x_1, \ldots, x_n)$，则定义$\mathrm{P}(x, y) \log [\mathrm{P}(x, y)]$的值为0.

联合熵性质

联合熵**不小于**每个独立的熵

在一个集合中，所有变量的联合熵大于或等于这个集合中任意一个变量的独立熵。

$$H(X,Y) \geq \max[H(X),H(Y)]$$

$$H(X_1,\dots,X_n) \geq \max[H(X_1),\dots,H(X_n)]$$

X|Y	0	1	X
0	¼	¼	½
1	½	0	½
Y	¾	¼	

$H(X,\mathrm{Y})$ =-(1/2log(1/2)+2*1/4log(1/4))=1.5

$H(\mathrm{Y})$ =-(3/4log(3/4)+1/4log(1/4))<1.5

$H(\mathrm{X})$ =-(1/2log(1/2)+1/2log(1/2))=1<1.5

联合熵性质

联合熵**不大于**每个独立熵的和

在一个集合中，所有变量的联合熵小于或等于这个集合中所有变量的独立熵之和。

$$H(X,Y) \leq H(X) + H(Y)$$

$$H(X_1, \dots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

条件熵（Conditional Entropy）

条件熵量化了在已知随机变量X的条件下, 描述随机变量Y所需的信息量。

如果 $H(Y|X=x)$ 为随机变量Y取特定值X条件下的熵，那么$H(Y|X)$就是$H(Y|\mathrm{X}=x)$ 在X取遍所有可能的x后取平均的结果。给定随机变量X与Y,

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

条件熵性质

条件熵等于零

$H(Y|X) = 0$ 当且仅当Y的值完全取决于X的值。

例如，X服从伯努利分布，Y=1-X。

相互独立的随机变量的条件熵

$H(Y|X) = H(X)$ 当且仅当Y和X相互独立。

条件熵性质

链式法则

$\boldsymbol{H(X,Y)=H(X)+H(Y|X)}$

证明:

$$H(X,Y)=-\sum_{x}\sum_{y}p(x,y)\log p(x,y)$$

$$=-\sum_{x}\sum_{y}p(x,y)\log p(x)-\sum_{x}\sum_{y}p(x,y)\log p(y|x)$$

$$=-\sum_{x}p(x)\log p(x)-\sum_{x}\sum_{y}p(x,y)\log p(y|x)$$

$$=\boldsymbol{H(X)+H(Y|X)}$$

条件熵性质

贝叶斯规则

$H(Y|X) = H(X|Y) - H(X) + H(Y)$

证明:

(1) $H(Y|X) = H(X,Y) - H(X)$

(2) $H(X|Y) = H(Y,X) - H(Y)$

$(1) - (2)$:

$H(Y|X) - H(X|Y) = H(X,Y) - H(X) - [H(Y,X) - H(Y)]$

$H(Y|X) - H(X|Y) = -H(X) + H(Y)$

互信息(Mutual Information)

