《信息论基础》[1] 笔记

纳文琪 [1]

2018 年 12 月 11 日

1 熵、相对熵与互信息

1.1 熵

熵 是随机变量不确定度的度量，离散型随机变量 X 的熵定义为:

$$
\begin{aligned}
H(X) &= -\sum_{x\in X} p(x)\log p(x) \\
&= -E\log p(x)
\end{aligned} \tag{1}
$$

熵不依赖于 X 的实际取值，值依赖于其概率分布。性质:

$$H(X) \geq 0 \tag{2}$$

1.2 联合熵与条件熵

联合熵 与联合分布相对应，定义为:

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) \\
&= -E\log p(x,y)
\end{aligned} \tag{3}
$$

条件熵 与条件概率对应，定义为:

$$
\begin{aligned}
H(Y|X) &= \sum_{x\in X} p(x)H(Y|X=x) \\
&= E\log p(Y|X)
\end{aligned} \tag{4}
$$

链式法则 定位为:

$$H(X,Y) = H(X) + H(Y|X) \tag{5}$$

其含义是：一对随机变量的熵等于其中一个随机变量的熵加上另一个随机变量的条件熵。

熵的链式法则

$$H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1) \tag{6}$$

1.3 相对熵与互信息

相对熵 (relative entropy, KL **距离**) 是两个随机分布之间距离的度量，也就是说它表示两个随机变量之间的距离。定义为:

$$\begin{aligned} D(p \parallel q) &= -\sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \\ &= E \log \frac{p(X)}{q(X)} \end{aligned} \tag{7}$$

相对熵总是非负的，且当且仅当 $p = q$ 时为零。它不对称，也不满足三角不等式，但将其视作分布之间的“距离”往往会很有用。

自信息 ?

互信息 是一个随机变量包含另一个随机变量信息量的度量。定义为:

$$\begin{aligned} I(X;Y) &= -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= D(p(x,y) \parallel p(x)p(y)) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned} \tag{8}$$

互信息 I(X;Y) 是在给定 Y 的知识的条件下 X 的不确定度 (熵) 的缩减量; 而且，X 含有 Y 的信息量等同于 Y 含有 X 的信息量。

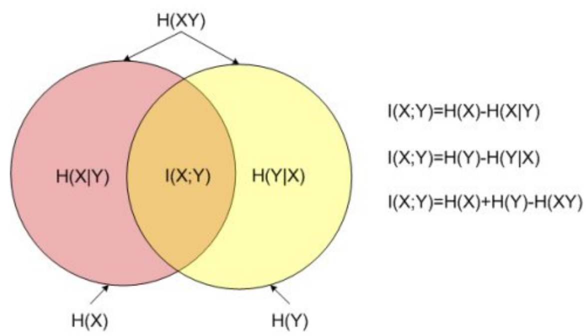

图 1: 熵与互信息的关系

互信息的链式法则

$$I(X_1, X_2, \cdots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \cdots, X_1) \tag{9}$$

条件相对熵

相对熵的链式法则

2 数据压缩

信源编码 关于随机变量 X 的信源编码 C，是从 X 的取值空间 X 到 D^* 的一个映射，其中 D^* 表示 D 元字幕表上有限长度的字符串所构成的集合。

信源编码的期望长度

$$L(C) = E[l(x)] \tag{10}$$

其中 $l(x)$ 是 x 的码字长度。

3 信道容量

离散信道 是由输入字母表 X、输出字母表 Y 和概率转移矩阵 $p(y|x)$ 构成的系统，其中 $p(y|x)$ 表示发送字符 x 的条件下收到字符 y 的概率。

离散无记忆信道（DMC） 如果离散信道输出的概率分布仅依赖于它所对应的输入，而与先前信道的输入或输出条件独立，则称信道是独立的。

DMC 的信道容量　定义为:

$$C = \max_{p(x)} I(X;Y) \tag{11}$$

3.1　信道容量的几个例子

无噪声二元信道　任何一个传输的比特都能被无误差第接收到。信道容量 $C = 1$

二元对称信道 (BSC)　这个信道的输入字符以概率 p 互补。其信道容量是 $C = 1 - H(p)$

码率　如果输入序列的长度为 k，而输出序列的长度为 n，则码率为:

$$r = \frac{k}{n}, (r < 1) \tag{12}$$

消息 (Word)　是一个字符序列。

码字 (Codewords)　是由消息组成的向量。(?)

码 (Code)　是一组码字的集合。

信道的 (M,n) **码**　指的是将 M 个消息生成长度为 n 的码字的一种编码。

(M,n) **码的码率**　为:

$$R = \frac{\log M}{n} \tag{13}$$

单位为“bits/传输”。

Example 3.4　The code $C = 00000, 10100, 11110, 11001$ is a block code of block length n = 5. Here q = 2(binary code), n = 5 and therefore, k = 2 and M = 4. Since there 4 codewords in the code, it can be used to represent two bit binary numbers. Such as:

Uncoded bits	Codewords
00	00000
01	10100
10	11110
11	11001

汉明距离　它表示两个（相同长度）字对应位不同的数量。

线性码　具有以下属性:

- 两个码字的和同样属于码
- 全 0 消息也是一个码字
- 两个码字的最小汉明距离与任何非零码字的最小权重相等，即 $d^* = w^*$

信道编码定理 (香农第二定理)　对于 DMC，小于信道容量 C 的所有码率都是可达的。具体来说，对任意码率 $R < C$，存在一个 $(2^{nR}, n)$ 码序列，它的最大误差概率为 $\lambda^{(n)} \to 0$。

重复码 (Repetition Code**)**　中每个输入被简单重复 n 次，n 为奇数。例如，n=3 时，有:
$0 \to 000, 1 \to 111$。它的码率为: $r = \frac{1}{n}$。其解码策略是多数解码（Majority Decoding），即当接收到的 n 个比特信息中 0 的个数大于 1 的个数的时候，就解码为 0，反之亦然。

信道容量定理　频率为 WHz，被谱密度为 $N_0/2$ 的附加白噪声（AWGN）干扰的连续信道的信道容量是:

$$C = W \log_2(1 + \frac{P}{N_0 W}) bits/second \tag{14}$$

P 是平均传输功率。

3.2　其他

香农限制 (Shannon Limit)

Bandwidth Eifficiency Diagram

域　是一个可以在其上进行加法、减法、乘法和除法运算而结果不会超出域的集合。

有限域　是仅含有限个元素的域。有限域中元素的个数称为有限域的阶。每个有限域的阶必为素数的幂，即有限域的阶可表示为 p^n（p 是素数、n 是正整数），该有限域通常称为 Galois 域 (Galois Fields)，记为 $GF(p^n)$。

参考文献

[1] T.M. Cover and J.A. Thomas, 信息论基础, 电子与电气工程丛书. 机械工业出版社, 2005. 1