第一章随机变量及其信息度量

第五节互信息或互熵

陈兴同

中国矿业大学 数学学院

2021 年 8 月

内容提要

1. 什么是互熵

内容提要

内容提要

1. 什么是互熵
2. 马氏链
3. 条件互信息

内容提要

定义 1.5.1：互信息定义

为了确定两个随机变量包含的公共信息量我们引入互熵，它实际上是两个事件间互自信息量的概率平均值。
设随机变量 X, Y 有联合分布律 $p(x, y)$，并且有边缘分布律 $p_X(x), p_Y(y)$，则称

1.32

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p_X(x)p_Y(y)} \qquad \cancel{(1.1)}$$

为两个随机变量 X 与 Y 的**互熵**或**互信息量**。

互熵其它定义

（1） 互信息也可以写成相对熵 $I(X;Y) = D(p||p_X p_y)$。

（2） 互信息也可以利用条件概率写成

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(y|x)}{p_Y(y)} \quad \cancel{(1.2)}$$

$$= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x|y)}{p_X(x)}. \quad \cancel{(1.3)}$$

1.33

（3） 规定 $I(X;X) = H(X)$。

多个随机变量的互熵

也可以定义多个随机向量间的互信息。比如：（1）$I(X,Y;Z)$ 表示随机向量 (X,Y) 与随机变量 Z 之间的互信息

$$I(X,Y;Z)=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}}p(x,y,z)\log\frac{p(x,y,z)}{p_{XY}(x,y)p_Z(z)}.$$

（2）$I(X;Y,Z)$ 表示随机向量 (Y,Z) 与随机变量 X 之间的互信息

$$I(X;Y,Z)=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}}p(x,y,z)\log\frac{p(x,y,z)}{p_X(x)p_{YZ}(y,z)}.$$

（3）而 $I(X,Y;U,V)$ 表示随机向量 (X,Y) 与随机向量 (U,V) 之间的互信息；
$I(X_1,X_2,\cdots,X_n;Y)$ 表示随机向量 $X_1,X_2,\cdots,X_n$ 与随机变量 Y 之间的互信息。

命题 1.5.1：互熵简单性质

互信息具有如下一些简单性质：

（1） $I(X;Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)$。

（2） $I(X;Y)=H(X)+H(Y)-H(X,Y)$。

（3） $I(X;Y)\geq 0$；当且仅当 X,Y 相互独立时等号成立。

（4） $I(X;Y)\leq \min\{H(X),H(Y)\}$。

（5） $I(X;Y)=I(Y;X)$。

（6） $I(X;X)=H(X)$。

练习:

两个随机变量的互信息何时为零?

例题 1.5.1

已知每 100 人中有 2 人患有某种疾病。为查明病情要进行某项指标的化验，这种化验结果对于有病的人总是呈阳性的，对于健康的人来说一半呈阳性半呈阴性。试问这种化验对于查明病情能提供多少信息？

解：

设随机变量

$$X=\begin{cases}0 & \text{健康的人}\\ 1 & \text{有病的人}\end{cases}, Y=\begin{cases}0 & \text{化验结果为阴性}\\ 1 & \text{化验结果为阳性}\end{cases},$$

于是

$$X\sim p_X(x)=\begin{pmatrix}0 & 1\\ 0.98 & 0.02\end{pmatrix}, H(X)=0.098\text{nats}.$$

并且条件分布矩阵为

$$p(y|x)=\begin{pmatrix}0.5 & 0.5\\ 0 & 1\end{pmatrix}.$$

续解：

本题相当于求“病情” X 与“化验结果” Y 的公共信息量是多少，即互信息 $I(X;Y)$。由乘法公式 $p(x,y)=p_X(x)p(y|x)$ 得 X,Y 联合分布律

$$(X,Y)\sim p(x,y)=\begin{pmatrix}0.49 & 0.49\\ 0 & 0.02\end{pmatrix},$$

从而联合熵 $H(X,Y)=0.7773\text{nats}$。
由边缘分布求得 Y 分布律:$p_Y(0)=0.49, p_Y(1)=0.51$，从而 $H(Y)=0.6929\text{nat}$。所求的互信息是

$$I(X;Y)=H(X)+H(Y)-H(X,Y)=0.0137\ \text{nats}.$$

定义 1.5.3：马氏链（条件独立性）

为了讨论条件互信息的性质，需要定义条件独立性或马氏链。如果三个随机变量 X, Y, Z 之间有如下条件概率

1.35

$$p(z|y, x) = p(z|y), \forall x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}, \qquad (\cancel{2.1})$$

成立，则称随机变量 X, Y, Z 构成**马氏链**，记成 $X \to Y \to Z$。可以解释为：随机变量 Z 的取值只受到前一个相邻的随机变量 Y 取值影响，与更前面的随机变量 X 取值无关。

