

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

AEP 的推论：数据
压缩

高概率集与典型集

第 4 讲 渐近均分性

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

AEP 的推论：数据压缩

高概率集与典型集

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

AEP 的推论：数据
压缩

高概率集与典型集

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

AEP 的推论：数据压缩

高概率集与典型集

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

AEP 的推论：数据
压缩

高概率集与典型集

定理 1.1

设 $H(p_1, \dots, p_n)$ 为满足下列性质的函数:

- (1) H 在 p_i 上连续, 其中 $i = 1, \dots, n$.
- (2) 若对所有的 $i = 1, \dots, n$ 都有 $p_i = \frac{1}{n}$, 则 H 关于 n 单调递增.
- (3) (分组公理 / Grouping axiom) 对于 $P = (p_1, \dots, p_n)$, 令 $Q = (q_1, \dots, q_s)$ 定义为 $q_j \triangleq \sum_{i=n_{j-1}+1}^{n_j} p_i$ (其中 $j = 1, \dots, s$, 且 $0 = n_0 < n_1 < \dots < n_s = n$), 则:

$$H(P) = H(Q) + \sum_{j=1}^s q_j H\left(\frac{p_{n_{j-1}+1}}{q_j}, \dots, \frac{p_{n_j}}{q_j}\right).$$

那么 $H(p_1, \dots, p_n)$ 必然具有形如 $-K \cdot \sum_{i=1}^n p_i \cdot \log p_i$ 的形式, 其中 K 为常数因子.

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

AEP 的推论: 数据压缩

高概率集与典型集

我们简要讨论一下不确定性度量的这三个公理。前两条性质看起来相当合理：概率的微小变化应该对不确定性只有微小的影响；而具有 n 个可能结果的实验的不确定性，应该大于具有 $m < n$ 个可能结果的实验的不确定性。在极极端情况 $n = 1$ 中（例如实验结果已知），此时不存在不确定性。然而，第三条公理并不那么显而易见。我们可以用下面的例子来说明：假设我们在酒店里寻找一个人。设 $p_{n_{j-1}+i}$ （其中 $i \leq n_j - n_{j-1}$ ）表示此人在 j 楼 i 号房间的概率，那么 q_j 就是此人在 j 楼的概率。当我们已经知道此人在 j 楼时，寻找此人的剩余不确定性即为：

$$H\left(\frac{p_{n_{j-1}+1}}{q_j}, \dots, \frac{p_{n_j}}{q_j}\right).$$

辛钦 (Khinchin) 提出了以下一组公理, 它们同样可以在差一个常数因子的情况下确定熵:

(1') $H(p, 1-p)$ 在 p 上连续.

(2') $H(p_1, \dots, p_n)$ 是对称的 (对称性).

(3') $H(p_1, \dots, p_n) \leq H\left(\frac{1}{n}, \dots, \frac{1}{n}\right)$.

(4') (分组公理):

$$\begin{aligned} & H(q_{11}, \dots, q_{1m_1}, q_{21}, \dots, q_{2m_2}, \dots, q_{n1}, \dots, q_{nm_n}) \\ &= H(p_1, \dots, p_n) + \sum_{i=1}^n p_i \cdot H\left(\frac{q_{i1}}{p_i}, \dots, \frac{q_{im_i}}{p_i}\right) \end{aligned}$$

其中 $p_i = q_{i1} + \dots + q_{im_i}$.

这里的关键是用更弱的 “对称性” 性质替换了单调性. 辛钦的公理后来由法捷耶夫 (Faddeev) 进行了简化.

定理 1.2: (Faddeev 1956)

设 $H(p_1, \dots, p_n)$ 为满足下列性质的函数:

(1'') $H(p, 1-p)$ 对于 $0 \leq p \leq 1$ 是连续的.

(2'') $H(p_1, \dots, p_n)$ 对所有 $n \in \mathbb{N}$ 都是对称的.

(3'') 若 $p_n = q_1 + q_2 > 0$, 则:

$$H(p_1, \dots, p_{n-1}, q_1, q_2) = H(p_1, \dots, p_n) + p_n \cdot H\left(\frac{q_1}{p_n}, \frac{q_2}{p_n}\right).$$

则 $H(p_1, \dots, p_n) = -K \cdot \sum_{i=1}^n p_i \cdot \log p_i$, 其中常数因子 $K \in \mathbb{R}$.

