第八章 信息论基础

第 23 讲 信息论基础及其在机器学习中的应用

黄定江

DaSE @ ECNU

djhuang@dase.ecnu.edu.cn

1. 23.1 熵、相对熵和互信息
2. 23.2 连续分布的微分熵和最大熵
3. 23.3 信息论在数据科学中的应用

信息论的基本问题

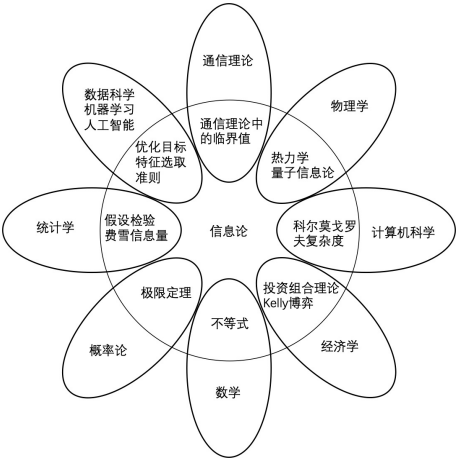
信息论是通信的基础理论之一。它解答了通信理论中的两个基本问题:

- 临界数据压缩的值
- 临界通信传输速率的值

然而，信息论的影响力远不止于此，它在很多学科都具有奠基性的贡献：

- 统计物理（热力学）
- 计算机科学（Kolmogorov 复杂度或算法复杂度）
- 统计推断（奥卡姆剃刀：最简洁的解释最佳）
- 概率和统计等学科（关于最优化假设检验和估计的误差指数）

信息论与其它学科的关系图



信息论与机器学习的关系

信息论和机器学习就是一枚硬币的两面。——David MacKay, information theory，inference and learning algorithms

在机器学习中一般有两类学习准则：

- 一类是如经验风险、经验误差、经验损失的经验函数。
- 另一类是如信息熵、交叉熵、相对熵、互信息的基于信息论中熵的函数。

信息论指导机器学习中的很多算法的设计和改进，比如用交叉熵损失作为损失函数，利用互信息进行特征选择等。以信息理论为基础的机器学习在理论上更具有优势:

- 我们可以将机器学习或深度模型中的编码解码模型看做是一个通信系统，输入为信源，这样信息论中的一些度量也可以作为学习算法的度量。
- 信息瓶颈理论认为深度学习模型具有特征拟合和特征压缩两个阶段，用互信息评价特征的保真度和压缩率。

学习就是一个熵减的过程。——Shun Watanabe

本讲主要关心的问题

人咬狗是新闻，狗咬人不是新闻，除非一条狗咬了很多人。

- 如何度量事件 (消息) 包含的信息量？
- 如何度量事件给出的关于另一个事件的信息量？

信息论的奠基之作和经典之作:

- 《通信的数学原理》(1948, Claude Elwood Shannon)
- 《噪声下的通信》(1949, Claude Elwood Shannon)
- 《信息论基础》(1990, Thomas M.Cover, Joy A.Thomas)

这里涉及几个基础概念：

- 香农信息：信息是对事物运动状态或存在方式的不确定性的描述.
- 消息：随机事件
- 信源：随机变量

信息论主要研究的是对一个信号包含信息的多少进行量化。

- 事件的信息量应当是事件发生概率的函数。
- 小概率事件, 不确定性大, 一旦出现使人感到意外, 因此产生的信息量就大, 特别是几乎不可能出现的事件一旦出现, 必然产生极大的信息量;
- 大概率事件, 是预料之中的事件, 不确定性小, 即使发生, 也没什么信息量, 特别是概率为 1 的确定事件发生以后, 不会给人以任何信息量。

信息是个相当宽泛的概念，很难用一个简单的定义将其完全准确地把握，然而对于任何一个概率分布可以定一个称为熵的量。它具有许多特性符合度量信息的直观要求，这个概念可以推广到互信息，互信息是一种测度，用来度量一个随机变量包含另一个随机变量的信息量。熵恰好变成了一个随机变量的自信息。相对熵是个更广泛的量，它是刻画两个概率分布之间的距离的一种度量，而互信息又是它的特殊情形，以上所有的这些量密切相关，存在许多简单的共性。

23.1.1 自信息和熵

自信息

随机事件的自信息量 $I(x_i)$ 是该事件发生概率 $p(x_i)$ 的函数, 并且 $I(x_i)$ 应该满足以下公理化条件:

- $I(x_i)$ 是 $p(x_i)$ 的严格递减函数. 当 $p(x_1) < p(x_2)$ 时,$I(x_1) > I(x_2)$, 概率越小, 事件发生的不确定性越大, 事件发生以后所包含的自信息量越大.
- 极限情况下, 当 $p(x_i) = 0$ 时, $I(x_i) \to \infty$; 当 $p(x_i) = 1$ 时,$I(x_i) = 0$。
- 由两个相对独立的不同的消息所提供的信息量应等于它们分别提供的信息量之和, 即自信息量满足可加性.$f(p_1 p_2) = f(p_1) + f(p_2)$

可以证明, 满足以上公理化条件的函数形式是对数形式.

