信息理论

第一部分：信息的度量

余官定　教授

浙江大学
信息与电子工程学院

第一讲：随机变量的熵和互信息

- 掌握随机事件的自信息，互信息的概念及物理意义
- 了解条件事件的互信息与联合事件的互信息
- 掌握随机变量的熵的概念以及物理意义
- 了解随机变量的条件熵和联合熵及其性质
- 掌握随机变量互信息定义以及互信息的性质

概率论基础

随机变量的概率空间 $\{X, \mathcal{X}, q(x)\}$

- $\mathcal{X}$: X的取值空间, $\mathcal{X} = \{x_k; k = 1, 2, \cdots, K\}$
- $q(x)$: 事件$\{X = x\}$发生的概率, $q(x) \geq 0, \sum_{x \in \mathcal{X}} q(x) = 1$

联合变量对 (X, Y)

二维随机变量 $\{(X, Y), \mathcal{X} \times \mathcal{Y}, p(x, y)\}$

- $p(x, y) = P\{X = x, Y = y\}$
- $\mathcal{X} = \{x_k; k = 1, 2, \cdots, K\}$, $\mathcal{Y} = \{y_j; j = 1, 2, \cdots, J\}$

概率论基础

- $p(x_k, y_j) \geq 0$
- $\sum_k \sum_j p(x_k, y_j) = 1$
- $\sum_k p(x_k, y_j) = \omega(y_j)$
- $\sum_j p(x_k, y_j) = q(x_k)$

条件概率

$$p(y_j|x_k) = p(Y = y_j|X = x_k) = \frac{p(x_k, y_j)}{q(x_k)}$$

$$p(x_k|y_j) = p(X = x_k|Y = y_j) = \frac{p(x_k, y_j)}{\omega(y_j)}$$

事件的自信息

信息量是信息论的重要概念，事件的信息量基于该事件发生的概率。

定义：对于概率空间$\{X, \mathcal{X}, q(x)\}$, 事件$\{X = x_k\}$ 的自信息定义为

$$I(x_k) = -\log_a q(x_k)$$

单位：当$a = 2$时，为比特(bit), 当$a = e$时, 为奈特(nat)。

定义为概率的负对数的优点：

1. 符合概率越小，信息量越大的要求。
2. 对数函数是比较简单的函数，容易进行数学处理。
3. 对数函数的可加性符合生活中关于信息可叠加性的经验。

事件的自信息

事件自信息的本质

1. 事件发生后对外界（观察者）所提供的信息量。
2. 事件发生前外界（观察者）为确证该事件的发生所需要的信息量，也是外界为确证该事件所需要付出的代价。
3. 事件的自信息并不代表事件的不确定性，事件本身没有不确定性可言，它要么是观察的假设和前提，要么是观察的结果。

事件的自信息

事件自信息的性质

性质1：$q(x_k)$越大，$I(x_k)$越小，概率越小的事件其自信息越大。
举例：某一个城市的天气经常下雨，即下雨天的概率比较大，那么当该城市气象台预报第二天将要降雨的时候，此时信息量比较小；相反，如果预报第二天是个晴天的时候，此时这个信息的信息量是比较大的。

性质2：$q(x_k) = 1$，$I(x_k) = 0$，确定事件的自信息为0。
举例：“明天的最低气温低于30度”，这句话的信息量几乎为0，因为该事件发生的概率基本为1。

性质3：$q(x_k) \to 0$，$I(x_k) = \infty$。
举例：“张三买彩票中了500万”，这句话的信息量是巨大的，因为中彩票头奖的概率基本上趋向于0。

事件的条件自信息

事件条件自信息的定义

二维随机变量$\{(X,Y),\mathcal{X}\times\mathcal{Y},p(x,y)\}$

事件$\{Y=y_j\}$发生的条件下事件$\{X=x_k\}$的条件自信息定义为：

$$I(x_k|y_j)=-\log p(x_k|y_j)$$

事件条件自信息的本质

1. 事件$\{Y=y_j\}$发生后，$\{X=x_k\}$如果再发生需要的“新”的信息量。
2. 事件$\{Y=y_j\}$发生后，如果$\{X=x_k\}$ 又发生了，则提供给观察者的“新”的信息量。

事件的条件自信息

例子1：x_k: 杭州下雨，y_j：上海下雨。

$I(x_k)$:杭州下雨需要的信息量，$I(x_k|y_j)$:上海下雨后杭州下雨需要的信息量。

$q(x_k)=0.5$；$p(x_k|y_j)=0.75$，则$I(x_k)=1\text{bit}$，$I(x_k|y_j)=\log_2\left(\frac{4}{3}\right)\text{bit}$，$I(x_k)>I(x_k|y_j)$

例子2：x_k: 杭州下雨，y_j：上海晴天。

$I(x_k)$:杭州下雨需要的信息量，$I(x_k|y_j)$:上海晴天时杭州下雨需要的信息量。

$q(x_k)=0.5$；$p(x_k|y_j)=0.25$，则$I(x_k)=1\text{bit}$，$I(x_k|y_j)=2\text{bit}$，$I(x_k)<I(x_k|y_j)$

例子3：（无关事件）x_k: 杭州下雨，y_j：北京下雨。

$$I(x_k)=I(x_k|y_j)$$

事件的互信息

二维随机变量$\{(X,Y),\mathcal{X}\times\mathcal{Y},p(x,y)\}$, 事件$\{Y=y_j\}$与事件$\{X=x_k\}$ 之间的互信息定义为：

$$I(x_k;y_j)=I(x_k)-I(x_k|y_j)=-\log q(x_k)-\{-\log p(x_k|y_j)\}$$

事件互信息的本质

事件$\{Y=y_j\}$发生后对事件$\{X=x_k\}$不确定性的降低。

事件互信息的性质

对称性 $I(x_k;y_j)=I(y_j;x_k)$

$$\log\frac{p(x_k|y_j)}{q(x_k)}=\log\frac{p(x_k,y_j)}{q(x_k)\omega(y_j)}=\log\frac{p(y_j|x_k)}{\omega(y_j)}$$

事件互信息的例子

例子1：x_k: 杭州下雨，y_j：上海下雨。

$I(x_k)$:杭州下雨需要的信息量，$I(x_k|y_j)$:上海下雨后杭州下雨需要的信息量。

$q(x_k)=0.5$；$p(x_k|y_j)=0.75$，则$I(x_k)=1\text{bit}$，$I(x_k|y_j)=\log_2\left(\frac{4}{3}\right)\text{bit}$，$I(x_k;y_j)>0$

例子2：x_k: 杭州下雨，y_j：上海晴天。

$I(x_k)$:杭州下雨需要的信息量，$I(x_k|y_j)$:上海晴天时杭州下雨需要的信息量。

$q(x_k)=0.5$；$p(x_k|y_j)=0.25$，则$I(x_k)=1\text{bit}$，$I(x_k|y_j)=2\text{bit}$，$I(x_k;y_j)<0$

例子3：（无关事件）x_k: 杭州下雨，y_j：北京下雨。

$$I(x_k)=I(x_k|y_j),\ I(x_k;y_j)=0$$

事件的联合自信息

事件联合自信息的定义

二维随机变量$\{(X,Y),\mathcal{X}\times\mathcal{Y},p(x,y)\}$

事件$\{Y=y_j\}$和$\{X=x_k\}$的联合自信息定义为：

$$I(x_k,y_j)=-\log p(x_k,y_j)$$

表示事件$\{X=x_k\}$和$\{Y=y_j\}$同时发生需要的信息量，或者同时发生后对外界提供的信息量。

例子1：x_k：杭州下雨，y_j：上海下雨。$I(x_k,y_j)$为杭州和上海同时下雨需要的信息量。

事件的条件互信息

在给定$Z=z$的条件下，事件$X=x$与$Y=y$之间的条件互信息为：

$$I(x;y|z)=\log\frac{p(x|y,z)}{p(x|z)}=\log\frac{p(x,y|z)}{p(x|z)\cdot p(y|z)}$$

表示事件$Z=z$发生时，事件$X=x$与$Y=y$相互之间提供的信息量。

例子：x：杭州下雨，y：上海下雨，z：宁波下雨。

- $q(x)=q(y)=q(z)=0.125$
- $p(x|y)=0.25,\ p(x|z)=0.25,\ p(y|z)=0.25$
- $p(x|y,z)=0.5$

则$I(x)=3\text{bit},\ I(x|y)=2\text{bit},\ I(x;y)=1\text{bit};\ I(x;y|z)=1\text{bit}$

事件的联合互信息

定义：联合事件$\{Y=y,\ Z=z\}$与事件$\{X=x\}$之间的互信息为：

$$I(x;y,z)=\log\frac{p(x|y,z)}{p(x)}=\log\frac{p(x,y,z)}{p(x)p(y,z)}$$

表示事件$\{Y=y,\ Z=z\}$联合提供给事件$\{X=x\}$ 的信息量

例子：x：杭州下雨，y：上海下雨，z：宁波下雨。

$I(x)$: 杭州下雨需要的信息量，$I(x;y,z)$：上海下雨和宁波下雨这两个事件同时提供给杭州下雨这个事件的信息量。

$q(x)=0.125; p(x|y)=0.25$, 则

$$I(x;y)=I(x)-I(x|y)=1$$

$p(x|y,z)=0.5$, 则

$$I(x;y,z)=I(x)-I(x|y,z)=2>I(x;y)$$

事件联合互信息的链式法则

$$I(x;y,z) = I(x;y) + I(x;z|y)$$

事件$\{Y=y, Z=z\}$联合提供给事件$\{X=x\}$的信息量，等于事件$\{Y=y\}$ 提供给事件$\{X=x\}$的信息量加上事件$\{Y=y\}$已知的条件下，事件$\{Z=z\}$ 提供给$\{X=x\}$ 的新信息量

证明：

$$\begin{aligned} I(x;y,z) &= \log \frac{p(x|y,z)}{p(x)} \\ &= \log \frac{p(x|y)p(x|y,z)}{p(x)p(x|y)} \\ &= \log \frac{p(x|y)}{p(x)} + \log \frac{p(x|y,z)}{p(x|y)} \\ &= I(x;y) + I(x;z|y) \end{aligned}$$

小结

1.事件的自信息　$I(x_k) = -\log q(x_k)$

2.事件的条件自信息　$I(x_k|y_j) = -\log p(x_k|y_j)$

3.事件的互信息　$I(x_k;y_j) = I(x_k) - I(x_k|y_j) = \log \dfrac{p(x_k|y_j)}{q(x_k)}$

4.事件的联合自信息　$I(x_k,y_j) = -\log p(x_k,y_j)$

5.事件的条件互信息

$$I(x;y|z) = \log \frac{p(x|y,z)}{p(x|z)} = \log \frac{p(x,y|z)}{p(x|z)p(y|z)}$$

6.事件的联合互信息

$$I(x;y,z) = I(x) - I(x|y,z) = \log \frac{p(x|y,z)}{p(x)} = \log \frac{p(x,y,z)}{p(x)p(y,z)}$$

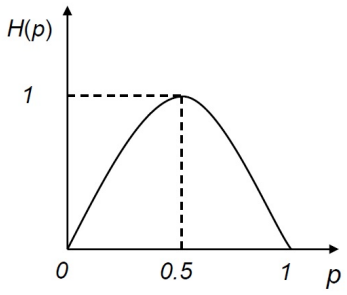
事件的自信息，条件自信息都大于**0**，而事件的互信息可小于、大于、等于**0**!