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

AEP 的推论：数据压缩

高概率集与典型集

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

AEP 的推论：数据
压缩

高概率集与典型集

在信息论中, 与大数定律类似的是渐近均分性 (AEP), 它是弱大数定律的直接结果。大数定律针对独立同分布 (i.i.d.) 随机变量, 当 n 很大时, $\frac{1}{n} \sum_{i=1}^n X_i$ 近似于期望值 EX 。渐近均分性表明 $\frac{1}{n} \log \frac{1}{p(X_1, X_2, \dots, X_n)}$ 近似于熵 H , 其中 $X_1, X_2, \dots, X_n$ 为 i.i.d. 随机变量, $p(X_1, X_2, \dots, X_n)$ 是观察序列 $X_1, X_2, \dots, X_n$ 出现的概率。因而, 当 n 很大时, 一个观察序列出现的概率 $p(X_1, X_2, \dots, X_n)$ 近似等于 2^{-nH} 。

这促使我们将全体序列组成的集合划分成两个子集, 其一是典型集, 其中样本熵近似于真实熵; 其二是非典型集, 包含其余的序列。我们将主要关注典型集, 这是因为任何基于典型序列的性质都是以高概率成立的, 并且决定着大样本的平均行为。

首先举个例子。设随机变量 $X \in \{0, 1\}$ 的概率密度函数为 $p(1) = p$ 和 $p(0) = q$ 。若 $X_1, X_2, \dots, X_n$ 为 i.i.d., 且服从 $p(x)$, 则序列 $x_1, x_2, \dots, x_n$ 出现的概率为 $\prod_{i=1}^n p(x_i)$ 。比如, 序列 $(1, 0, 1, 1, 0, 1)$ 出现的概率是 $p^{\sum x_i} q^{n - \sum x_i} = p^4 q^2$ 。很显然, 并非所有长度为 n 的 2^n 个序列都具有相同的概率。

然而, 我们能够预测出实际观测到的序列的概率, 即可以求出观测结果 $X_1, X_2, \dots, X_n$ 的概率

$$p(X_1, X_2, \dots, X_n),$$

其中 $X_1, X_2, \dots, X_n$ 为 i.i.d. $\sim p(x)$ 。这是一个自引用的问题, 但仍然是可以明确定义的。显然, 我们是在寻求服从同一概率分布的事件的概率, 而结论是 $p(X_1, X_2, \dots, X_n)$ 将以高的概率接近于 2^{-nH} 。

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

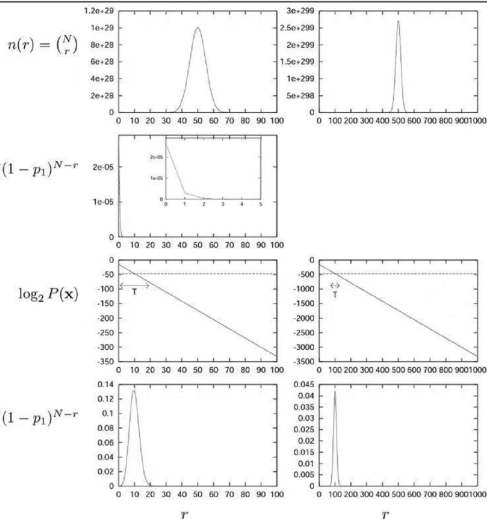
AEP 的推论: 数据压缩

高概率集与典型集

对此, 概括为 “几乎一切事件都令人同等的意外。” 换言之, 当 $X_1, X_2, \dots, X_n$ 为 i.i.d. $\sim p(x)$ 则

$$Pr\{(X_1, X_2, \dots, X_n) : p(X_1, X_2, \dots, X_n) = 2^{-n(H \pm \epsilon)}\} \approx 1$$

在这个例子中, $p(X_1, X_2, \dots, X_n) = p^{\sum X_i} q^{n - \sum X_i}$ 可以简单地说序列中 1 出现的个数近似等于 np (以很高的概率), 且所有这样的序列 (粗略地) 有相同的概率 $2^{-nH(p)}$.

$N = 100$ $N = 1000$ 

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

AEP 的推论：数据压缩

高概率集与典型集

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

AEP 的推论：数据压缩

高概率集与典型集

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

AEP 的推论：数据
压缩

高概率集与典型集

定义 3.1: 随机变量的收敛

设 $X_1, X_2, \dots$ 为一随机变量序列. 下面我们考虑序列 $\{X_n\}$ 收敛到 X 的三种不同方式:

1. 如果对任意的 $\epsilon > 0$, $P\{|X_n - X| > \epsilon\} \rightarrow 0$, 则称为依概率收敛.
2. 如果 $E(X_n - X)^2 \rightarrow 0$, 则称为均方收敛.
3. 如果 $P\{\lim_{n \rightarrow \infty} X_n = X\} = 1$, 则称为概率 1 (或称几乎处处) 收敛.