定义 1

随机事件的自信息量定义为该事件发生概率的对数的负值. 设事件 x_i 的概率为 $p(x_i)$, 则它的自信息量定义为

$$I(x_i) = -\log p(x_i) = \log \frac{1}{p(x_i)}$$

$I(x_i)$ 代表两种含义: 在事件 x_i 发生以前, 等于事件 x_i 发生的不确定性的大小; 在事件 x_i 发生以后, 表示事件 x_i 所含有或所能提供的信息量。

自信息量的单位与所用对数的底有关.

比特: 通常取对数的底为 2 , 信息量的单位为比特 (bit). 比特是信息论中最常用的信息量单位, 当取对数的底为 2 时, 2 常省略.

奈特: 取自然对数 (以 e 为底) , 自信息量的单位为奈特 (nat) . 理论推导中或用于连续信源时用以 e 为底的对数比较方便.

$$1 \text{ nat} = \log_2 e \text{ bit} = 1.443 \text{ bit}$$

哈特莱: 工程上用以 10 为底较方便. 若以 10 为对数底, 则自信息量的单位为哈特莱 (Hartley) , 用来纪念哈特莱首先提出用对数来度量信息.

$$1 \text{ Hartley} = \log_2 10 \text{ bit} = 3.322 \text{ bit}$$

一般:

$$1 \ r\text{进制单位} = \log_2 r \text{ bit}$$

例 1

(1) 英文字母中“a”出现的概率为 0 .064 ,“c”出现的概率为 0.022 , 分别计算它们的自信息量.

(2) 假定前后字母出现是互相独立的, 计算“ac”的自信息量.

(3) 假定前后字母出现不是互相独立的, 当"a"出现以后,"c"出现的概率为 0.04 , 计算"a"出现以后,"c"出现的自信息量.

(1) 英文字母中“a”出现的概率为 0 .064，“c”出现的概率为 0.022，分别计算它们的自信息量.
(2) 假定前后字母出现是互相独立的, 计算“ac”的自信息量.
(3) 假定前后字母出现不是互相独立的, 当"a”出现以后,"c”出现的概率为 0.04，计算"a”出现以后,"c”出现的自信息量.

解

(1) $I(a) = -\log_2 0.064 = 3.96bit \qquad I(c) = -\log_2 0.022 = 5.51bit$

(2) 由于前后字母出现是互相独立的,“ac”出现的概率为 0.064×0.022，所以

$$I(ac) = -\log_2(0.064 \times 0.022) = -(\log_2 0.064 + \log_2 0.022)$$

$$= I(a) + I(c) = 9.47bit$$

(3)“a”出现的条件下“, c”出现的概率变大, 它的不确定性变小.

$$I(c|a) = -\log_2 0.04 = 4.64bit$$

定义 2

设 x_i, y_j 是两个随机事件。x_i 和 y_j 的联合自信息 *(joint self-information)* 定义为积事件 $x_i y_j$ 的自信息，即

$$I(x_i y_j) = -\log p(x_i y_j)$$

是两个事件所提供的总的信息量。

定义 3

设 x_i, y_j 是两个随机事件。x_i 和 y_j 的条件自信息 *(conditional self-information)* 定义为条件事件 $x_i|y_j$ 的自信息，即

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

是在已知条件 y_j 时 x_i 所含的新的信息量。

熵

自信息量:

信源发出某一具体消息所含有的信息量

平均自信息量:

表征整个信源的不确定度. 又称为信息熵、信源熵, 简称熵.

因为信源具有不确定性, 所以把信源用随机变量来表示, 用随机变量的概率分布来描述信源的不确定性.

概率空间：

通常把一个随机变量的所有可能的取值和这些取值对应的概率 $[X, p(X)]$ 称为它的概率空间.

