第一章随机变量的信息度量

第六节熵函数的凹凸性

陈兴同

中国矿业大学 数学学院

2021 年 8 月

内容提要

1. 两个凸集

内容提要

内容提要

构造两个凸集

先构造有限字符空间 $\mathcal{X},\mathcal{Y}$ 上两个概率分布集合.

（1）设 $\mathcal{P}$ 是定义在同一字符空间 $\mathcal{X}$ 上的全体概率分布集合:

$$\mathcal{P}=\left\{p=(p_1,p_2,\cdots,p_N)\middle|\sum_i p_i=1,p_i\geq 0,i=1,2,\cdots,N\right\}. \qquad 137 \quad \cancel{(1.1)}$$

（2）设 $\mathcal{Q}$ 是定义在字符空间 $\mathcal{X}\times\mathcal{Y}$ 上以 $p(y|x)$ 为元素的全体条件分布矩阵集合:

$$\mathcal{Q}=\{Q|Q\in\mathbb{R}^{N\times M}\text{是条件分布矩阵 (1-11)}\}. \qquad 138 \quad \cancel{(1.2)}$$

引理 1.6.1

对于（1-37）和（1-38）中定义的集合 $\mathcal{P}$ 和 $\mathcal{Q}$ 分别是空间 $\mathbb{R}^N$ 和 $\mathbb{R}^{N\times M}$ 中两个闭凸集。

定理 1.6.1：熵函数凹凸性

熵函数（1-22）是概率分布集合（1-37）上的凹函数。

证明：

设 $p = (p_1, p_2, \cdots, p_N)$ 及 $q = (q_1, q_2, \cdots, q_N) \in \mathcal{P}, \lambda \in [0, 1]$，则凸组合 $\lambda p + (1 - \lambda)q \in \mathcal{P}$。由于函数 $f(x) = x \log x$ 当 $x \geq 0$ 时是一个严格凸函数，故有

$\lambda f(p_i) + (1 - \lambda)f(q_i) \geq f(\lambda p_i + (1 - \lambda)q_i)$ 即

$$\lambda p_i \log p_i + (1 - \lambda)q_i \log q_i \geq [\lambda p_i + (1 - \lambda)q_i] \log [\lambda p_i + (1 - \lambda)q_i],$$

续证明:

两边求和得

$$\sum_{i=1}^{N}[\lambda p_i \log p_i + (1-\lambda) q_i \log q_i] \geq \sum_{i=1}^{N}[\lambda p_i + (1-\lambda) q_i] \log [\lambda p_i + (1-\lambda) q_i$$

两边同乘负号得

$$-\lambda \sum_{i=1}^{N} p_i \log p_i - (1-\lambda) \sum_{i=1}^{N} q_i \log q_i$$

$$\leq -\sum_{i=1}^{N}[\lambda p_i + (1-\lambda) q_i] \log [\lambda p_i + (1-\lambda) q_i],$$

$$\lambda H(p) + (1-\lambda) H(q) \leq H\left(\lambda p + (1-\lambda) q\right),$$

即 $H(p)$ 是凹函数。

定理 1.6.2：相对熵凹凸性

相对熵 $D(p||q)$ 是概率分布集合 (??) 中概率分布对 (p, q) 的凸函数，或者说是 $\mathcal{P} \times \mathcal{P}$ 上的凸函数。即如果 $(p^{(1)}, q^{(1)})$ 与 $(p^{(2)}, q^{(2)})$ 是 $\mathcal{P} \times \mathcal{P}$ 中两对概率分布，并且 $\lambda \in [0, 1]$，则有

1.37

$$
\begin{aligned}
&D(\lambda p^{(1)} + (1-\lambda)p^{(2)}||\lambda q^{(1)} + (1-\lambda)q^{(2)}) \\
&\leq \lambda D(p^{(1)}||q^{(1)}) + (1-\lambda)D(p^{(2)}||q^{(2)}).
\end{aligned}
$$

(1.39)

证明：

设 $a_1 = \lambda p_i^{(1)}, a_2 = (1-\lambda)p_i^{(2)}, b_1 = \lambda q_i^{(1)}, b_2 = (1-\lambda)q_i^{(2)}$，利用对数和不等式

$$(a_1 + a_2) \log \frac{a_1 + a_2}{b_1 + b_2} \le a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2},$$

可得

$$\left[\lambda p_i^{(1)} + (1-\lambda)p_i^{(2)}\right] \log \left(\frac{\lambda p_i^{(1)} + (1-\lambda)p_i^{(2)}}{\lambda q_i^{(1)} + (1-\lambda)q_i^{(2)}} \right)$$

$$\le \lambda p_i^{(1)} \log \frac{p_i^{(1)}}{q_i^{(1)}} + (1-\lambda)p_i^{(2)} \log \frac{p_i^{(2)}}{q_i^{(2)}}.$$