随机变量的熵

定义：随机变量的熵定义为随机变量各个事件的平均自信息：

$$H(X) = E[I(X)] = \sum_{x \in \mathcal{X}} q(x) I(x) = -\sum_{x \in \mathcal{X}} q(x) \log q(x)$$

熵与自信息的区别：熵针对的是随机变量，自信息针对具体的事件。



[例子]：二元随机变量X的概率分布$q(x_1) = p, q(x_2) = 1 - p$，则

$$H(X) = -p \log p - (1-p) \log(1-p)$$

$p = 0, 1; H(X) = 0$　确定性变量的熵为0。

$p = 0.5; H(X) = 1$　等概率变量的随机性最大，所以熵最大。

随机变量熵的物理意义

熵是随机变量不确定性的度量

例子：随机变量$X=1$表示明天莫斯科下雪，$X=0$表示莫斯科不下雪；$Y=1$表示北京下雪，$Y=0$表示北京不下雪；$Z=1$表示香港下雪，$Z=0$ 表示香港不下雪。

设$p(X=1)=0.8, p(X=0)=0.2; p(Y=1)=0.5, p(Y=0)=0.5; p(Z=1)=0.0001, p(Z=0)=0.9999$。

我们可以看到$H(Y)>H(X)>H(Z)$，即北京下不下雪很难说，莫斯科很有可能下雪，但是香港几乎不会下雪！

随机变量的联合熵

随机变量的联合熵

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

表示两个随机变量的不确定性度量

例子：$H(X,Y)$表示香港下雪和北京下雪两个变量的不确定度。

随机变量的条件熵

随机变量的条件熵

定义1：给定$Y=y$条件下，X的条件熵为

$$H(X|y)=\sum_{x\in\mathcal{X}}p(x|y)I(x|y)=-\sum_{x\in\mathcal{X}}p(x|y)\log p(x|y)$$

例子：随机变量$X=1$表示杭州下雨，$X=0$表示杭州不下雨；$Y=1$表示上海下雨，$Y=0$ 表示上海不下雨。

设$p(X=1)=0.5;p(X=0)=0.5;p(1|1)=0.75,p(0|1)=0.25$，则$H(X|Y=1)<H(X)$。

$p(X=1)=0.25;\ p(X=0)=0.75;\ p(1|1)=0.5,\ p(0|1)=0.5$，则$H(X|Y=1)>H(X)$。

$Q:H(X|y)$与$H(X)$的关系？

随机变量的条件熵

随机变量的条件熵

定义2：随机变量X相对于随机变量Y的条件熵为

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

[性质]：X和Y统计独立时，$H(X|Y) = H(X)$

随机变量联合熵的链式法则

定理2.1.1：随机变量X和Y的联合熵的链式法则

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

$$\begin{aligned} H(X,Y) &= E[I(X,Y)] = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\ &= H(X) + H(Y|X) \end{aligned}$$

性质：X和Y统计独立时，$H(X,Y) = H(X) + H(Y)$

$$H(X,Y,Z) = H(X) + H(Y,Z|X) = H(X) + H(Y|X) + H(Z|X,Y)$$

熵的性质

$$\mathbf{X} \sim \begin{pmatrix} x_1 & x_2 & \dots & x_K \\ p_1 & p_2 & \dots & p_K \end{pmatrix}$$

$$H(X) \triangleq H_K(p_1, p_2, \cdots, p_K) \triangleq H_K(P) = -\sum_{k=1}^{K} p_k \log p_k$$

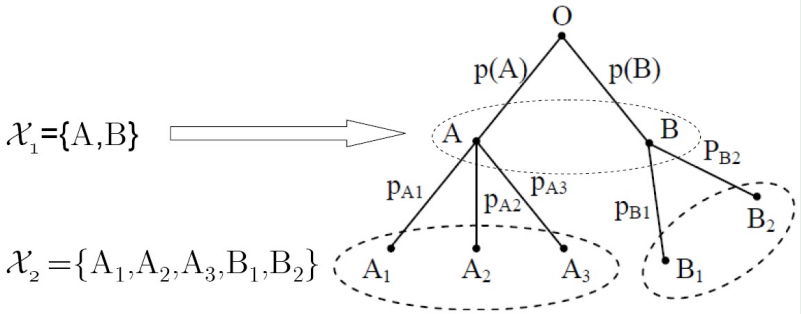
熵的性质

$$H(X) \triangleq H_K(p_1, p_2, \cdots, p_K) \triangleq H_K(\mathbf{P})$$
$$= -\sum_{k=1}^{K} p_k \log p_k$$

1. $H_K(\mathbf{P})$对概率矢量$\mathbf{P}$的分量是对称的。
2. 非负性，即$H_K(\mathbf{P}) \geq 0$。
3. 确定性，即若$\mathbf{P} = (p_1, p_2, \cdots, p_K)$中有一个分量为1，其余均为零，则$H_K(\mathbf{P}) = 0$。
4. 可扩展性，即$\lim_{\epsilon \to 0} H_{K+1}(p_1, p_2, \cdots, p_K - \epsilon, \epsilon) = H_K(p_1, p_2, \cdots, p_K)$

熵的性质

可加性



$$H(X_2)\big|_{X_2\in\mathcal{X}_2} = H(X_1)\big|_{X_1\in\mathcal{X}_1} + H(X_2\mid X_1)\Big|_{X_2\in\mathcal{X}_2}^{X_1\in\mathcal{X}_1}$$

熵的性质

$$
\begin{aligned}
H(X_1)\mid_{X_1\in\mathcal{X}_1} &= H(P(A),P(B))\\
H(X_2)\mid_{X_2\in\mathcal{X}_2} &= H(P(A)P_{A_1},P(A)P_{A_2},P(A)P_{A_3},P(B)P_{B_1},P(B)P_{B_2})\\
H(X_2\mid X_1)\mid_{X_2\in\mathcal{X}_2}^{X_1\in\mathcal{X}_1} &= \sum_{x_1\in\mathcal{X}_1} P(x_1)H(X_2\mid x_1)\\
&= P(A)H(P_{A_1},P_{A_2},P_{A_3})+P(B)H(P_{B_1},P_{B_2})
\end{aligned}
$$

对变量X可以进行多步分层的观察，每一步都可从上一步的观察结果中得到更为细致的结果，变量X在最后的观察结果集合中的不确定性等于第一次观察结果的不确定性，加上其后每次观察结果在前一次观察结果已知的前提下的条件不确定性。

熵的性质

$$
\begin{aligned}
H_M &= -\sum_{k=1}^{K}\sum_{j=1}^{m_k} p_k q_{jk} \log p_k q_{jk} \\
&= -\sum_{k=1}^{K}\left(\sum_{j=1}^{m_k} q_{jk}\right) p_k \log p_k - \sum_{k=1}^{K}\sum_{j=1}^{m_k} p_k q_{jk} \log q_{jk} \\
&= H_K(p_1, p_2, \cdots, p_K) + \sum_{k=1}^{K} p_k H_{m_k}(q_{jk})
\end{aligned}
$$

$$
\begin{aligned}
H(X_1, X_2) &= H(X_2) + H(X_1 \mid X_2) = H(X_2) \\
H(X_1, X_2) &= H(X_1) + H(X_2 \mid X_1)
\end{aligned}
$$

熵的性质

极值性

$$H_K(p_1, p_2, \cdots, p_K) \leq H_K\left(\frac{1}{K}, \frac{1}{K}, \cdots, \frac{1}{K}\right) = \log K$$

证明：

$H_K(p_1, p_2, \cdots, p_K) \leq -\sum_{k=1}^{K} p_k \log q_k$ 对任何概率矢量$\mathbf{q}$均成立。因为：

$H_K(p_1, p_2, \cdots, p_K) + \sum_{k=1}^{K} p_k \log q_k$

$= \sum_{k=1}^{K} p_k \log \frac{q_k}{p_k} \leq \log e \cdot \sum_{k=1}^{K} p_k \left(\frac{q_k}{p_k} - 1\right)$,因$(\ln x \leq x - 1)$

$= 0$ 令$q_k = \frac{1}{K}, k = 1, 2, \cdots, K$即得。

熵的性质

条件熵≤熵：增加条件使熵减少

$$H(p_1, p_2, \cdots, p_K) \leq -\sum_{k=1}^{K} p_k \log q_k$$

$$\begin{aligned} H(X|Y) &= E\{H(X|y)\} = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x|y) \\ &= -\sum_{y\in\mathcal{Y}} \omega(y)\left\{\sum_{x\in\mathcal{X}} p(x|y)\log p(x|y)\right\} \\ &\leq -\sum_{y\in\mathcal{Y}} \omega(y)\left\{\sum_{x\in\mathcal{X}} p(x|y)\log q(x)\right\} \\ &= -\sum_{x\in\mathcal{X}} q(x)\log q(x) = H(X) \end{aligned}$$