引理 3.2: 马尔可夫不等式

对任意非负随机变量 X 以及任意的 $t > 0$, 我们有

$$P\{X \geq t\} \leq \frac{EX}{t}. \quad (3.1)$$

证明.

如果 X 有分布 $F(x)$,

$$EX = \int_0^\infty x dF = \int_0^t x dF + \int_t^\infty x dF \geq \int_t^\infty x dF \geq \int_t^\infty t dF = tP\{X \geq t\}.$$

从而我们有 (3.1). □

引理 3.3: 切比雪夫不等式

设随机变量 Y 的均值和方差分别为 μ 和 σ^2 . 则有对任意 $\epsilon > 0$,

$$P\{|Y - \mu| > \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}.$$

证明.

在马尔可夫不等式中我们取 $X = (Y - \mu)^2$, 可得

$$\begin{aligned} P\{(Y - \mu)^2 > \epsilon^2\} &\leq P\{(Y - \mu)^2 \geq \epsilon^2\} \\ &= \frac{E(Y - \mu)^2}{\epsilon^2} \\ &= \frac{\sigma^2}{\epsilon^2}, \end{aligned}$$

又注意到 $P\{(Y - \mu)^2 > \epsilon^2\} = P\{|Y - \mu| > \epsilon\}$, 我们有

$$P\{|Y - \mu| > \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}.$$

□

定理 3.4: 弱大数定律

设 $Z_1, Z_2, \dots, Z_n$ 为 i.i.d. 随机变量序列, 其均值和方差分别为 μ 和 σ^2 , 令 $\bar{Z}_n = \frac{1}{n} \sum_{i=1}^n Z_i$ 为样本均值. 证明:

$$P\{|\bar{Z}_n - \mu| > \epsilon\} \leq \frac{\sigma^2}{n\epsilon^2}.$$

因此, 当 $n \rightarrow \infty$ 时, $P\{|\bar{Z}_n - \mu| > \epsilon\} \rightarrow 0$. 这就是著名的弱大数定律.

证明.

在切比雪夫不等式中取 $Y = \bar{Z}_n$, 且注意到 $E\bar{Z}_n = \mu$ 以及 $\text{Var}(\bar{Z}_n) = \frac{\sigma^2}{n}$ ($\bar{Z}_n$ 是 n 个独立同分布的随机变量 $\frac{Z_i}{n}$, 每个的方差为 $\frac{\sigma^2}{n^2}$), 我们有

$$P\{|\bar{Z}_n - \mu| > \epsilon\} \leq \frac{\sigma^2}{n\epsilon^2}.$$

从而得证. □

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

AEP 的推论：数据压缩

高概率集与典型集

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

AEP 的推论：数据
压缩

高概率集与典型集

定理 4.1: AEP

若 $X_1, X_2, \dots, X_n$ 为 i.i.d. $\sim p(x)$, 则

$$-\frac{1}{n} \log p(X_1, X_2, \dots, X_n) \rightarrow H(X) \text{ 依概率.}$$

证明.

独立随机变量的函数依然是独立随机变量. 因此, 由于 X_i 是 i.i.d., 从而 $\log p(X_i)$ 也是 i.i.d. 的. 从而由大数定律,

$$\begin{aligned} -\frac{1}{n} \log p(X_1, X_2, \dots, X_n) &= -\frac{1}{n} \sum_i \log p(X_i) \\ &\rightarrow -E \log p(X) \text{ 依概率} \\ &= H(X). \end{aligned}$$

这就证明了该定理.



定义 4.2

关于 $p(x)$ 的典型集 $A_\epsilon^{(n)}$ 是序列 $(x_1, x_2, \dots, x_n) \in \mathcal{X}^n$ 的集合, 且满足性质:

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \dots, x_n) \leq 2^{-n(H(X)-\epsilon)}.$$

作为渐近均分性的一个推论, 可以证明典型集 $A_\epsilon^{(n)}$ 有如下性质:

定理 4.3

1. 如果 $(x_1, x_2, \dots, x_n) \in A_\epsilon^{(n)}$, 则
$$H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \dots) \leq H(X) + \epsilon.$$
2. 当 n 充分大时, $P\{A_\epsilon^{(n)}\} > 1 - \epsilon.$
3. $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$, 其中 $|A|$ 表示集合 A 中的元素个数.
4. 当 n 充分大时, $|A_\epsilon^{(n)}| \geq (1 - \epsilon)2^{n(H(X)-\epsilon)}.$

证明.