引理 1.5.1（马氏链的条件）：

下列每个条件都是随机变量 X, Y, Z 构成马氏链的充要条件。

（1） $p(x, z|y) = p(x|y)p(z|y)$ 对任何 x, y, z 都成立，这称为 X, Z 关于 Y **条件独立**；

（2） 联合概率公式 $p(x, y, z) = p(x)p(y|x)p(z|y)$ 对任何 x, y, z 都成立；

（3） 随机变量 Z, Y, X 构成马氏链。

证明：

$$
\begin{aligned}
X \to Y \to Z \quad &\Leftrightarrow \quad p(z|y,x) = p(z|y) \Leftrightarrow \frac{p(x,y,z)}{p_{XY}(x,y)} = p(z|y) \\
&\Leftrightarrow \quad \frac{p(x,y,z)}{p_Y(y)} = \frac{p_{XY}(x,y)}{p_Y(y)}.p(z|y) \\
&\Leftrightarrow \quad p(x,z|y) = p(x|y)p(z|y).
\end{aligned}
$$

多个随机变量的马氏链:

更一般的情况是：如果对任何 $i = 2, 3, \cdots, n-1$ 总有

$$p(x_{i+1}|x_i, \cdots, x_2, x_1) = p(x_{i+1}|x_i),$$

则称随机变量序列 $X_1, X_2, \cdots, X_n$ 构成马氏链，记成 $X_1 \to X_2 \to \cdots \to X_n$。这也相当于联合概率公式为

$$p(x_1, x_2, \cdots, x_i) = p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_i|x_{i-1}).$$

引理 1.5.2：马氏链的子链

设 $U \to X \to Y \to V$，则

（1） $U \to X \to Y$；$X \to Y \to V$；

（2） $U \to X \to V$；$U \to Y \to V$。

证明：

只证明第（2）条中第一个。事实上：

$$\begin{aligned} p(u,v|x) &= \frac{p(u,x,v)}{p_X(x)} = \frac{\sum_y p(u,x,y,v)}{p_X(x)} \\ &= \frac{\sum_y p(u,x)p(y|u,x)p(v|u,x,y)}{p_X(x)} \\ &= \frac{\sum_y p(u,x)p(y|x)p(v|y)}{p_X(x)} \\ &= \frac{p(u,x)}{p_X(x)} \sum_y p(y|x)p(v|y) \end{aligned}$$

续证明：

$$
\begin{aligned}
&= p(u|x)\sum_y p(y|x)p(v|y) \\
&= p(u|x)\sum_y \frac{p(x,y)}{p_X(x)}p(v|x,y) \\
&= p(u|x)\sum_y \frac{p(x,y,v)}{p_X(x)} \\
&= p(u|x)\frac{p(x,v)}{p_X(x)} = p(u|x)p(v|x).
\end{aligned}
$$

这些性质在信道容量与编码定理等内容中用到。

练习：

设 X, Y 是两个随机变量，并且 Z 是 Y 的函数即 $Z = f(Y)$，则 $X \to Y \to Z$。

条件互信息定义/5.2

类似于条件熵，也引入条件互信息。
在给定随机变量 Z 的条件下随机变量 X, Y 的**条件互信息**定义为

$$I(X;Y|Z) = \sum_{z\in\mathcal{Z}} p_Z(z) \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_{X,Y|Z}(x,y|z) \log \frac{p(x,y|z)}{p_{X|Z}(x|z)p_{Y|Z}(y|z)}, \quad \cancel{(3.1)}$$

134

其中 $p(x,y,z)$ 是 X, Y, Z 的联合分布，而 $p_{X,Y|Z}(x,y|z), p_{X|Z}(x|z), p_{Y|Z}(y|z)$ 分别是在给定随机变量 Z 的条件下的条件分布。

相关计算公式:

定义中的联合分布及边缘分布、条件分布计算公式是

$$
\begin{aligned}
p_{XZ}(x,z) &= \sum_{y\in\mathcal{Y}} p(x,y,z), p_{YZ}(y,z) = \sum_{x\in\mathcal{X}} p(x,y,z) \\
p_Z(z) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y,z), p_{X,Y|Z}(x,y|z) = \frac{p(x,y,z)}{p_Z(z)} \\
p_{X|Z}(x|z) &= \frac{p_{XZ}(x,z)}{p_Z(z)}, p_{Y|Z}(y|z) = \frac{p_{YZ}(y,z)}{p_Z(z)}.
\end{aligned}
$$

更一般条件互信息：

也可以定义更一般的条件互信息，比如:
（1）$I(X,Y;Z|U)$ 表示在给定随机变量 U 条件下随机向量 (X,Y) 与随机变量 Z 之间的条件互信息

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

$$\sum_{u\in\mathcal{U}} p_U(u) \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \sum_{z\in\mathcal{Z}} p_{X,Y,Z|U}(x,y,z|u) \log \frac{p_{X,Y,Z|U}(x,y,z|u)}{p_{X,Y|U}(x,y|u)p_{Z|U}(z|u)}.$$