熵的性质

凸性

$H_k(P)$ 是$P = (p_1, p_2, \cdots, p_k)$ 的严格上凸（∩,凹,Concave)函数，即对任何θ, $0 < \theta < 1$, 和任何二个K维概率矢量$P_1, P_2, P_1 \neq P_2$, 有

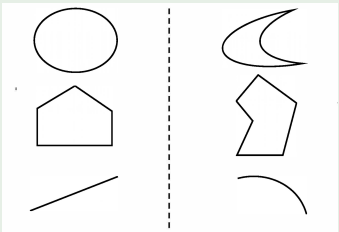
$$H_k(\theta P_1 + (1-\theta)P_2) > \theta H_k(P_1) + (1-\theta)H_k(P_2)$$

凸集与凸函数

令$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_k);\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_k)$是$k$维实矢量空间集合$R$中的任何两个矢量。如果对任何实数$\theta(0\leq\theta\leq1)$，有

$$\theta\boldsymbol{\alpha}+(1-\theta)\boldsymbol{\beta}\in R$$

则称R是凸集合。



概率矢量组成的集合称为概率空间，概率空间是一个凸集合。

凸集与凸函数

定义在凸集合R上的实值矢量函数f被称为上凸函数，当且仅当对任何两个矢量$\boldsymbol{\alpha}, \boldsymbol{\beta}$以及实数$\theta(0 \leq \theta \leq 1)$

$$\theta f(\boldsymbol{\alpha}) + (1-\theta) f(\boldsymbol{\beta}) \leq f\left[\theta \boldsymbol{\alpha} + (1-\theta)\boldsymbol{\beta}\right]$$

如果不等式中符号翻转，则称该函数为下凸函数。

凸集与凸函数

凸函数的性质

- 如果$f_1(\boldsymbol{\alpha}), f_2(\boldsymbol{\alpha}), \cdots, f_L(\boldsymbol{\alpha})$是上凸函数，以及$C_1, C_2, \cdots, C_L$是任意正数，则

$$\sum_l C_l f_l(\boldsymbol{\alpha})$$

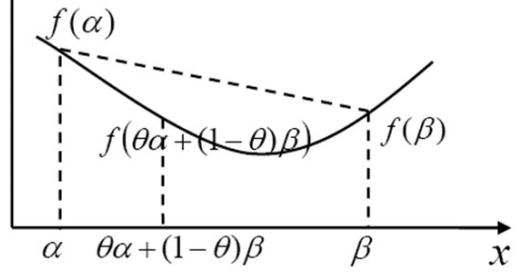
  也是上凸函数

- 一元函数$f(\alpha)$上凸的充要条件是在所定义的区间中满足

$$\frac{d^2 f(\alpha)}{d\alpha^2} \leq 0$$

凸函数（下凸函数，Convex Function）

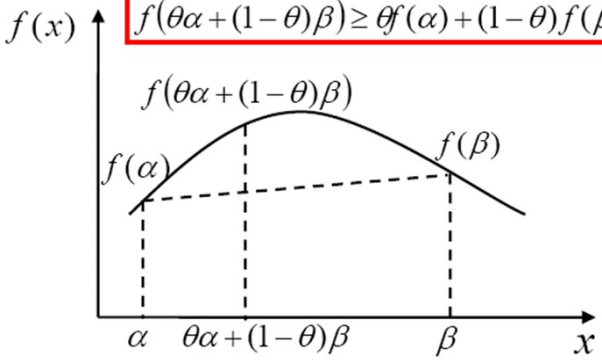
$$f(\theta\alpha+(1-\theta)\beta)\le \theta f(\alpha)+(1-\theta)f(\beta),\ \forall\theta\in[0,1]$$



$$\left.\frac{\partial f(x)}{\partial x}\right|_{x=x^*}=0,\ \text{if } f(x)\text{在}x=x^*\text{处取极小值}$$

凹函数（上凸函数，Concave Function）

$$f\left(\theta\alpha+(1-\theta)\beta\right)\geq\theta f(\alpha)+(1-\theta)f(\beta),\ \forall\theta\in[0,1]$$



$$\left.\frac{\partial f(x)}{\partial x}\right|_{x=x^*}=0,\ \text{if } f(x)\text{在}x=x^*\text{处取极大值}$$

凸函数的性质

- Jensen不等式

  令$(\alpha_1, \alpha_2, \cdots, \alpha_L)$是凸集中的一组矢量，$f(\boldsymbol{\alpha})$是该凸集上的一个上凸函数，$(\theta_1, \theta_2, \cdots, \theta_L)$是一组概率分布，则

$$\sum_{l=1}^{L} \theta_l f(\alpha_l) \leq f\left[\sum_{l=1}^{L} \theta_l \alpha_l\right]$$

凸集与凸函数

非负凸集上的上凸函数取极大值的充要条件

$f(\alpha)$是定义在K维非负凸集$\{\mathcal{R}^+\}^K$上的上凸函数，若$f(\alpha)$对于任一分量连续可导，则$f(\alpha)$在$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_K)$处取得极大值的充要条件是

$$\frac{\partial f(\alpha)}{\partial \alpha_k} \begin{cases} = 0, & \forall \alpha_k > 0 \\ < 0, & \forall \alpha_k = 0 \end{cases}$$

凸集与凸函数

概率空间上的上凸函数取极大值的充要条件

$f(\alpha)$是定义在K维概率空间上的上凸函数，若$f(\alpha)$对于任一分量连续可导，则$f(\alpha)$在$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_K)$ 处取得极大值的充要条件是

$$\frac{\partial f(\alpha)}{\partial \alpha_k} \begin{cases} = \lambda, & \forall \alpha_k > 0 \\ < \lambda, & \forall \alpha_k = 0 \end{cases}$$

随机变量的加权熵

随机变量的加权熵

香农熵仅考虑事件发生的客观概率，无法描述主观意义上对事件判断的差别

$$\mathbf{X} \sim \begin{pmatrix} x_1 & x_2 & \dots & x_K \\ w_1 & w_2 & \dots & w_K \\ p_1 & p_2 & \dots & p_K \end{pmatrix}$$

$$H_W(X) = -\sum_{k=1}^{K} w_k p_k \log p_k$$

随机变量的Rényi熵

随机变量的Rényi熵

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left(\sum_{k=1}^{K} p_k^\alpha \right)$$

当$\alpha = 0$

$$H_0(X) = \log K$$

当$\alpha \to 1$

$$H_1(X) = -\sum_{k=1}^{K} p_k \log p_k$$

小结

1. 随机变量的熵$H(X)$
2. 随机变量的条件熵$H(X \mid Y)$
3. 随机变量的联合熵$H(X,Y)$ 联合熵=熵+条件熵
4. 随机变量熵的性质

随机变量间的平均互信息

二个事件$X = x$与$Y = y$之间相互提供的信息量定义为：

$$I(x,y) = I(x) - I(x|y) = \log \frac{p(x|y)}{q(x)}$$

$I(x,y)$可正，可负，可为0。

随机变量$\{(X,Y), \mathcal{X} \times \mathcal{Y}, p(x,y)\}$相互提供的平均信息量称之为二者之间的平均互信息，简称互信息：

$$I(X;Y) = E\{I(x;y)\} = \sum_x \sum_y p(x,y) \log \frac{p(x|y)}{q(x)}$$

互信息的性质

非负性

$$I(X;Y) \geq 0$$

证明：

$$\begin{aligned} I(X;Y) &= \sum_x \sum_y p(x,y) \log \frac{p(x|y)}{q(x)} \\ &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x) \cdot \omega(y)} \\ &\geq -\sum_x \sum_y p(x,y) \left\{ \frac{q(x)\omega(y)}{p(x,y)} - 1 \right\} \\ &= 0 \end{aligned}$$

互信息的性质

对称性

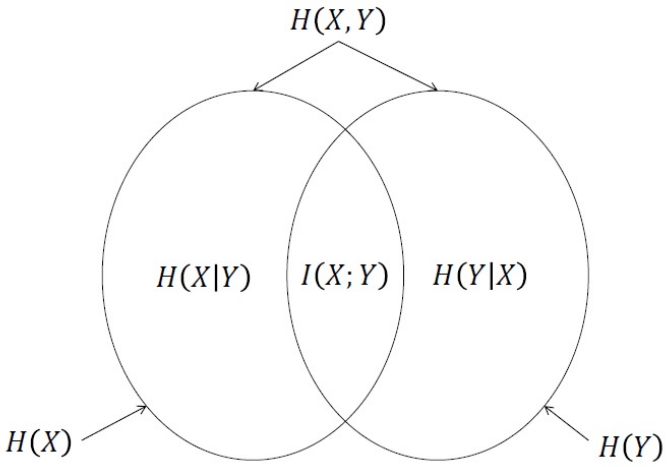
$$I(X;Y) = I(Y;X)$$

互信息与熵的关联性

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

互信息$I(X;Y)$等于X总的不确定性，减去Y已知以后，X 中还留存的剩余不确定性。从而互信息代表了Y提供给X的信息量

互信息与熵的关联性



互信息的性质

互信息与熵的大小关系

$$I(X;Y) \leq H(X)$$

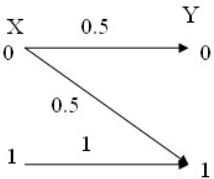
等号成立的条件是X是Y的确定性函数；

$$I(X;Y) \leq H(Y)$$

等号成立的条件是Y是X的确定性函数。

互信息的例子

$X=0$表示人健康，$X=1$表示人有病。$Y=0$表示化验结果阴性，$Y=1$表示阳性。问这项化验对于查明病情提供多少信息？



若$p(X=0)=0.98, p(X=1)=0.02$:

$$H(X)=0.1414,\ p(Y=0)=0.49,\ p(Y=1)=0.51$$

$$p(X=0|Y=0)=1, p(X=1|Y=0)=0$$

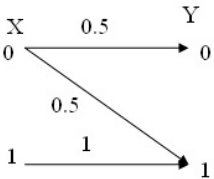
$$p(X=0|Y=1)=0.96, p(X=1|Y=1)=0.04$$

$$H(X|Y)=0.51\times H(0.04,0.96)=0.1236$$

$$I(X;Y)=H(X)-H(X|Y)=0.0178$$

互信息的例子

$X=0$表示人健康，$X=1$表示人有病。$Y=0$表示化验结果阴性，$Y=1$表示阳性。问这项化验对于查明病情提供多少信息？



若$p(X=0)=0.5, p(X=1)=0.5$:

$$H(X)=1\text{bit}, p(Y=0)=0.25, p(Y=1)=0.75$$

$$p(X=0|Y=0)=1, p(X=1|Y=0)=0$$

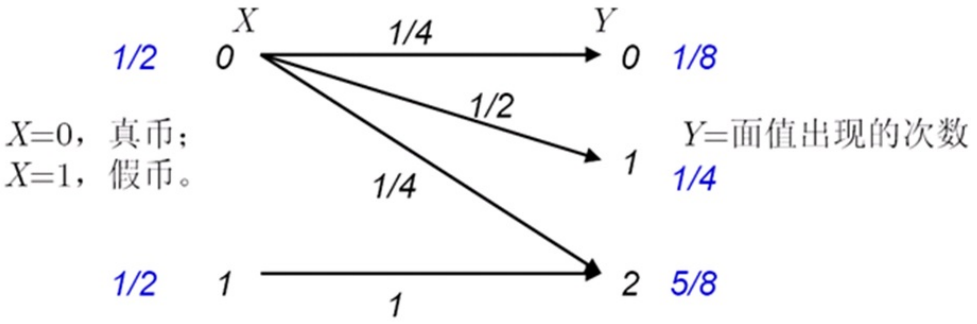
$$p(X=0|Y=1)=1/3, p(X=1|Y=1)=2/3$$

$$H(X|Y)=0.75\times H(1/3,2/3)=0.6885$$

$$I(X;Y)=H(X)-H(X|Y)=0.3115$$

互信息的例子

有两个硬币，一个是真币（一面国徽，一面面值），另一个是假币（两面都是面值）。随机抽取一个硬币，抛掷2 次。问出现面值的次数对于硬币的识别提供多少信息？



互信息的例子

$$
\begin{aligned}
H(Y) &= -(\frac{1}{8}\log\frac{1}{8} + \frac{1}{4}\log\frac{1}{4} + \frac{5}{8}\log\frac{5}{8}) \\
&= \frac{11}{4} - \frac{5}{8}\log 5 \\
H(Y|X) &= \Pr(X=0)H(Y|X=0) + \Pr(X=1)H(Y|X=1) \\
&= \frac{1}{2}H(Y|X=0) \\
&= -\frac{1}{2}(\frac{1}{4}\log\frac{1}{4} + \frac{1}{4}\log\frac{1}{4} + \frac{1}{2}\log\frac{1}{2}) = \frac{3}{4} \\
I(X;Y) &= H(Y) - H(Y|X) = 2 - \frac{5}{8}\log 5 = 0.55\text{bit}
\end{aligned}
$$

互信息的例子

$$
\begin{aligned}
H(X) &= 1 \\
H(X|Y) &= \Pr(Y=0)H(X|Y=0) + \Pr(Y=1)H(X|Y=1) + \Pr(Y=2)H(X|Y=2) \\
&= \Pr(Y=2)H(X|Y=2) \\
&= (\frac{1}{5}\log 5 + \frac{4}{5}\log\frac{5}{4}) \times \frac{5}{8} \\
&= \frac{5}{8}\log 5 - 1 \\
I(X;Y) &= H(X) - H(X|Y) = 2 - \frac{5}{8}\log 5
\end{aligned}
$$

条件互信息

事件的条件互信息

$$I(x;y|z) = \log \frac{p(x|y,z)}{p(x|z)} = \log \frac{p(x,y|z)}{p(x|z)p(y|z)}$$

在随机变量Z已知的条件下变量X与Y相互提供的信息量为

$$\begin{aligned} I(X;Y|Z) &= E\left\{I(x;y|z)\right\} \\ &= \sum_x \sum_y \sum_z p(x,y,z) \log \frac{p(x|y,z)}{p(x|z)} \end{aligned}$$

联合互信息

事件的联合互信息

$$I(x;y,z) = I(x) - I(x|y,z) = \log \frac{p(x|y,z)}{p(x)} = \log \frac{p(x,y,z)}{p(x)p(y,z)}$$

随机变量Y和Z共同提供给变量X的信息量为

$$\begin{aligned} I(X;Y,Z) &= \sum_x \sum_y \sum_z p(x,y,z) \log \frac{p(x|y,z)}{p(x)} \\ &= \sum_x \sum_y \sum_z p(x,y,z) \log \frac{p(x|z)p(x|y,z)}{p(x)p(x|z)} \\ &= I(X;Z) + I(X;Y|Z) \end{aligned}$$

相对熵（散度）

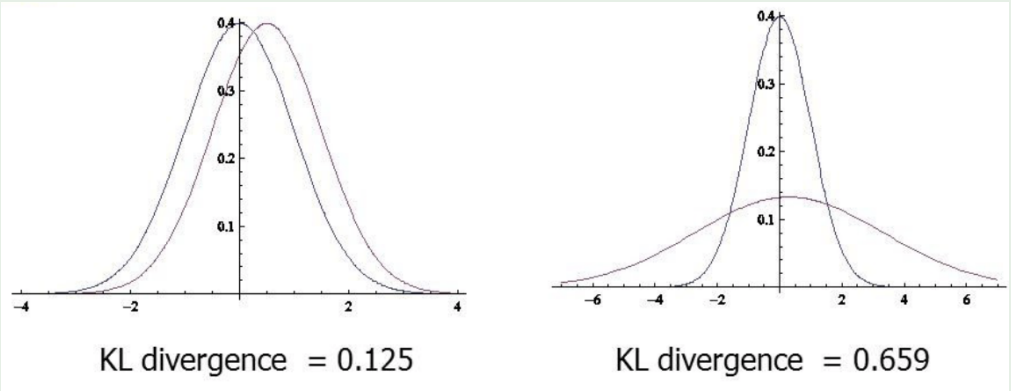
定义在相同字符表$\mathcal{X}$上的两个概率分布$\{p(x)\}$和$\{q(x)\}$之间的相对熵（散度），或称Kullback-Leibler距离，表示为：

$$D(p//q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_p \left\{ \log \frac{p(x)}{q(x)} \right\}$$

表示实际分布$\{p(x)\}$与假定分布$\{q(x)\}$之间的平均差距，因而也称鉴别熵。

相对熵（散度）

例子


0.4
0.3
0.2
0.1
−4
−2
2
4
KL divergence = 0.125
0.4
0.3
0.2
0.1
−6
−4
−2
2
4
6
KL divergence = 0.659

相对熵（散度）

例子

一个方法P得到四个类别的概率分别是$0.1, 0.2, 0.3, 0.4$；另一个方法Q得到四个类别的概率分别是$0.4, 0.3, 0.2, 0.1$。那么这两个分布的KL散度就是

$$D(P//Q) = 0.1 \times \log_2\left(\frac{0.1}{0.4}\right) + 0.2 \times \log_2\left(\frac{0.2}{0.3}\right)$$

$$+0.3 \times \log_2\left(\frac{0.3}{0.2}\right) + 0.4 \times \log_2\left(\frac{0.4}{0.1}\right)$$

相对熵的性质

$$
\begin{aligned}
(1)\quad D(p//q) &= \sum_x p(x)\log\frac{p(x)}{q(x)} = -\sum_x p(x)\log\frac{q(x)}{p(x)} \\
&\geq -\sum_x p(x)\left(\frac{q(x)}{p(x)}-1\right) = 0 \\
(2)\quad D(p//q) &\neq D(q//p) \\
(3)\quad I(X;Y) &= \sum_{x,y} p(xy)\log\frac{p(xy)}{p(x)p(y)} \\
&= D(p(xy)//p(x)p(y)) \geq 0
\end{aligned}
$$

相对熵的性质

(4) $$\begin{aligned} H(X) &= -\sum_x p(x)\log p(x) \\ &= -\sum_x p(x)\log\frac{p(x)}{1/K} + \log K \\ &= H(U) - D(X//U) \end{aligned}$$

其中U是均匀分布。

(5)如果P_1, P_2是独立分布，并且联合分布是$P = P_1P_2$,如果Q_1, Q_2是独立分布，并且联合分布是$Q = Q_1Q_2$，那么

$$D(P//Q) = D(P_1//Q_1) + D(P_2//Q_2)$$

相对熵的应用

相对熵的应用

假设数据通过未知分布$p(x)$生成，我们想要对$p(x)$ 进行建模，即使用参数分布$q(x|\theta)$来近似该分布

$$\min_{\theta} D(p//q) = \frac{1}{N}\sum_{n=1}^{N}\{-\log q(x_n|\theta) + \log p(x_n)\}$$

关于疑义度的Fano不等式

定义在相同字符表$\{0,1,\cdots,K-1\}$上的两个随机变量X和$\hat{X}$，其中$\hat{X}$是对X的某种估计，估计错误概率定义为：

$$P_E = \sum_{k=0}^{K-1} \sum_{\substack{j=0 \\ j\neq k}}^{K-1} \Pr\left\{X=k, \hat{X}=j\right\}$$

则$\hat{X}$已知条件下X的疑义度$H(X|\hat{X})$满足下述不等式：

$$H(X|\hat{X}) \leq H(P_E) + P_E \log(K-1)$$

关于疑义度的Fano不等式

$$
\begin{aligned}
&H(X|\hat{X}) - H(P_E) - P_E \log(K-1) \\
&= -\sum_{k=0}^{K-1}\sum_{j=0}^{K-1} p(k,j)\log p(k|j) + \sum_{k=0}^{K-1}\sum_{\substack{j=0\\ j\neq k}}^{K-1} p(k,j)\log P_E \\
&+ \sum_{k=0}^{K-1} p(k,k)\log(1-P_E) - \sum_{k=0}^{K-1}\sum_{\substack{j=0\\ j\neq k}}^{K-1} p(k,j)\log(K-1) \\
&= \sum_{k=0}^{K-1}\sum_{\substack{j=0\\ j\neq k}}^{K-1} p(k,j)\log\frac{P_E}{(K-1)p(k|j)} + \sum_{k=0}^{K-1} p(k,k)\log\frac{1-P_E}{p(k|k)} \\
&\leq \sum_{k=0}^{K-1}\sum_{\substack{j=0\\ j\neq k}}^{K-1} p(k,j)\left[\frac{P_E}{(K-1)p(k|j)} - 1\right] + \sum_{k=0}^{K-1} p(k,k)\left[\frac{1-P_E}{p(k|k)} - 1\right] \\
&= \frac{P_E}{(K-1)}\sum_{k=0}^{K-1}\sum_{\substack{j=0\\ j\neq k}}^{K-1} p(j) + (1-P_E)\sum_{k=0}^{K-1} p(j) - \sum_{k=0}^{K-1}\sum_{j=0}^{K-1} p(k,j) = 0
\end{aligned}
$$