两边对 $i = 1, 2, \cdots, N$ 求和即得不等式（1-39）。

互信息公式变形

为了分析互信息的的凹凸性，将两个随机变量 X, Y 的互信息定义式（1-32）进行改写。

$$\begin{aligned}
I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p_X(x)p_Y(y)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_X(x)p(y|x)\log\frac{p(y|x)}{p_Y(y)} \\
&= \sum_{x\in\mathcal{X}} p_X(x)\sum_{y\in\mathcal{Y}} p(y|x)\log\frac{p(y|x)}{\sum_{x'\in\mathcal{X}} p_X(x')p(y|x')} \\
&= \sum_{i=1}^{N} p_X(x_i)\sum_{j=1}^{M} p(y_j|x_i)\log\frac{p(y_j|x_i)}{\sum_{k=1}^{N} p_X(x_k)p(y_j|x_k)} \\
&= \sum_{i=1}^{N} p_i\sum_{j=1}^{M} p_{j|i}\log\frac{p_{j|i}}{\sum_{k=1}^{N} p_k p_{j|k}},
\end{aligned}$$

续：互信息公式变形

这个表达式说明，互信息 $I(X;Y)$ 是概率分布 $p=(p_1,p_2,\cdots,p_N)$ 及条件分布矩阵 $Q=(p_{j|i})$ 的函数，因此可以写成

$$I(X;Y)=f(p,Q), p\in\mathcal{P}, Q\in\mathcal{Q}. \qquad \cancel{(3.1)}\ 1.40$$

定理 1.6.3：互信息凹凸性

（1） *对给定的条件分布矩阵* $Q \in \mathcal{Q}$，*由（1-40）确定的互信息* $f(p, Q)$ *是分布* $p \in \mathcal{P}$ *的凹函数。*

（2） *对给定的分布* $p \in \mathcal{P}$，*由（1-40）确定的互信息* $f(p, Q)$ *又是条件分布矩阵* $Q \in \mathcal{Q}$ *的凸函数。*

证明：

第一条证明： 任取两个分布 $p^{(1)}, p^{(2)}) \in \mathcal{P}$ 及 $\lambda \in [0,1]$，则 $p^{(\lambda)} = \lambda p^{(1)} + (1-\lambda)p^{(2)}) \in \mathcal{P}$；又由于条件分布矩阵 $Q \in \mathcal{Q}$ 已知，故由全概率公式可得字符空间 $\mathcal{Y}$ 上的三个分布：

$$q^{(1)} = p^{(1)}Q\text{，它的元素为} q_j^{(1)} = \sum_i p_i^{(1)} p_{j|i},$$

$$q^{(2)} = p^{(2)}Q\text{，它的元素为} q_j^{(2)} = \sum_i p_i^{(2)} p_{j|i},$$

$$q^{(\lambda)} = p^{(\lambda)}Q\text{，它的元素为} q_j^{(\lambda)} = \sum_i p_i^{(\lambda)} p_{j|i},$$

续证明：

并且 $q^{(\lambda)} = \lambda q^{(1)} + (1-\lambda)q^{(2)}$。从而根据互信息定义得

$$
\begin{aligned}
& f\left(\lambda p^{(1)} + (1-\lambda)p^{(2)}, Q\right) - \left[\lambda f(p^{(1)}, Q) + (1-\lambda) f(p^{(2)}, Q)\right] \\
&= f(p^{(\lambda)}, Q) - \left[\lambda f(p^{(1)}, Q) + (1-\lambda) f(p^{(2)}, Q)\right] \\
&= \sum_i \sum_j p_i^{(\lambda)} p_{j|i} \log \frac{p_{j|i}}{q_j^{(\lambda)}} - \lambda \sum_i \sum_j p_i^{(1)} p_{j|i} \log \frac{p_{j|i}}{q_j^{(1)}} \\
&\quad -(1-\lambda) \sum_i \sum_j p_i^{(2)} p_{j|i} \log \frac{p_{j|i}}{q_j^{(2)}} \\
&= \lambda D(q^{(1)} || q^{(\lambda)}) + (1-\lambda) D(q^{(2)} || q^{(\lambda)}) \geq 0,
\end{aligned}
$$

所以

$$
f(\lambda p_1 + (1-\lambda)p_2, Q) \geq \lambda f(p_1, Q) + (1-\lambda) f(p_2, Q).
$$

这就证明了第（1）条。

续证明：

第二条证明： 任取两个条件分布矩阵 $Q^{(1)}, Q^{(2)} \in \mathcal{Q}$ 及 $\lambda \in [0,1]$，则 $Q^{(\lambda)} = \lambda Q^{(1)} + (1-\lambda)Q^{(2)} \in \mathcal{Q}$; 并且可得：