假设随机变量 X 有 q 个可能的取值 x_i ,$i = 1, 2, \cdots, q$, 各种取值出现的概率为 $p(x_i)$, $i = 1, 2, \cdots, q$, 它的概率空间表示为

$$\begin{pmatrix} X \\ p(X) \end{pmatrix} = \begin{pmatrix} X = x_1 & \cdots & X = x_i & \cdots & X = x_q \\ p(x_1) & \cdots & p(x_i) & \cdots & p(x_q) \end{pmatrix}$$

这里要注意，$p(x_i)$ 满足概率空间的基本特性: 非负性 $0 \leqslant p(x_i) \leqslant 1$ 和完备性 $\sum_{i=1}^{q} p(x_i) = 1$。

定义 4

随机变量 X 的每一个可能取值的自信息 $I(x_i)$ 的统计平均值定义为随机变量 X 的信息熵.

$$H(X) = \mathbb{E}[I(x_i)] = -\sum_{i=1}^{q} p(x_i) \log p(x_i)$$

这里 q 为 X 的所有可能取值的个数。

熵的单位也是与所取的对数底有关, 根据所取的对数底不同, 可以是比特/ 符号、奈特/ 符号、哈特莱/ 符号或者是 r 进制单位/ 符号, 通常用比特/ 符号为单位.

例 2

假设随机变量 X 的概率分布为 $p(x_i) = 2^{-i}, i = 1, 2, 3, \cdots$，求 $H(X)$。

解

$$H(X) = \sum_{i=1}^{\infty} 2^{-i} \log_2 \frac{1}{2^{-i}} = \sum_{i=1}^{\infty} i 2^{-i} = 2 \text{比特/符号}$$

熵编码

- 信息论的研究目标之一是如何用最少的编码表示传递信息. 假设我们要传递一段文本信息，这段文本中包含的符号都来自于一个字母表 $\mathbb{A}$，我们就需要对字母表 $\mathbb{A}$ 中的每个符号进行编码. 以二进制编码为例，我们常用的 ASCII 码就是用固定的 8 比特来编码每个字母. 但这种固定长度的编码方案不是最优的. 一种高效的编码原则是字母的出现概率越高，其编码长度越短. 比如对字母 a, b, c 分别编码为 0,10,110.
- 给定一串要传输的文本信息，其中字母 x 的出现概率为 $p(x)$，其最佳编码长度为 $-\log_2 p(x)$，整段文本的平均编码长度为 $-\sum_x p(x)\log_2 p(x)$，即底为 2 的熵.
- 在对分布 $p(x)$ 的符号进行编码时，熵 $H(p)$ 也是理论上最优的平均编码长度，这种编码方式称为熵编码（Entropy Encoding）.
- 由于每个符号的自信息通常都不是整数，因此在实际编码中很难达到理论上的最优值. 霍夫曼编码（Huffman Coding）和算术编码（Arithmetic Coding）是两种最常见的熵编码技术。

熵函数

熵函数

信息熵 $H(X)$ 是随机变量 X 的概率分布的函数, 所以又称为熵函数.

如果把概率分布 $p(x_i), i = 1, 2, \cdots, q$, 记为 $p_1, p_2, \cdots, p_q$, 则熵函数又可以写成概率向量 $p = (p_1, p_2, \cdots, p_q)$ 的函数形式, 记为 $H(\boldsymbol{p})$.

$$H(X) = -\sum_{i=1}^{q} p(x_i) \log p(x_i) = H(p_1, p_2, \cdots, p_q) = H(\boldsymbol{p})$$

因为概率空间的完备性, 即 $\sum_{i=1}^{q} p(x_i) = 1$, 所以 $H(\boldsymbol{p})$ 是 $(q-1)$ 元函数.

当 $q = 2$ 时, 因为 $p_1 + p_2 = 1$, 若令其中一个概率为 p, 则另一个概率为 $(1-p)$, 熵函数可以写成 $H(p)$.

熵函数的性质

性质 1

对称性

$$H(p_1, p_2, \cdots, p_q) = H(p_2, p_1, \cdots, p_q) = \cdots = H(p_q, p_1, \cdots, p_{q-1})$$

也就是说概率向量 $p = (p_1, p_2, \cdots, p_q)$ 各分量的次序可以任意变更, 熵值不变。对称性说明熵函数仅与信源的总体统计特性有关。

性质 2

确定性

$$H(1, 0) = H(1, 0, 0) = H(1, 0, 0, 0) = \cdots = H(1, 0, \cdots, 0) = 0$$

在概率向量 $p = (p_1, p_2, \cdots, p_q)$ 中, 只要有一个分量为 1 , 其他分量必为 0 , 它们对熵的贡献均为 0 , 因此熵等于 0, 也就是说确定信源的平均不确定度为 0 。

性质 3

非负性

$$H(\boldsymbol{p}) = H(p_1, p_2, \cdots, p_q) \geq 0$$

对确定信源, 等号成立.