关于疑义度的Fano不等式

证明

$$H(E,X|\hat{X}) = H(X|\hat{X}) + H(E|X,\hat{X}) = H(X|\hat{X})$$

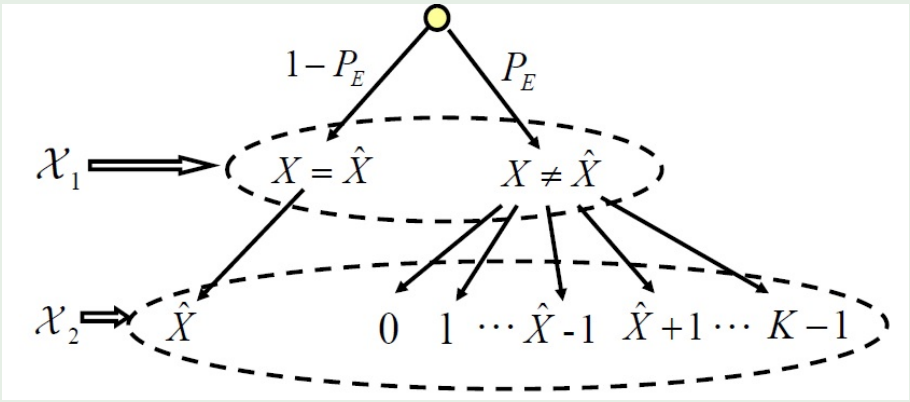
$$H(E,X|\hat{X}) = H(E|\hat{X}) + H(X|E,\hat{X})$$

$$H(E|\hat{X}) \leq H(E) = H(P_E)$$

$$H(X|E,\hat{X}) = P_E H(X|E=1,\hat{X}) + (1-P_E)H(X|E=0,\hat{X})$$

$$\leq P_E \log(K-1)$$

关于疑义度的Fano不等式



$$H(X \mid \hat{X})\,|_{X\in\mathcal{X}_2} = H(X_1|\hat{X})\,|_{X_1\in\mathcal{X}_1} + H(X \mid X_1, \hat{X})\,|_{X\in\mathcal{X}_2}^{X_1\in\mathcal{X}_1}$$

$$\leq H(P_E) + P_E \log(K-1)$$

关于疑义度的Fano不等式

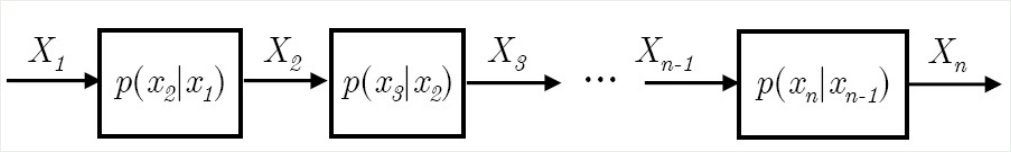
Fano不等式的物理意义

已知$\hat{X}$条件下，对X还存在的不确定性$H(X \mid \hat{X})$可分为两个部分。第一部分是估计$\hat{X}$ 是否准确，这部分的不确定性为$H(P_E)$；第二部分是如果估计是不准确的，这时X可能取值有$K-1$个，这部分的不确定性为$P_E(K-1)$。

Fano不等式在证明香农信道编码定理之逆定理时是必须应用的。

马尔可夫链

定义



如果随机变量序列$X_1, X_2, \cdots, X_n$的联合概率分布可以写成如下形式：

$$p(x_1, x_2, \cdots, x_n) = p(x_1)p(x_2 \mid x_1) \cdots p(x_n \mid x_{n-1})$$

则称这n个随机变量构成马尔可夫链，记为：

$$X_1 \to X_2 \to \cdots \to X_n$$

马尔可夫过程

一类重要的随机过程，它的原始模型马尔可夫链，由俄国数学家A.A. 马尔可夫于1907年提出。人们在实际中常遇到具有下述特性的随机过程：在已知它目前的状态（现在）的条件下，它未来的演变（将来）不依赖于它以往的演变（过去）。这种已知“现在”的条件下，“将来”与“过去”独立的特性称为马尔可夫性，具有这种性质的随机过程叫做马尔可夫过程。

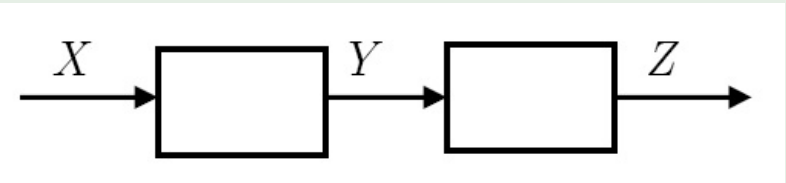
马尔可夫过程

荷花池中一只青蛙的跳跃是马尔可夫过程的一个形象化的例子。青蛙依照它瞬间而起的念头从一片荷叶上跳到另一片荷叶上，因为青蛙是没有记忆的，当现在所处的位置已知时，它下一步跳往何处和它以往走过的路径无关。如果将荷叶编号并用$X_0, X_1, X_2, \cdots$ 分别表示青蛙最初处的荷叶号码及第一次、第二次、……跳跃后所处的荷叶号码，那么$\{X_n, n \geq 0\}$ 就是马尔可夫过程。

马尔可夫过程

液体中微粒所作的布朗运动，传染病受感染的人数，原子核中一自由电子在电子层中的跳跃，人口增长过程等等都可视为马尔可夫过程。还有些过程（例如某些遗传过程）在一定条件下可以用马尔可夫过程来近似。

马尔可夫链



1. $p(xyz) = p(x)p(y \mid x)p(z \mid y) \Leftrightarrow p(z \mid xy) = p(z \mid y)$
2. $p(xz \mid y) = \dfrac{p(xyz)}{p(y)} = \dfrac{p(x,y)p(z \mid y)}{p(y)} = p(x \mid y)p(z \mid y) \Leftrightarrow I(X;Z \mid Y) = 0$
3. $p(xyz) = p(x)p(y \mid x)p(z \mid y) = \dfrac{p(x)p(xy)p(yz)p(z)}{p(x)p(y)p(z)} = p(z)p(y \mid z)p(z \mid y) \Leftrightarrow Z \to Y \to X$

数据处理定理

定理： 如果$X \to Y \to Z$，则

$$I(X;Y) \geq I(X;Z)$$

$$I(X;Y) \geq I(X;Y \mid Z)$$

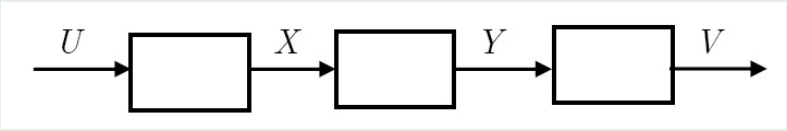
证明： $I(X;YZ) = I(X;Y) + I(X;Z \mid Y) = I(X;Z) + I(X;Y \mid Z)$

由于$X \to Y \to Z \Rightarrow I(X;Z \mid Y) = 0$，故

$$\begin{array}{c} I(X;Y) \geq I(X;Z) \\ I(X;Y) \geq I(X;Y \mid Z) \end{array} \quad \Big\| \quad \begin{array}{c} I(Y;Z) \geq I(X;Z) \\ I(Y;Z) \geq I(Y;Z \mid X) \end{array}$$

物理意义：增加数据处理的次数，不会使信息量增加

四变量马尔可夫链



定理： 如果$U \to X \to Y \to V$，则$I(X;Y) \geq I(U;V)$

证明：

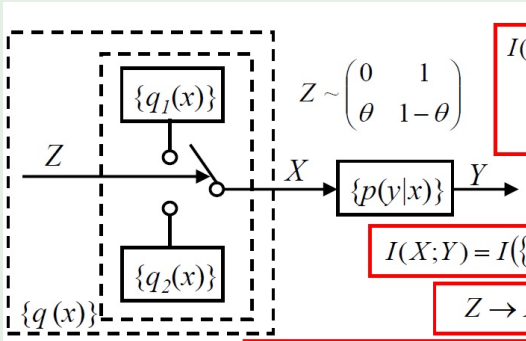
$$
\begin{array}{c}
U \to X \to Y \\
\Leftrightarrow Y \to X \to U \\
\Rightarrow I(Y;X) \geq I(Y;U) \\
\Leftrightarrow I(X;Y) \geq I(U;Y) \quad (1)
\end{array}
\Bigg\| \quad
\begin{array}{c}
U \to Y \to V \\
\Rightarrow I(U;Y) \geq I(U;V) \quad (2) \\
\\
\text{由(1)和(2)即得。}
\end{array}
$$

互信息的凸性(1)

$$
\begin{aligned}
I(X;Y) &= \sum_x \sum_y p(xy) \log \frac{p(x \mid y)}{q(x)} \\
&= \sum_x \sum_y q(x)p(y \mid x) \log \frac{p(y \mid x)}{\omega(y)} \\
&= \sum_x \sum_y q(x)p(y \mid x) \log \frac{p(y \mid x)}{\sum_x q(x)p(y \mid x)} \\
&= I\left(\{q(x)\}, \{p(y \mid x)\}\right)
\end{aligned}
$$