定义： 在概率论和信息论中，两个随机变量的互信息（Mutual Information，简称MI）是变量间相互依赖性的量度。

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

其中 $p(x,y)$是X和Y的联合概率分布函数，而 $p(y)$ 分别是 X 和 Y 的边缘概率分布函数。

互信息与熵的关系

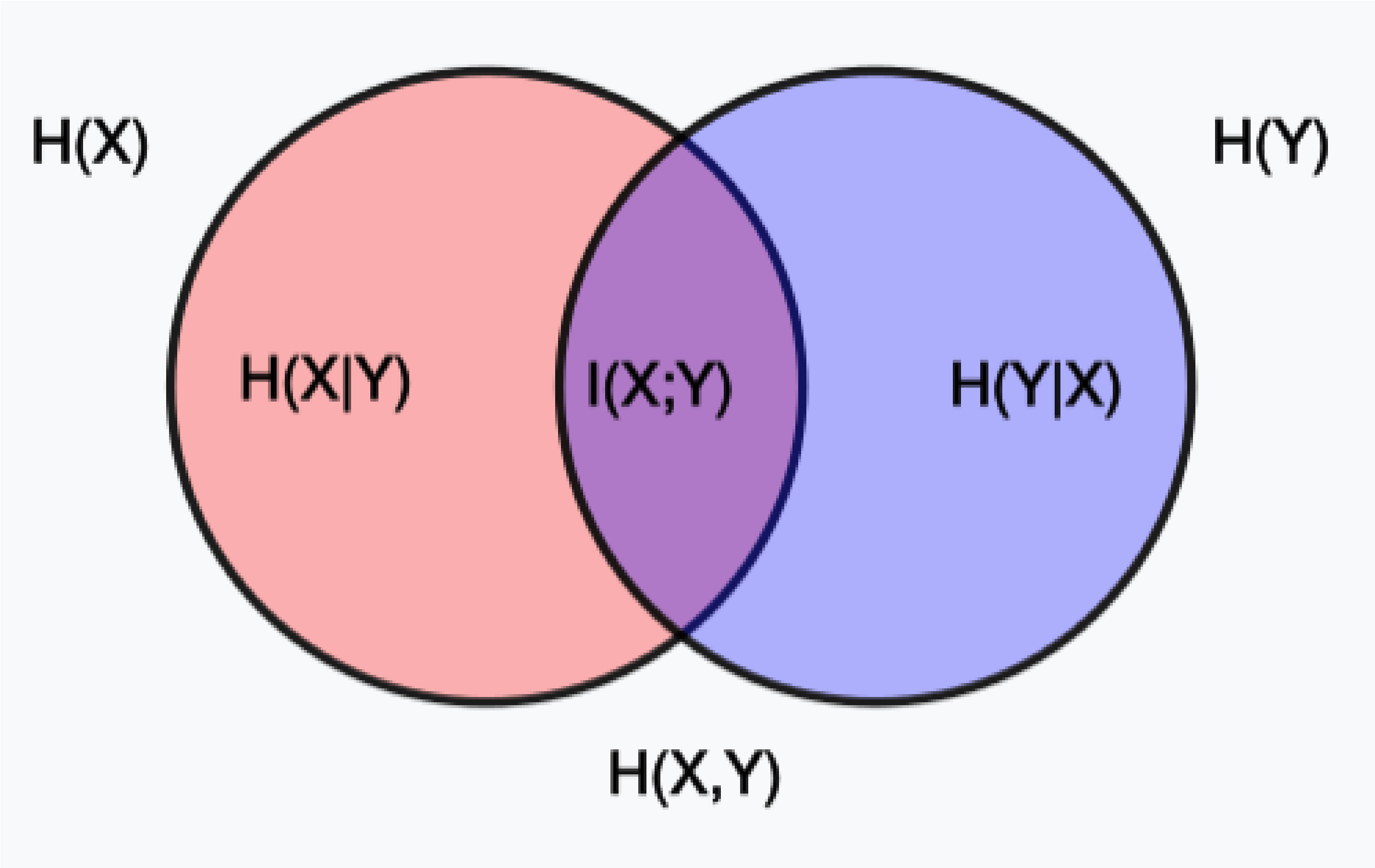
$$I(X;Y) = H(X) - H(X|Y)$$

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

$$= H(X) + H(Y) - H(X,Y)$$

$$= H(X,Y) - H(X|Y) - H(Y|X)$$

互信息与熵的关系



相对熵（Relative Entropy）

相对熵（relative entropy）又称为KL散度(Kullback–Leibler divergence)。KL散度是两个概率分布P和Q差别的非对称性的度量。 P为真实分布，Q为假设分布的“距离” （但相对熵不满足对称性和三角不等式）。

$$D_{KL}(P||Q) = -\sum P(x)\log\frac{Q(x)}{P(x)}$$

$$D_{KL}(P||Q) = \sum P(x)\log\frac{P(x)}{Q(x)}$$

相对熵性质

非负性

$D_{KL}(P||Q) \geq 0$, $D_{KL}(P||Q) = 0$ 当且仅当$P = Q$

不对称性

$D_{KL}(P||Q) \neq D_{KL}(Q||P)$

一种方案:

$Let\ D(P,Q) = [D_{KL}(P||Q) + D_{KL}(Q||P)]/2$

交叉熵（Cross Entropy）

基于相同事件测度的两个概率分布 p和 q的交叉熵是指，当基于一个“非真实”（相对于真实分布 p）的概率分布 q进行编码时，在事件集合中唯一标识一事件所需要信息量。

$$H(p,q) = -\sum p(x)\log q(x)$$

交叉熵与相对熵

$$H(p,q) = H(p) + D_{KL}(P||Q)$$