信源熵是自信息的数学期望, 自信息是非负值, 所以信源熵必定是非负的. 离散信源熵才有这种非负性, 连续信源的相对熵则可能出现负值.

性质 4

扩展性

$$\lim_{\epsilon \to 0} H_{q+1}(p_1, p_2, \cdots, p_q - \epsilon, \epsilon) = H_q(p_1, p_2, \cdots, p_q)$$

这是因为 $\lim_{\epsilon \to 0} \epsilon \log \epsilon = 0$

这个性质的含义是: 增加一个基本不会出现的小概率事件, 信源的熵保持不变. 虽然小概率事件出现给予收信者的信息量很大, 但在熵的计算中, 它占的比重很小, 可以忽略不计, 这也是熵的总体平均性的体现.

性质 5

连续性

$$\lim_{\epsilon \to 0} H(p_1, p_2, \cdots, p_{q-1} - \epsilon, p_q + \epsilon) = H(p_1, p_2, \cdots, p_q)$$

即信源概率空间中概率分量的微小波动, 不会引起熵的变化.

性质 6

递增性

$$H(p_1, p_2, \cdots, p_{n-1}, q_1, q_2, \cdots, q_m) =$$
$$H(p_1, p_2, \cdots, p_n) + p_n H(\frac{q_1}{p_n}, \frac{q_2}{p_n}, \cdots, \frac{q_m}{p_n})$$

这个性质表明, 假如有一信源的 n 个元素的概率分布为 $p_1, p_2, \cdots, p_n$, 其中某个元素 x_n 又被划分成 m 个元素, 这 m 个元素的概率之和等于元素 x_n 的概率, 这样得到的新信源的熵增加了一项, 增加的一项是由于划分产生的不确定性.

例 3

利用递增性计算 $H(1/2, 1/8, 1/8, 1/8, 1/8)$.

解

$$
\begin{aligned}
& H(1/2, 1/8, 1/8, 1/8, 1/8) \\
=& H(1/2, 1/2) + \frac{1}{2} \times H(1/4, 1/4, 1/4, 1/4) \\
=& 1 + \frac{1}{2} \times 2 \\
=& 2\text{比特/符号}
\end{aligned}
$$

命题 1

极值性

$$H(p_1, p_2, \cdots, p_n) \leq H(\frac{1}{n}, \cdots, \frac{1}{n}) = \log n \tag{1}$$

式中 n 是随机变量 X 的可能取值的个数.

极值性表明离散信源中各消息等概率出现时熵最大, 这就是最大离散熵定理. 连续信源的最大熵则还与约束条件有关.

极值性可看成

$$H(p_1, p_2, \cdots, p_n) \leqslant -\sum_{i=1}^{n} p_i \log_2 q_i \tag{2}$$

的特例情况.

下面先证明式(2)

证明.

利用 Jensen 不等式, 有

$$H(p_1,p_2,\cdots,p_n)+\sum_{i=1}^{n}p_i\log_2 q_i$$

$$=-\sum_{i=1}^{n}p_i\log_2 p_i+\sum_{i=1}^{n}p_i\log_2 q_i=\sum_{i=1}^{n}p_i\log_2\frac{q_i}{p_i}\leqslant\log_2\sum_{i=1}^{n}\left(p_i\cdot\frac{q_i}{p_i}\right)=0$$

当 $\frac{q_i}{p_i}=1,\quad i=1,2,\cdots,n$ 时, 等号成立。证毕。 □

式(2)表明: 任一随机变量的概率分布 p_i, 对其他概率分布 q_i 定义的自信息 $-\log_2 q_i$ 的数学期望, 必不小于概率分布 p_i 本身定义的熵 $H(p_1,p_2,\cdots,p_n)$.

证明.

如果取 $q_i = \frac{1}{n}, i = 1, 2, \cdots, n$ 时，由式(2)就得到

$$H(p_1, p_2, \cdots, p_n) \leq H(\frac{1}{n}, \cdots, \frac{1}{n}) = \log n$$

当 $p_i = \frac{1}{n}, i = 1, 2, \cdots, n$ 时，等号成立。 □

信息量的单位比特和计算机术语中位的单位比特的关系:

当信源输出的消息等概分布时, 信源熵达到最大值——1 比特/ 符号.

因此当二元数字是由等概的二元信源输出时，每个二元数字提供 1 bit 的信息量. 否则, 每个二元数字提供的信息量小于 1 bit .