互信息$I(X;Y)$是关于输入分布$\{q(x)\}$和转移概率矩阵$\{p(y \mid x)\}$的函数

互信息的凸性(2)



$$I(X;Y|Z) = p(Z=0)I(X;Y|Z=0) + p(Z=1)I(X;Y|Z=1) = \theta I(\{q_1(x)\}) + (1-\theta)I(\{q_2(x)\})$$

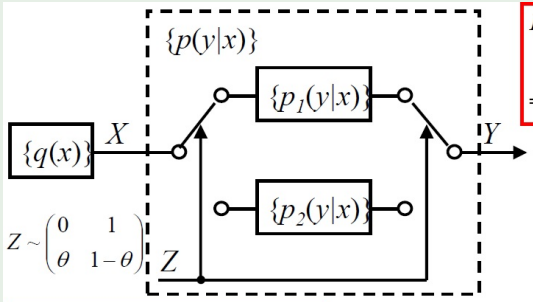
$$I(X;Y) = I(\{q(x)\}) = I(\{\theta q_1(x) + (1-\theta)q_2(x)\})$$

$$Z \to X \to Y \Rightarrow I(X;Y) \geq I(X;Y|Z)$$

$$I(\{\theta q_1(x) + (1-\theta)q_2(x)\}) \geq \theta I(\{q_1(x)\}) + (1-\theta)I(\{q_2(x)\})$$

定理一： 当转移概率矩阵$\{p(y \mid x)\}$给定时，互信息$I(X;Y) = I(\{q(x)\})$是输入分布的上凸（凹，concave）函数。

互信息的凸性(3)



$$I(X;Y|Z) = p(Z=0)I(X;Y|Z=0) + p(Z=1)I(X;Y|Z=1)$$
$$= \theta I(\{p_1(y|x)\}) + (1-\theta) I(\{p_2(y|x)\})$$

$$I(X;YZ) = \overbrace{I(X;Z)}^{=0} + I(X;Y|Z)$$
$$= I(X;Y) + \underbrace{I(X;Z|Y)}_{>=0}$$
$$\therefore I(X;Y) \le I(X;Y|Z)$$

定理二： 当输入分布$\{q(x)\}$给定时，互信息$I(X;Y) = I(\{P(y|x)\})$是转移概率矩阵的下凸（凸，convex）函数。

本讲小结

1. 随机事件的自信息，条件自信息，联合自信息，互信息
2. 随机变量的熵，条件熵，联合熵以及熵的性质
3. 随机变量互信息定义以及互信息的性质

作业：

- 2.1；2.2；2.4；2.5；2.6；2.9；2.12；2.13

第二讲：连续随机变量的互信息和微分熵

- 掌握连续随机变量互信息的定义
- 掌握连续随机变量微分熵的定义
- 掌握微分熵极值以及熵功率

针对离散随机变量，本讲将讨论连续随机变量

连续随机变量的概率密度

$p_{XY}(x,y)$ 是连续随机变量XY的联合概率密度函数

$p_X(x)$ 是X的边际概率密度

$p_Y(y)$ 是Y的边际概率密度

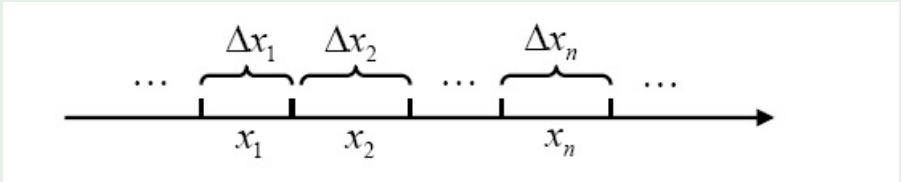
则：

$$p_X(x) = \int_{-\infty}^{+\infty} p_{XY}(x,y)dy$$

$$p_Y(y) = \int_{-\infty}^{+\infty} p_{XY}(x,y)dx$$

$$p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)}$$

连续随机变量的离散化



$X \in (-\infty, +\infty)$

$$\Rightarrow \begin{pmatrix} \cdots & x_i & \cdots \\ \cdots & p_X(x_i)\Delta x_i & \cdots \end{pmatrix}$$

$Y \in (-\infty, +\infty)$

$$\Rightarrow \begin{pmatrix} \cdots & y_j & \cdots \\ \cdots & p_Y(y_j)\Delta y_j & \cdots \end{pmatrix}$$

$(X, Y) \in R^2$

$$\Rightarrow \begin{pmatrix} \cdots & (x_i, y_j) & \cdots \\ \cdots & p_{XY}(x_i, y_j)\Delta x_i \Delta y_j & \cdots \end{pmatrix}$$

连续随机变量的互信息

联合连续随机变量$((X,Y),R^2,p_{XY}(x,y))$之间的互信息：

$$
\begin{aligned}
I(X;Y) &= \sum_{i=-\infty}^{+\infty}\sum_{j=-\infty}^{+\infty}[p_{XY}(x_i,y_j)\Delta x_i\Delta y_j]\log\frac{[p_{XY}(x_i,y_j)\Delta x_i\Delta y_j]}{[p_X(x_i)\Delta x_i][p_Y(y_j)\Delta y_j]} \\
&= \sum_{i=-\infty}^{+\infty}\sum_{j=-\infty}^{+\infty}\left(p_{XY}(x_i,y_j)\log\frac{p_{XY}(x_i,y_j)}{p_X(x_i)p_Y(y_j)}\right)\Delta x_i\Delta y_j \\
&\xrightarrow{\Delta x_i\to 0,\Delta y_j\to 0} \iint p_{XY}(x,y)\log\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}dxdy
\end{aligned}
$$

连续随机变量的互信息

$$I(X;Y|Z) = \iiint p_{XYZ}(x,y,z) \log \frac{p_{XY|Z}(x,y|z)}{p_{X|Z}(x|z)p_{Y|Z}(y|z)} dxdydz$$

$$I(X;YZ) = \iiint p_{XYZ}(x,y,z) \log \frac{p_{XYZ}(x,y,z)}{p_X(x)p_{YZ}(yz)} dxdydz$$

性质：

1) $I(X;Y) \geq 0$

2) $I(X;Y) = I(Y;X), I(X;Y|Z) = I(Y;X|Z)$

3) $I(X;YZ) = I(X;Y) + I(X;Z|Y) = I(X;Z) + I(X;Y|Z)$

4) *if* $X \to Y \to Z$, *then*

$I(X;Y) \geq I(X;Z); \quad I(X;Y) \geq I(X;Y|Z)$

连续随机变量的微分熵

连续随机变量$(X, R, p_X(x))$的离散化熵值：

$$
\begin{aligned}
H_\Delta(X) &= -\sum_{i=-\infty}^{+\infty} p_X(x_i)\Delta x_i \log\left(p_X(x_i)\Delta x_i\right) \\
&= -\sum_{i=-\infty}^{+\infty} [p_X(x_i)\log p_X(x_i)]\Delta x_i - \sum_{i=-\infty}^{+\infty} p_X(x_i)\Delta x_i \log \Delta x_i \\
&\xrightarrow{\Delta x_i \to 0} -\int p_X(x)\log\ p_X(x)dx + \infty
\end{aligned}
$$

定义连续随机变量X的微分熵

$H_C(X) = h(X) \triangleq -\int p_X(x)\log p_X(x)dx$

微分熵的本质和性质

1)微分熵$H_C(X)$不反映连续随机变量X的不确定性。连续随机变量的不确定性一般都是无穷大。但微分熵的确在一定程度上反映了该连续随机变量的相对不确定性。

2)$H_C(X)$可正，可负，可为0

例子：$p(x)=\begin{cases}\dfrac{1}{b-a}\,, x\in[a,b]\\ 0\,, x\notin[a,b]\end{cases}$ $H_C(X)=\int_a^b \dfrac{1}{b-a}\ln(b-a)dx=\ln(b-a)$

条件微分熵和联合微分熵及其性质

$$
\begin{aligned}
H_C(X,Y) &= -\iint p_{XY}(x,y)\log\ p_{XY}(x,y)dxdy \\
H_C(X|Y) &= -\iint p_{XY}(x,y)\log\ p_{X|Y}(x|y)dxdy \\
H_C(X,Y) &= H_C(X)+H_C(Y|X) = H_C(Y)+H_C(X|Y) \\
H_C(U^N) &= H_C(U_1,U_2,\cdots U_N) = \sum_{n=1}^{N} H_C(U_n|U_1U_2\cdots U_{n-1}) \\
&= \sum_{n=1}^{N} H_C(U_n|U^{n-1}) \\
I(X,Y) &= H_C(X)-H_C(X|Y) \\
&= H_C(Y)-H_C(Y|X) \\
&= H_C(X)+H_C(Y)-H_C(X,Y)
\end{aligned}
$$

微分熵的例子

已知某线段的长度在$[0, L]$之间均匀分布，现用精度为Δ 的尺子去量它，问量得的结果对于线段的真实长度提供多少信息？

令：X=线段的长度，Y=测量的结果

则：测量前$H_C(X) = \log L$

测量后$H_C(X|Y) = \log \Delta$

故：$I(X;Y) = H_C(X) - H_C(X|Y) = \log \dfrac{L}{\Delta}$

微分熵在线性变化下不具有不变性

对于离散随机变量X，令$Y = f(X)$是$X \to Y$上的一对一函数，则$H(X) = H(Y)$

但对于连续随机变量：

$$H_C(Y) = -\int p(y) \log\ p(y) dy = -\int p(x) \log p(x) f'(x) dx \neq H_C(X)$$

即使对于线性变换，微分熵也不具有不变性。

互信息的例子

例：二维正态分布的密度函数为

$$p_{XY}(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-m_x)^2}{\sigma_x^2}-\frac{2\rho(x-m_x)(y-m_y)}{\sigma_x\sigma_y}+\frac{(y-m_y)^2}{\sigma_y^2}\right]\right\}$$