（1）字符空间 $\mathcal{Y}$ 上的三个分布：

$$q^{(1)} = pQ^{(1)}，\text{它的元素为} q_j^{(1)} = \sum_{i=1}^{N} p_i p_{j|i}^{(1)},$$

$$q^{(2)} = pQ^{(2)}，\text{它的元素为} q_j^{(2)} = \sum_{i=1}^{N} p_i p_{j|i}^{(2)},$$

$$q^{(\lambda)} = pQ^{(\lambda)}，\text{它的元素为} q_j^{(\lambda)} = \sum_{i=1}^{N} p_i p_{j|i}^{(\lambda)},$$

以及 $q^{(\lambda)} = \lambda q^{(1)} + (1-\lambda)q^{(2)}$。

续证明：

（2）集合 $\mathcal{X} \times \mathcal{Y}$ 上的三个二维联合分布矩阵：

$$P^{(\lambda)} = \mathsf{diag}(p)Q^{(\lambda)}，\text{它的元素为} P_{ij}^{(\lambda)} = p_i p_{j|i}^{(\lambda)},$$

$$P^{(1)} = \mathsf{diag}(p)Q^{(1)}，\text{它的元素为} P_{ij}^{(1)} = p_i p_{j|i}^{(1)},$$

$$P^{(2)} = \mathsf{diag}(p)Q^{(2)}，\text{它的元素为} P_{ij}^{(2)} = p_i p_{j|i}^{(2)},$$

以及 $P^{(\lambda)} = \lambda P^{(1)} + (1-\lambda)P^{(2)}$。

续证明：

（3）集合 $\mathcal{X} \times \mathcal{Y}$ 上的三个乘积分布矩阵：

$$R^{(\lambda)} = p^{\mathsf{T}} q^{(\lambda)}，\text{它的元素为} R_{ij}^{(\lambda)} = p_i q_j^{(\lambda)},$$
$$R^{(1)} = p^{\mathsf{T}} q^{(1)}，\text{它的元素为} R_{ij}^{(1)} = p_i q_j^{(1)},$$
$$R^{(2)} = p^{\mathsf{T}} q^{(2)}，\text{它的元素为} R_{ij}^{(2)} = p_i q_j^{(2)},$$

以及 $R^{(\lambda)} = \lambda R^{(1)} + (1-\lambda) R^{(2)}$。

续证明：

从而根据互信息定义得

$$f\left(p, \lambda Q^{(1)} + (1-\lambda)Q^{(2)}\right) - \left[\lambda f(p, Q^{(1)}) + (1-\lambda) f(p, Q^{(2)})\right]$$

$$= f(p, Q^{(\lambda)}) - \left[\lambda f(p, Q^{(1)}) + (1-\lambda) f(p, Q^{(2)})\right]$$

$$= \sum_i \sum_j p_i p_{j|i}^{(\lambda)} \log \frac{p_{j|i}^{(\lambda)}}{q_j^{(\lambda)}} - \lambda \sum_i \sum_j p_i p_{j|i}^{(1)} \log \frac{p_{j|i}^{(1)}}{q_j^{(1)}}$$

$$-(1-\lambda) \sum_i \sum_j p_i p_{j|i}^{(2)} \log \frac{p_{j|i}^{(2)}}{q_j^{(2)}}$$

续证明：

从而根据互信息定义得

$$
\begin{aligned}
&= \sum_i \sum_j p_i p_{j|i}^{(\lambda)} \log \frac{p_i p_{j|i}^{(\lambda)}}{p_i q_j^{(\lambda)}} - \lambda \sum_i \sum_j p_i p_{j|i}^{(1)} \log \frac{p_i p_{j|i}^{(1)}}{p_i q_j^{(1)}} \\
&-(1-\lambda) \sum_i \sum_j p_i p_{j|i}^{(2)} \log \frac{p_i p_{j|i}^{(2)}}{p_i q_j^{(2)}} \\
&= \sum_i \sum_j P_{ij}^{(\lambda)} \log \frac{P_{ij}^{(\lambda)}}{R_{ij}^{(\lambda)}} - \lambda \sum_i \sum_j P_{ij}^{(1)} \log \frac{P_{ij}^{(1)}}{R_{ij}^{(1)}} \\
&-(1-\lambda) \sum_i \sum_j P_{j|i}^{(2)} \log \frac{P_{ij}^{(2)}}{Q_{ij}^{(2)}} \\
&= D\left(P^{(\lambda)} || R^{(\lambda)}\right) - \lambda D\left(P^{(1)} || R^{(1)}\right) - (1-\lambda) D\left(P^{(2)} || R^{(2)}\right) \leq 0,
\end{aligned}
$$

续证明：

最后的不等式是由于相对熵是概率分布对 $P^{(1)}, R^{(1)} \in \mathcal{Q}$ 和 $P^{(2)}, R^{(2)} \in \mathcal{Q}$ 的凸函数，所以

$$f(p, \lambda Q^{(1)} + (1-\lambda)Q^{(2)}) \leq \lambda f(p, Q^{(1)}) + (1-\lambda) f(p, Q^{(2)}).$$

这就证明了第（2）条。

练习：

证明引理 1.6.1。