命题 2

上凸性

$H(\boldsymbol{p})$ 是严格的上凸函数，设 $\boldsymbol{p}=(p_1,p_2,\cdots,p_q)$, $\boldsymbol{p}'=(p'_1,p'_2,\cdots,p'_q)$, $\sum_{i-1}^{q}p_i=1$, $\sum_{i-1}^{q}p'_i=1$, 则对于任意小于 1 的正数 α,$0<\alpha<1$, 以下不等式成立:

$$H[\alpha\boldsymbol{p}+(1-\alpha)\boldsymbol{p}']>\alpha H(\boldsymbol{p})+(1-\alpha)H(\boldsymbol{p}')$$

证明.

由 $0 \leqslant p_i \leqslant 1, 0 \leqslant p'_i \leqslant 1$, 且 $0 < \alpha < 1$，所以 $0 \leqslant \alpha p_i + (1-\alpha)p'_i \leqslant 1$，得 $\sum_{i=1}^{q} (\alpha p_i + (1-\alpha)p'_i) = 1$，所以 $\alpha p + (1-\alpha)p'$ 可以看作时一种新的概率分布。

$$H(\alpha p + (1-\alpha)p') = -\sum_{i=1}^{q} (\alpha p_i + (1-\alpha)p'_i) \log (\alpha p_i + (1-\alpha)p'_i)$$

$$= -\alpha \sum_{i=1}^{q} p_i \log (\alpha p_i + (1-\alpha)p'_i) - (1-\alpha) \sum_{i=1}^{q} p'_i \log_2 (\alpha p_i + (1-\alpha)p'_i)$$

$$\geqslant -\alpha \sum_{i=1}^{q} p_i \log_2 p_i - (1-\alpha) \sum_{i=1}^{q} p'_i \log_2 p'_i$$

$$\geqslant \alpha H(p) + (1-\alpha) H(p')$$

证明续.

当 $p \neq p'$ 时，有 $\frac{\alpha p_i+(1-\alpha)p_i'}{p_i} \neq 1$，式(2)中等号不成立，所以

$$H(\alpha p + (1-\alpha)p') > \alpha H(p) + (1-\alpha)H(p') \tag{3}$$

成立。证毕。 □

上凸函数在定义域内的极值必为极大值, 可以利用熵函数的这个性质证明熵函数的极值性.

23.1.2 联合熵和条件熵

一个随机变量的不确定性可以用熵来表示，这一概念可以方便地推广到多个随机变量。

二维随机变量 XY 的概率空间表示为

$$\begin{bmatrix} XY \\ p(XY) \end{bmatrix} = \begin{bmatrix} x_1y_1 & \cdots & x_iy_j & \cdots & x_ny_n \\ p(x_1y_1) & \cdots & p(x_iy_j) & \cdots & p(x_ny_n) \end{bmatrix}$$

其中，$p(x_iy_j)$ 满足概率空间的非负性和完备性：$0 \le p(x_iy_j) \le 1$，$\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_iy_j) = 1$。

联合熵

定义 5

[联合熵] 二维随机变量 XY 的联合熵定义为联合自信息的数学期望，它是二维随机变量 XY 的不确定性的度量。

$$H(XY) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i y_j) I(x_i y_j) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i y_j) \log p(x_i y_j)$$

条件熵

定义 6

[条件熵] 考虑在给定 $X = x_i$ 的条件下，随机变量 Y 的不确定性为

$$H(Y|x_i) = -\sum_j p(y_j|x_i) \log p(y_j|x_i)$$

对 $H(Y|x_i)$ 的所有可能值进行统计平均，就得出给定 X 时，Y 的条件熵 $H(Y|X)$：

$$\begin{aligned} H(Y|X) &= \sum_i p(x_i) H(Y|x_i) \\ &= -\sum_i \sum_j p(x_i) p(y_j|x_i) \log p(y_j|x_i) \\ &= -\sum_i \sum_j p(x_i y_j) \log p(y_j|x_i) \end{aligned}$$

性质 7

联合熵和条件熵有如下关系：

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

证明.

$$\begin{aligned} H(XY) =& \mathbb{E}(\log \frac{1}{p(xy)}) = \mathbb{E}(\log \frac{1}{p(x)p(y|x)}) \\ =& \mathbb{E}(\log \frac{1}{p(x)} - \log \frac{1}{p(y|x)}) \\ =& \mathbb{E}(\log \frac{1}{p(x)}) + \mathbb{E}(\log \frac{1}{p(y|x)}) \\ =& H(X) + H(Y|X) \end{aligned}$$

□

推论 1

当二维随机变量 X，Y 相互独立时，联合熵等于 X 和 Y 各自熵之和：

$$H(XY) = H(X) + H(Y)$$

证明.