其中相关系数$\rho=\dfrac{E\left[(x-m_x)(y-m_y)\right]}{\sigma_x\sigma_y}$

则：$I(X;Y)=-\dfrac{1}{2}\ln(1-\rho^2)$

当$\rho=1$时，$I(X;Y)=-\dfrac{1}{2}\ln(1-\rho^2)\to\infty$，$I(X;Y)\le H(X)$

当$\rho=0$时，$I(X;Y)=-\dfrac{1}{2}\ln(1-\rho^2)=0$

证明：

$$I(X;Y) = E\left\{\log \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\right\}$$

$$= E\left\{-\log\sqrt{1-\rho^2} + \left\{-\frac{1}{2(1-\rho^2)}\left[x^2 - 2\rho xy + y^2\right] + \frac{1}{2}\left[x^2+y^2\right]\right\}\right\}$$

$$= -\frac{1}{2}\log(1-\rho^2) + E\left\{-\frac{1}{2(1-\rho^2)}\left[\rho^2x^2 - 2\rho xy + \rho^2y^2\right]\right\}$$

$$= -\frac{1}{2}\log(1-\rho^2) - \frac{1}{2(1-\rho^2)}\left[\rho^2 E\left\{x^2\right\} - 2\rho E\left\{xy\right\} + \rho^2 E\left\{y^2\right\}\right]$$

$$= -\frac{1}{2}\log(1-\rho^2) - \frac{1}{2(1-\rho^2)}\left[\rho^2 - 2\rho^2 + \rho^2\right]$$

$$= -\frac{1}{2}\log(1-\rho^2)$$

微分熵的极大化：峰值受限

定理：设X取值限于$(-M, M)$,即$\int_{-M}^{M} p(x)dx = 1$,这时微分熵$H_C(X) \leq \ln 2M$,等号在均匀分布时达到。

证明(Lagrange乘子法)：

$$
\begin{aligned}
&J\left(p(x)\right) \triangleq H_C(X) - \lambda \int_{-M}^{+M} p(x)dx \\
&= -\int_{-M}^{+M} p(x) \ln p(x)dx - \lambda \int_{-M}^{+M} p(x)dx \\
&= -\int_{-M}^{+M} p(x) \ln \left(e^{\lambda} p(x)\right) dx \\
&\leq -\int_{-M}^{+M} p(x) \left(\frac{1}{e^{\lambda} p(x)} - 1\right) dx \\
&= \frac{2M}{e^{\lambda}} - 1 = const.
\end{aligned}
$$

等号成立的条件为$\frac{1}{e^{\lambda} p(x)} = 1$,即$p(x) = e^{-\lambda} = const$,为均匀分布。

由约束条件$\int_{-M}^{M} p(x)dx = 1$得,$p(x) = \frac{1}{2M}$.故$H_C(X) \leq \ln(2M)$

微分熵的极大化：平均功率受限

定理：在方差σ^2一定条件下，当X服从正态分布时，微分熵最大，
即$H_C(X) \leq \ln(\sqrt{2\pi e}\sigma)$。

证明(Lagrange乘子法)：
$J(p(x)) \triangleq H_C(X) - \lambda_1 \int_{-\infty}^{+\infty} p(x)dx - \lambda_2 \int_{-\infty}^{+\infty} p(x)(x-m)^2 dx$
$= \int_{-\infty}^{+\infty} p(x) \ln\left(\frac{e^{\lambda_1} e^{-\lambda_2(x-m)^2}}{p(x)}\right) dx \leq \int_{-\infty}^{+\infty} p(x)\left(\frac{e^{\lambda_1} e^{-\lambda_2(x-m)^2}}{p(x)} - 1\right) dx$
$= const.$

等号成立的条件为$p(x) = e^{\lambda_1} e^{-\lambda_2(x-m)^2}$。
根据约束条件$\int_{-\infty}^{+\infty} p(x)dx = 1$及$\int_{-\infty}^{+\infty} p(x)(x-m)^2 dx = \sigma^2$
得$p(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$ 此时$H_C(X) = \frac{1}{2}\ln\left(2\pi e\sigma^2\right)$

熵功率

熵功率的定义

$\overline{\sigma_x}^2 = \frac{1}{2\pi e} e^{2H_C(X)}$

高斯随机变量的熵功率

高斯随机变量$X \sim p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}x^2\right\}$的微分熵为

$H_C(X) = \frac{1}{2}\ln(2\pi e\sigma^2)$

其熵功率为

$\overline{\sigma_x}^2 = \frac{1}{2\pi e} e^{2H_C(X)} = \sigma^2$刚好为高斯随机变量的方差。

熵功率不等式

$$H_C(X) \leq \ln(\sqrt{2\pi e}\sigma)$$

$$\Updownarrow$$

$$\overline{\sigma_x}^2 = \frac{1}{2\pi e} e^{2H_C(X)} \leq \sigma^2$$

1. 随机变量的熵功率一般不大于功率，只有高斯变量其熵功率与功率相等。

2. 功率一定时，高斯变量的熵功率最大，与功率相等。

本讲小结

- 连续随机变量互信息的定义
- 连续随机变量微分熵的定义
- 熵功率定义
- 微分熵极值以及熵功率不等式

作业：

2.19；2.20

第三讲：平稳离散信源的熵

- 了解平稳离散信源的概念
- 掌握平稳离散信源熵的定义
- 了解马尔可夫信源的概念
- 掌握马尔可夫信源熵的计算方法

将随机变量熵的定义推广到随机过程

平稳随机过程

定义

对于任意的n，任意的$t_1, t_2, \cdots, t_n \in T$ 和h，

若$(X(t_1), X(t_1), \cdots, X(t_n))$与$(X(t_1+h), X(t_2+h), \cdots, X(t_n+h))$ 具有同样的分布，则称随机过程$\{X(t)\}$是平稳随机过程。

平稳随机过程的性质

- $E(X(t_n)) = E(X(t_n+h)) = E(X(0)) = Const.$
- $X(t)$的均值和方差对于所有t都一样。

平稳信源的定义

$$\cdots X_{-1}, X_0, X_1, X_2, \cdots, X_n, \cdots$$

平稳信源：任意长度片段的联合概率分布与时间起点无关

$$Pr(X_1 X_2 \cdots X_L) = Pr(X_{1+n} X_{2+n} \cdots X_{L+n})$$

简单无记忆信源：不同时间的随机变量不相关

$$Pr(X_1 X_2 \cdots X_L) = \prod_{i=1}^{L} Pr(X_i)$$

m阶马尔可夫信源：（$m = 1$：马尔可夫信源）

$$Pr(X_l \mid X_{l-1} X_{l-2} \cdots X_0) = Pr(X_l \mid X_{l-1} X_{l-2} \cdots X_{l-m})$$

平稳信源的熵

如果一个平稳信源发出长度为N的序列$X_1, X_2, \cdots, X_n$，令N 维随机矢量$\mathrm{X} = (X_1, X_2, \cdots, X_n)$，则

$$H(\mathrm{X}) = H(X_1, X_2, \cdots, X_n) = -\sum p(x_1, x_2, \cdots, x_n) \log p(x_1, x_2, \cdots, x_n)$$

$H(\mathrm{X})$随N增长而增长，趋向无穷大

平均每符号熵： $H_N(\mathrm{X}) \triangleq \dfrac{1}{N} H(\mathrm{X}) = \dfrac{1}{N}(X_1 X_2 \cdots X_N)$

熵速率： $H_\infty(\mathrm{X}) = \lim_{N\to\infty} H_N(\mathrm{X})$

平均条件熵： $H(X_N \mid X_{N-1} X_{N-2} \cdots X_1)$

平稳信源熵的性质

1. $H(X_N \mid X_{N-1}X_{N-2}\cdots X_1)$随$N$增大而单调不增
2. $H_N(\mathrm{X})$随N增大也单调不增
3. $H_N(\mathrm{X}) \geq H(X_N \mid X_{N-1}X_{N-2}\cdots X_1)$
4. $H_\infty(\mathrm{X}) = \lim_{N\to\infty} H_N(\mathrm{X}) = \lim_{N\to\infty} H(X_N \mid X_{N-1}X_{N-2}\cdots X_1)$

平稳信源熵的性质

(1)$H(X_N \mid X_{N-1}X_{N-2}\cdots X_1)$随$N$增大而单调不增

利用$H(X \mid Y) \leq H(X)$

$$
\begin{aligned}
&H(X_N \mid X_1, X_2, \cdots, x_{N-1}) \\
&\leq H(X_N \mid X_2, \cdots, x_{N-1}) \\
&\leq H(X_{N-1} \mid X_1, \cdots, x_{N-2})
\end{aligned}
$$

平稳信源熵的性质

(2)$H_N(\mathbf{X}) \geq H(X_N \mid X_{N-1}X_{N-2}\cdots X_1)$

$$
\begin{aligned}
H_N(\mathbf{X}) &= \frac{1}{N}H(X_1, X_2, \cdots, X_{N-1}, X_N) \\
&= \frac{1}{N}\{H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_N \mid X_1, X_2, \cdots, X_{N-1})\} \\
&\geq H(X_N \mid X_1, \cdots, X_{N-2}, X_{N-1})
\end{aligned}
$$

平稳信源熵的性质

(3)$H_N(\mathrm{X})$随N增大也单调不增

$$
\begin{aligned}
H_N(\mathrm{X}) &= \frac{1}{N}\left\{H(X_1, X_2, \cdots, X_{N-1}) + H(X_N \mid X_1, X_2, \cdots, X_{N-1})\right\} \\
&= \frac{1}{N}\left\{(N-1)H_{N-1}(\mathrm{X}) + H(X_N \mid X_1, X_2, \cdots, X_{N-1})\right\} \\
&\leq \frac{1}{N}\left\{(N-1)H_{N-1}(\mathrm{X}) + H_N(\mathrm{X})\right\} \\
&\therefore H_N(\mathrm{X}) \leq H_{N-1}(\mathrm{X})
\end{aligned}
$$

平稳信源熵的性质

(4)$H_\infty(\mathrm{X}) = \lim_{N\to\infty} H_N(\mathrm{X}) = \lim_{N\to\infty} H(X_N \mid X_{N-1}X_{N-2}\cdots X_1)$

$$
\begin{aligned}
&(N+K)H_{N+K}(\mathrm{X}) = H(X_{N+K}X_{N+K-1}\cdots X_2X_1)\\
&= H(X_{N-1}X_{N-2}\cdots X_2X_1) + H(X_N \mid X_{N-1}X_{N-2}\cdots X_2X_1)\\
&\qquad + \cdots + H(X_{N+K} \mid X_{N+K-1}X_{N+K-2}\cdots X_2X_1)\\
&\le H(X_{N-1}X_{N-2}\cdots X_1) + (K+1)H(X_N \mid X_{N-1}X_{N-2}\cdots X_2X_1)\\
&H_{N+K}(\mathrm{X}) \le \frac{1}{N+K}H(X_{N-1}X_{N-2}\cdots X_1) + \frac{K+1}{N+K}H(X_N \mid X_{N-1}X_{N-2}\cdots X_2X_1)
\end{aligned}
$$