续更一般条件互信息：

（2）$I(X,Y;Z|U,V)$ 表示在给定随机向量 (U,V) 条件下随机向量 (X,Y) 与随机变量 Z 之间的互信息。
（3）$I(X;Y|U,V)$ 表示在给定随机向量 (U,V) 条件下随机变量 X 与随机向量 Y 之间的互信息。
（4）再如：
$I(X_1,X_2,\cdots,X_n;Y|Z_1,Z_2,\cdots,Z_n)$ 表示在给定随机向量 $(Z_1,Z_2,\cdots,Z_n)$ 条件下随机向量 $X_1,X_2,\cdots,X_n$ 与随机变量 Y 之间的互信息。

定理 1.5.1：条件熵的性质

$H(X,Y|Z)=H(X|Z)+H(Y|Z)$ 充要条件是 $X\to Z\to Y$。

证明：

利用条件熵定义得

$$
\begin{aligned}
& H(X|Z) + H(Y|Z) - H(X,Y|Z), \\
&= \sum_z p_Z(z) \sum_x p(x|z) \log \frac{1}{p(x|z)} + \sum_z p_Z(z) \sum_y p(y|z) \log \frac{1}{p(y|z)} \\
&- \sum_z p_Z(z) \sum_x \sum_y p(x,y|z) \log \frac{1}{p(x,y|z)}, \\
&= \sum_z p_Z(z) \sum_x \sum_y p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}, \\
&= \sum_z p_Z(z) D_z,
\end{aligned}
$$

1.36
~~(3.2)~~

其中 D_z 是条件联合分布 $p(x,y|z)$ 对条件乘积分布 $p(x|z)p(y|z)$ 的相对熵。

续证明：

充分性：如果 $X \to Z \to Y$，则有 $p(x,y|z) = p(x|z)p(y|z)$，从而相对熵 D_z 恒为 0，这证明了充分性。

必要性：如果 $H(X,Y|Z) = H(X|Z) + H(Y|Z)$，则式（3.2）必为 0，由于它的每一项都是非负的故有 $p_Z(z)D_z$ 恒为 0，因为 $p_Z(z)$ 恒不为 0，故条件相对熵 D_z 恒为 0，再由相对熵性质即得 $p(x,y|z) = p(x|z)p(y|z)$，这就证明了必要性。

定理 1.5.2：条件互信息性质

（1） 有多种表达式：

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

（2） $I(X;Y|Z) \geq 0$，等号成立当且仅当 X, Z, Y 构成马氏链即 $X \to Z \to Y$。

（3） $I(X;Y,Z) \geq I(X;Z)$ 等号成立当且仅当 $X \to Z \to Y$。

（4） $I(X;Y,Z) \geq I(X;Y)$，等号成立当且仅当 $X \to Y \to Z$。

（5） 如果 $X \to Z \to Y$，则 $I(X;Y) \leq I(X;Z), I(X;Y) \leq I(Y;Z)$，它们称为**数据处理不等式**。

定理 1.5.3：互信息链式法则

类似于联合熵的链式法则也有互信息的链式法则。

$$I(X_1, X_2, \cdots, X_n; Y) = I(X_1; Y)) + \sum_{i=2}^{n} I(X_i; Y | X_{i-1} \cdots X_1).$$

证明：

利用联合熵、条件熵的链式法则可得

$$
\begin{aligned}
& I(X_1, X_2, \cdots, X_n; Y) \\
= \; & H(X_1, X_2, \cdots, X_n) - H(X_1, X_2, \cdots, X_n | Y) \\
= \; & \left[H(X_1) + \sum_{i=2}^{n} H(X_i | X_1, \cdots, X_{i-1}) \right] \\
& - \left[H(X_1 | Y) + \sum_{i=2}^{n} H(X_i | X_1, \cdots, X_{i-1}, Y) \right] \\
= \; & [H(X_1) - H(X_1 | Y)] \\
& + \sum_{i=2}^{n} [H(X_i | X_1, \cdots, X_{i-1}) - H(X_i | X_1, \cdots, X_{i-1}, Y)] \\
= \; & I(X_1; Y) + \sum_{i=2}^{n} I(X_i; Y | X_1, X_2, \cdots, X_{i-1}).
\end{aligned}
$$

推论：柯尔莫哥洛夫公式

$$I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y | X_1) = I(X_2; Y) + I(X_1; Y | X_2).$$

练习：

考虑当随机变量序列构成马氏链时，联合熵、互信息的链式法则是什么形式

各种熵之间关系图

熵、联合熵、条件熵、互信息之间关系可以用文氏图来表示。

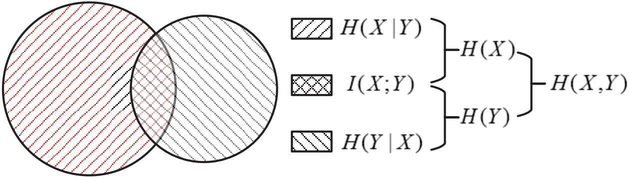

Figure: 信息熵图示

1.5