因为随机变量 X, Y 相互独立，所以有 $p\left(x_i y_j\right) = p\left(x_i\right) p\left(y_j\right)$

$$\begin{aligned} H(XY) =& E[-\log_2 p(xy)] \\ =& E[-\log_2 p(x)p(y))] \\ =& E[-\log_2 p(x) - \log_2 p(y))] \\ =& E[-\log_2 p(x)) + E[-\log_2 p(y)] \\ =& H(X) + H(Y) \end{aligned}$$

□

23.1.3 互信息和相对熵

事件对事件的互信息

定义 7

一个事件 y_j 所给出关于另一个事件 x_i 的信息定义为互信息，用 $I(x_i; y_j)$ 表示.

$$I\left(x_i; y_j\right) = I(x_i) - I\left(x_i|y_j\right) = \log_2 \frac{p\left(x_i|y_j\right)}{p\left(x_i\right)} \tag{4}$$

互信息 $I(x_i; y_j)$ 是已知事件 y_j 后所消除的关于事件 x_i 的不确定性，它等于事件 x_i 本身的不确定性 $I(x_i)$ 减去已知事件 y_j 后对 x_i 仍然存在的不确定性 $I(x_i|y_j)$. 互信息的引出，使信息的传递得到了定量的表示.

例 4

某地二月份天气出现的概率分别为晴 $1/2$，阴 $1/4$，雨 $1/8$，雪 $1/8$. 某一天有人告诉你：“今天不是晴天”，把这句话作为收到的消息 y_1，求收到 y_1 后，y_1 与各种天气的互信息量.

某地二月份天气出现的概率分别为晴 $1/2$，阴 $1/4$，雨 $1/8$, 雪 $1/8$. 某一天有人告诉你: “今天不是晴天”，把这句话作为收到的消息 y_1，求收到 y_1 后, y_1 与各种天气的互信息量.

解

把各种天气记作 x_1 (晴) ,x_2 (阴) , x_3 (雨) , x_4 (雪) . 收到消息 y_1 后, 各种天气发生的概率变成了后验概率:

$$p(x_1|y_1) = \frac{p(x_1y_1)}{p(y_1)} = 0$$

$$p(x_2|y_1) = \frac{p(x_2y_1)}{p(y_1)} = \frac{1/4}{1/4+1/8+1/8} = \frac{1}{2}$$

$$p(x_3|y_1) = \frac{p(x_3y_1)}{p(y_1)} = \frac{1/8}{1/4+1/8+1/8} = \frac{1}{4}$$

$$p(x_4|y_1) = \frac{1}{4}$$

某地二月份天气出现的概率分别为晴 $1/2$，阴 $1/4$，雨 $1/8$, 雪 $1/8$. 某一天有人告诉你: “今天不是晴天”，把这句话作为收到的消息 y_1，求收到 y_1 后，y_1 与各种天气的互信息量.

解

根据互信息量的定义，可计算出 y_1 与各种天气之间的互信息:

$$I(x_1;y_1)=\log_2\frac{p(x_1|y_1)}{p(x_1)}=\infty$$

$$I(x_2;y_1)=\log_2\frac{p(x_2|y_1)}{p(x_2)}=\log_2\frac{1/2}{1/4}=1\text{bit}$$

$$I(x_3;y_1)=\log_2\frac{p(x_3|y_1)}{p(x_3)}=\log_2\frac{1/4}{1/8}=1\text{bit}$$

$$I(x_4;y_1)=\log_2\frac{p(x_4|y_1)}{p(x_4)}=\log_2\frac{1/4}{1/8}=1\text{bit}$$

平均互信息

随机变量对随机变量的互信息

一个事件 y_j 所给出关于另一个事件 x_i 的信息定义为互信息，用 $I(x_i; y_j)$ 表示。

$$I(x_i; y_j) = I(x_i) - I(x_i|y_i) = \log \frac{p(x_i|y_i)}{p(x_i)}$$

定义 8

定义互信息 $I(x_i; y_j)$ 在 XY 的联合概率空间中的统计平均值为随机变量 X 和 Y 间的平均互信息。

$$I(X; Y) = \sum_x \sum_y p(x, y) I(x_i; y_j)$$

也称为互信息。

互信息有以下性质：

性质 8

[对称性]

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

对称性表示从 Y 中获得关于 X 的信息量等于从 X 中获得关于 Y 的信息量.