令$K\to\infty$，则$H_\infty(\mathrm{X}) \le H(X_N \mid X_{N-1}X_{N-2}\cdots X_2X_1) \le H_N(\mathrm{X})$

再令$N\to\infty$,则$H_\infty(\mathrm{X}) \le \lim_{N\to\infty} H(X_N \mid X_{N-1}X_{N-2}\cdots X_1) \le H_\infty(\mathrm{X})$

即$H_\infty(\mathrm{X}) = \lim_{N\to\infty} H_N(\mathrm{X}) = \lim_{N\to\infty} H(X_N \mid X_{N-1}X_{N-2}\cdots X_1)$

熵速率（熵率）

熵速率：

$$H_\infty(\mathrm{X}) = \lim_{N\to\infty} H_N(\mathrm{X}) = \lim_{N\to\infty} H(X_N \mid X_{N-1}X_{N-2}\cdots X_1)$$

$$H_\infty(\mathrm{X}) \le H(X_N \mid X_{N-1}X_{N-2}\cdots X_2X_1)$$

$$\le \cdots \le H(X_2 \mid X_1) \le H(X_2) \le \log K$$

熵的相对率：

$$\eta = \frac{H_\infty(\mathrm{X})}{\log K}, \quad \eta \le 1$$

信源的冗余度： $R = 1 - \eta$

熵速率的例子

当英文26个字母加上空格共27个字母等概时，最大熵为

$$H_0 = \log 27 = 4.766\text{比特/字母}$$

考虑到各个字符出现的概率，而不考虑字符间的依赖关系，得到

$$H_1 = 4.036$$

考虑到前后2-3个字符的关联性，可以统计出

$$H_2 = 3.326 \quad H_3 = 3.31$$

一般地，有$H_\infty = 1.4$，也就是说$\eta = \dfrac{H_\infty(X)}{\log K} = 0.29$

说明英文写作时，71%是由语言结构预先确定的信息，仅有29%是作者可以自由选择的。

计算机随机产生英文文章时，如何？

熵速率的例子

法文：3.98比特

西班牙文：4.01比特

英文：4.03比特

俄文：4.35比特

中文：9.65比特

优点？缺点？

熵速率的例子

联合国几乎所有的文件中，中文版是最薄的。

熵速率的例子

联合国几乎所有的文件中，中文版是最薄的。

汉语的信息量最大，也是最难学的，文言文更难学。

熵速率的例子

床前明月光，疑是地上霜。
举头望明月，低头思故乡。

熵速率的例子

床前明月光，疑是地上霜。
举头望明月，低头思故乡。

There is a bright moon high above my bed.
Thought it was ground frost.
Looking up I can see the bright moon.
And looking down I think about my home town.

熵速率的例子

窈窕淑女 君子好逑

熵速率的例子

窈窕淑女 君子好逑

Beautiful girls never fails to fascinate gentleman

Beautiful ladies, gentlemen's good mate

马尔可夫信源

m阶马尔可夫信源

$$\Pr(X_l|X_{l-1}X_{l-2}\cdots X_0) = \Pr(X_l|X_{l-1}X_{l-2}\cdots X_{l-m})$$

马尔可夫信源的状态空间

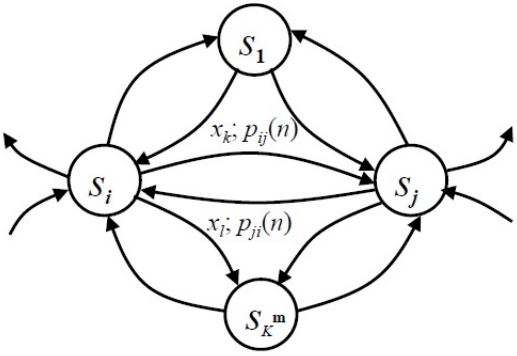
$$x^m = x_{i_{l-1}}x_{i_{l-2}}\cdots x_{i_{l-m}} \in \mathcal{X}^m, |\mathcal{X}^m| = K^m$$

马尔可夫信源的特点

1.状态空间有限

2.由状态转移函数确定：$s_{i_{n+1}} = F(s_{i_n}, x_{i_n})$

马尔可夫信源的状态图表示



$$s_i = \mathbf{x_i} = x_{i_1}x_{i_2}\cdots x_{i_m} \in \mathcal{X}^m \qquad s_j = \mathbf{x_j} = x_k x_{i_1}x_{i_2}\cdots x_{i_{m-1}} \in \mathcal{X}^m$$

$$\begin{aligned} p_{ij}(n) &= \Pr(X_n = x_k | X_{n-1}X_{n-2}\cdots X_{n-m} = x_{i_1}x_{i_2}\cdots x_{i_m}) \\ &= \Pr(S_n = s_j | S_{n-1} = s_i) \end{aligned}$$

时齐既约马尔可夫源的稳态状态分布

时齐（时不变）马尔可夫源：状态转移概率$p_{ij}(n)$与时间n无关。

既约（不可约）马尔可夫源：从任一状态出发，经有限步总可以到达任一其他状态。

状态转移概率矩阵： $P = [p_{ij}]_{K^m \times K^m}$

n时刻的状态概率分布： $Q(n) = (q_1(n), q_2(n), \cdots, q_{K^m}(n))$

其中$q_i(n) = \Pr(S_n = s_i)$

$$Q(n+1) = Q(n)P$$

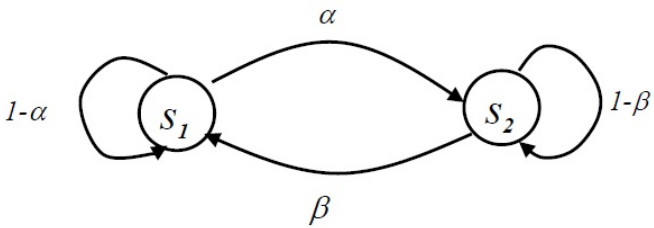
对时齐既约马尔可夫源，状态的稳态分布存在

$$\bar{Q} = \lim_{n\to\infty} Q(n+1) = \lim_{n\to\infty} Q(n)P = \bar{Q}P$$

马尔可夫源的熵率

$$
\begin{aligned}
H_\infty &= \lim_{N\to\infty} \frac{1}{N} H(X_1, X_2, \cdots, X_N) \\
&= \lim_{N\to\infty} H(X_N | X_{N_1}, X_{N-2}, \cdots, X_2, X_1) \\
&= \lim_{N\to\infty} H(X_N | X_{N_1}, X_{N-2}, \cdots, X_{N-m}) \\
&= H(X_{m+1} | X_m, X_{m-1}, \cdots, X_1) \\
&= \sum_{i=1}^{K^m} q(S = s_i) H(X | S = s_i) \\
&= H(X|S)
\end{aligned}
$$

二元一阶马尔可夫源的熵率



$$P = [P_{ij}]_{2\times 2} = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} \quad \begin{cases} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}^T \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} & \mu_1 = \dfrac{\beta}{\alpha+\beta} \\ \mu_1+\mu_2=1 & \mu_2 = \dfrac{\alpha}{\alpha+\beta} \end{cases}$$

$$H_\infty = \mu_1 H(X|s_1) + \mu_2 H(X|s_2)$$

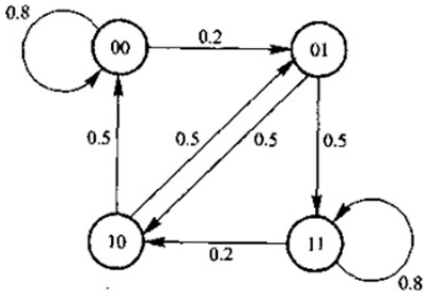
$$H(X|s_1) = -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$$

$$H(X|s_2) = -\beta \log \beta - (1-\beta)\log(1-\beta)$$

二阶二元马尔可夫信源

一个二元二阶马尔可夫信源，有$2^2 = 4$个状态，分别记为00, 01, 10, 11，其状态转移图和状态转移概率矩阵如下所示

$$\mathbf{P} = \begin{bmatrix} 0.8 & 0.2 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.2 & 0.8 \end{bmatrix}$$



二阶二元马尔科夫信源

则稳定状态概率$\mu_{00}, \mu_{01}, \mu_{10}, \mu_{11}$满足

$$\begin{bmatrix} \mu_{00} \\ \mu_{01} \\ \mu_{10} \\ \mu_{11} \end{bmatrix} = \mathrm{P}^{\mathrm{T}} \cdot \begin{bmatrix} \mu_{00} \\ \mu_{01} \\ \mu_{10} \\ \mu_{11} \end{bmatrix}$$

以及

$$\mu_{00} + \mu_{01} + \mu_{10} + \mu_{11} = 1$$

可以解出稳态概率为

$$\mu_{00} = \mu_{11} = \frac{5}{14}$$

$$\mu_{01} = \mu_{10} = \frac{1}{7}$$

二阶二元马尔科夫信源

该二元二阶马尔科夫信源的熵率为

$$H_\infty = \mu_{00}H(X|00) + \mu_{01}H(X|01) + \mu_{10}H(X|10) + \mu_{11}H(X|11)$$

其中

$$H(X|00) = H(X|11) = -0.8\log 0.8 - 0.2\log 0.2 = 0.723\text{bit}$$

$$H(X|01) = H(X|10) = -0.5\log 0.5 - 0.5\log 0.5 = 1\text{bit}$$

所以

$$H_\infty \approx 0.802\text{bit}$$

本讲小结

- 平稳离散信源的概念
- 平稳离散信源熵的定义
- 马尔可夫信源的概念
- 马尔可夫信源熵的定义

- 作业：2.23；2.25