性质 (1) 的证明可直接由 $A_\epsilon^{(n)}$ 的定义, 利用对数函数的单调性得到. 第二个性质由定理 4.1 可以得到. 为证明性质 (3), 我们有

$$\begin{aligned}
 1 &= \sum_{x \in \mathcal{X}^n} p(x) \\
 &\geq \sum_{x \in A_\epsilon^{(n)}} p(x) \\
 &\geq \sum_{x \in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} \\
 &= 2^{-n(H(X)+\epsilon)} |A_\epsilon^{(n)}|
 \end{aligned}$$

因此 $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$.



证明.

为证明 (4), 我们注意到对于充分大的 n , $P\{A_\epsilon^{(n)}\} > 1 - \epsilon$, 从而

$$\begin{aligned} 1 - \epsilon &< P\{A_\epsilon^{(n)}\} \\ &\leq \sum_{\mathbf{x} \in A_\epsilon^{(n)}} 2^{-n(H(X) - \epsilon)} \\ &= 2^{-n(H(X) - \epsilon)} |A_\epsilon^{(n)}|, \end{aligned}$$

其中第二个不等式由典型集的定义可得. 从而, $|A_\epsilon^{(n)}| \geq (1 - \epsilon)2^{n(H(X) - \epsilon)}$, 从而定理得证. $\square$

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

AEP 的推论：数据压缩

高概率集与典型集

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

AEP 的推论：数据
压缩

高概率集与典型集

设 $X_1, X_2, \dots, X_n$ 为 $\text{i.i.d} \sim p(x)$ 的随机变量. 我们将 $\mathcal{X}^n$ 中的所有序列划分成两个集合: 典型集 $A_\epsilon^{(n)}$ 及其补集. 我们现在要对 $\mathcal{X}^n$ 中的元素进行 01 编码, 我们希望能够有较短的平均码字长度.

将每个集合中的所有元素按某种顺序 (比如字典序) 排列. 然后给集合中的序列指定下标可以表示 $A_\epsilon^{(n)}$ 中的每个序列. 由于 $A_\epsilon^{(n)}$ 中序列个数 $\leq 2^{n(H+\epsilon)}$, 则表示它们需要的比特数不超过 $n(H+\epsilon) + 1$ (加 1 是因为 $n(H+\epsilon)$ 可能是非整数). 在所有这些序列前加上 0, 于是可知表示 $A_\epsilon^{(n)}$ 需要的比特数不超过 $n(H+\epsilon) + 2$. 类似地, 对不属于 $A_\epsilon^{(n)}$ 中的序列, 我们表示它们需要的比特数不超过 $n \log |\mathcal{X}| + 1$, 再在这些序列前加 1. 这样我们就获得了关于所有序列的一个编码方案.

注意, 上述编码方案有如下特征:

- ▶ 编码是 1-1 的, 且可以很方便地译码. 起始位作为标识位, 标明紧随码字的长度.
- ▶ 没有考虑非典型集的元素实际上少于 $\mathcal{X}^n$ 中元素的个数. 但这足以产生一个有效的描述. (后面我们会看到, 这是由于非典型集占的比例很小)
- ▶ 典型序列具有较短的描述长度.

下面用记号 x^n 表示序列 $x_1, x_2, \dots, x_n$. 设 $l(x^n)$ 表示相应于 x^n 的码字长度. 若 n 充分大, 使得 $P\{A_\epsilon^{(n)}\} \geq 1 - \epsilon$, 于是, 码字长度的数学期望为

$$\begin{aligned}
 E(l(X^n)) &= \sum_{x^n} p(x^n) l(x^n) \\
 &= \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) l(x^n) + \sum_{x^n \in (A_\epsilon^{(n)})^c} p(x^n) l(x^n) \\
 &\leq \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) (n(H + \epsilon) + 2) + \sum_{x^n \in (A_\epsilon^{(n)})^c} p(x^n) (n \log |\mathcal{X}| + 2) \\
 &= P\{A_\epsilon^{(n)}\} (n(H + \epsilon) + 2) + P(\{(A_\epsilon^{(n)})^c\}) (n \log |\mathcal{X}| + 2) \\
 &\leq n(H + \epsilon) + \epsilon n(\log |\mathcal{X}|) + 2\epsilon + 2 \\
 &= n(H + \epsilon').
 \end{aligned}$$

其中 $\epsilon' = \epsilon + \epsilon \log |\mathcal{X}| + \frac{2\epsilon + 2}{n}$, 适当选取 ϵ 和 n 时, ϵ' 可以任意小. 至此, 我们证明了如下的定理.