性质 9

[非负性]

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

当且仅当 $p(x,y) = p(x)p(y)$ 即 X 与 Y 独立时，互信息为 0

证明.

$$
\begin{aligned}
-I(X;Y) &= \sum_{i=1}^{n}\sum_{j=1}^{m} p\left(x_i y_j\right) \log_2 \frac{p\left(x_i\right) p\left(y_j\right)}{p\left(x_i y_j\right)} \\
&\leqslant \log_2 \sum_{i=1}^{n}\sum_{j=1}^{m} p\left(x_i y_j\right) \frac{p\left(x_i\right) p\left(y_j\right)}{p\left(x_i y_j\right)} \\
&= \log_2 \sum_{i=1}^{n}\sum_{j=1}^{m} p\left(x_i\right) p\left(y_j\right) = 0
\end{aligned}
$$

所以

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

证毕。 □

平均互信息是非负的, 说明给定随机变量 Y 后, 一般来说总能消除一部分关于 X 的不确定性.

相对熵：定义

相对熵是两个随机分布 $p(x)$ 和 $q(x)$ 之间距离的度量。统计学上对应于对数似然比的期望。

定义 9

定义同一个随机变量 x 的两个概率密度函数 $p(x)$ 和 $q(x)$ 间的相对熵为：

$$D(p||q) = \mathbb{E}_{x\sim p}\left[\log \frac{p(x)}{q(x)}\right] = \mathbb{E}_{x\sim p}[\log p(x) - \log q(x)]$$

在机器学习中，相对熵更常用的名称是 *Kullback-Leibler(KL)* 散度，记做 $D_{KL}(p||q)$。

- 在信息理论中，相对熵是用来度量使用基于 q 的编码来编码来自 p 的样本平均所需的额外的比特个数。典型情况下，p 表示数据的真实分布，q 表示数据的理论分布，模型分布，或 p 的近似分布。
- 给定一个字符集的概率分布，我们可以设计一种编码，使得表示该字符集组成的字符串平均需要的比特数最少。假设这个字符集是 X，对 $x \in X$，其出现概率为 $p(x)$，那么其最优编码平均需要的比特数等于这个字符集的熵：

$$H(x) = -\sum_{x \in X} p(x) \log \frac{1}{q(x)}$$

在同样的字符集上，假设存在另一个概率分布 $q(x)$，如果用概率分布 $p(x)$ 的最优编码（即字符 x 的编码长度等于 $\log \frac{1}{p(x)}$ ），来为符合分布 $p(x)$ 的字符编码，那么表示这些字符就会比理想情况多用一些比特数。

- 相对熵就是用来衡量这种情况下平均每个字符多用的比特数，因此可以用来衡量两个分布的距离，即：

$$D(p\|q) = -\sum_{x \in X} p(x) \log \frac{1}{p(x)} + \sum_{x \in X} p(x) \log \frac{1}{q(x)} = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

例 5

假如一个字符发射器，随机发出 *0* 和 *1* 两种字符，真实发出概率分布为 A，但实际不知道 A 的具体分布。通过观察，得到概率分布 B 与 C，各个分布的具体情况如下：

$$A(0) = 1/2, A(1) = 1/2$$

$$B(0) = 1/4, B(1) = 3/4$$

$$C(0) = 1/8, C(1) = 7/8$$

可以计算出得到如下：

$$D_{KL}(A\|B) = 1/2\log(\frac{1/2}{1/4}) + 1/2\log(\frac{1/2}{3/4}) = 1/2\log(4/3)$$

$$D_{KL}(A\|C) = 1/2\log(\frac{1/2}{1/8}) + 1/2\log(\frac{1/2}{7/8}) = 1/2\log(16/7)$$

由上式可知，按照概率分布 B 进行编码，要比按照 C 进行编码，平均每个符号增加的比特数目少。从分布上也可以看出，实际上 B 要比 C 更接近实际分布（因为其与 A 分布的相对熵更小）。

相对熵的性质

- KL 散度有很多有用的性质，可以证明它是非负的。
- KL 散度为 0 当且仅当 p 和 q 在离散型变量的情况下是相同的分布，或者在连续型变量的情况下是“几乎处处”相同的。
- KL 散度是非负的并且可以度量两个分布之间的差异。然而，它并不是距离，因为它不是对称的，不满足三角不等式。
- 联合分布 $p(X, Y)$ 和 $p(X)p(Y)$ 之间的 KL 散度可以作为 X 和 Y 的互信息的另一种定义：

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

23.1.4 互信息与熵的关系

性质 10

[互信息和熵的关系]

$$I(X;\,Y)=H(X)-H(X|\,Y)=H(\,Y)-H(\,Y|X)=H(X)+H(\,Y)-H(XY)$$

当 $X,\,Y$ 统计独立时,$I(X;\,Y)=0$.