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

AEP 的推论: 数据压缩

高概率集与典型集

定理 5.1

设 X^n 为服从 $p(x)$ 的 i.i.d. 序列, $\epsilon > 0$, 则存在一个编码将长度为 n 的序列 x^n 映射为比特串, 使得映射是 1-1 的 (因而可逆), 且对于充分大的 n , 有

$$E\left[\frac{1}{n}l(X^n)\right] \leq H(X) + \epsilon.$$

上述定理说明, 在平均意义下, 我们可以用大约 $nH(X)$ 比特来表示序列 X^n .

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

AEP 的推论：数据压缩

高概率集与典型集

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

AEP 的推论：数据
压缩

高概率集与典型集

定义 6.1

对每个 $n = 1, 2, \dots$, 设 $B_\delta^{(n)} \subset \mathcal{X}^n$ 为满足如下条件的元素个数最小的集合:

$$P\{B_\delta^{(n)}\} \geq 1 - \delta.$$

定理 6.2

设 $X_1, X_2, \dots, X_n$ 为服从 $p(x)$ 的 i.i.d. 序列. 对于 $\delta < \frac{1}{2}$ 以及任意的 $\delta' > 0$, 如果 $P\{B_\delta^{(n)}\} > 1 - \delta$, 则对于充分大的 n , 有

$$\frac{1}{n} \log |B_\delta^{(n)}| > H - \delta'.$$

证明.

设 $A_\varepsilon^{(n)}$ 为一个典型集, 则 $P(A_\varepsilon^{(n)}) > 1 - \epsilon$, 从而 $P(A_\varepsilon^{(n)} \cap B_\delta^{(n)}) \geq 1 - \varepsilon - \delta$.
我们有

$$\begin{aligned}
 1 - \varepsilon - \delta &\leq P(A_\varepsilon^{(n)} \cap B_\delta^{(n)}) \\
 &= \sum_{A_\varepsilon^{(n)} \cap B_\delta^{(n)}} p(x^n) \\
 &\leq \sum_{A_\varepsilon^{(n)} \cap B_\delta^{(n)}} 2^{-n(H-\varepsilon)} \\
 &= |A_\varepsilon^{(n)} \cap B_\delta^{(n)}| 2^{-n(H-\varepsilon)} \\
 &\leq |B_\delta^{(n)}| 2^{-n(H-\varepsilon)}.
 \end{aligned}$$

熵的公理化定义

渐近均分性

大数定律

渐近均分性定理

AEP 的推论: 数据压缩

高概率集与典型集

证明.

前式两边取对数, 我们有

$$\frac{1}{n} \log |B_{\delta}^{(n)}| \geq (H - \varepsilon) + \frac{1}{n} \log(1 - \varepsilon - \delta).$$

取 ε 充分小, 以及 n 充分大, 使得 $\varepsilon - \frac{1}{n} \log(1 - \varepsilon - \delta) < \delta'$, 从而得证. □

下面引进一个新记号以表示在一阶指数意义下的相等概念.

定义 6.3: 记

$a_n \doteq b_n$ 表示

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0.$$

下面引进一个新记号以表示在一阶指数意义下的相等概念.

定义 6.4: 记

$a_n \doteq b_n$ 表示

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0.$$

于是, 我们有: 如果 $\delta_n \rightarrow 0$ 且 $\epsilon_n \rightarrow 0$, 则

$$|B_{\delta_n}^{(n)}| \doteq |A_{\epsilon_n}^{(n)}| \doteq 2^{nH}.$$

例 6.5

为说明 $A_\epsilon^{(n)}$ 与 $B_\delta^{(n)}$ 的区别, 考虑一个 Bernoulli 序列 $X_1, X_2, \dots, X_n$, 其参数 $p = 0.9$. 此时, 典型序列元素中 1 所占的比例近似等于 0.9. 然而, 这并不包括全部都是 1 的序列, 虽然其出现的概率最大. 集合 $B_\delta^{(n)}$ 包含所有很可能出现的序列, 因而包括全部为 1 的序列.