证明.

$$\begin{aligned}
I(X;\,Y)&=\sum_{i=1}^{n}\sum_{j=1}^{m}p\left(x_iy_j\right)\log_2\frac{p\left(x_iy_j\right)}{p\left(x_i\right)p\left(y_j\right)}=\sum_{i=1}^{n}\sum_{j=1}^{m}p\left(x_iy_j\right)\left(\log_2\frac{1}{p\left(x_i\right)}+\log_2\frac{p\left(x_iy_j\right)}{p\left(y_j\right)}\right)\\
&=\sum_{i=1}^{n}\sum_{j=1}^{m}p\left(x_iy_j\right)\log_2\frac{1}{p\left(x_i\right)}+\sum_{i=1}^{n}\sum_{j=1}^{m}p\left(x_iy_j\right)\log_2\frac{p\left(x_iy_j\right)}{p\left(y_j\right)}\\
&=\sum_{i=1}^{n}p(x_i)\log_2\frac{1}{p(x_i)}-\sum_{i=1}^{n}\sum_{j=1}^{m}p\left(x_iy_j\right)\log_2\frac{1}{p(x_i|y_j)}=H(X)-H(X|\,Y)
\end{aligned}$$

性质 11

$$H(X) = I(X; X)$$

也就是随机变量 X 的熵是自己对自己的互信息。

性质 12

[极值性]

$$I(X; Y) \leqslant H(X), I(X; Y) \leqslant H(Y)$$

推论 2

条件熵和信息熵的关系

$$H(X|Y) \leq H(X) \tag{5}$$

$$H(Y|X) \leq H(Y) \tag{6}$$

推论 3

联合熵和信息熵的关系：

$$H(XY) \leq H(X) + H(Y) \tag{7}$$

23.1.5 熵、相对熵和互信息的链式法则

熵的链式法则

两个随机变量 X 和 Y 的联合熵等于 X 的熵加上在 X 已知条件下 Y 的条件熵, 这个关系可以方便地推广到 N 个随机变量的情况, 即

$$H(X_1X_2\cdots X_N) = H(X_1) + H(X_2|X_1) + \cdots + H(X_N|X_1X_2\cdots X_{N-1})$$

称为熵函数的链规则。

如果 N 个随机变量 $X_1, X_2, \cdots, X_N$ 相互独立, 则有

$$H(X_1X_2\cdots X_N) = \sum_{i=1}^{N} H(X_i) \tag{8}$$

互信息的链式法则

我们先定义条件互信息：

定义 10

随机变量 X 和 Y 在给定随机变量 Z 时的条件互信息为

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = \sum_{i=1}^{l}\sum_{j=1}^{m}\sum_{k=1}^{n} p(x_i,y_j,z_k)\log_2 \frac{p(x_i,y_j|z_k)}{p(x_i|z_k)p(y_j|z_k)}$$

互信息的链式法则

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

相对熵的链式法则

我们先定义条件相对熵：

定义 11

联合概率密度函数 $p(x,y)$ 和 $q(x,y)$ 的条件相对熵 $D(p(y|x)||q(y|x))$ 定义为条件概率密度函数 $p(y|x)$ 和 $q(y|x)$ 间关于 $p(x)$ 的平均相对熵，即

$$D(p(y|x)||q(y|x)) = \sum_{i=1}^{m} p(x_i) \sum_{j=1}^{n} p(y_j|x_i) \log_2 \frac{p(y|x)}{q(y|x)} = \sum_{i=1}^{m} \sum_{j=1}^{n} p(x_i, y_j) \log_2 \frac{p(y|x)}{q(y|x)}$$

相对熵的链式法则

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$$

23.1.6 信息不等式

信息不等式

设 $p(x), q(x)$ 是两个概率密度函数，则

$$D(p||q) \geq 0$$

当且仅当对任意 x，$p(x) = q(x)$ 时，等号成立。

证明.

$$\begin{aligned} -D(p||q) &= -\sum_x p(x) \log \frac{p(x)}{q(x)} = \sum_x p(x) \log \frac{q(x)}{p(x)} \\ &\leq \log \sum_x p(x) \frac{q(x)}{p(x)} = \log \sum_x q(x) \\ &= \log 1 = 0 \end{aligned}$$

□

信息不等式的推论

推论 4

对任意两个随机变量 X 和 Y，

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

当且仅当 X 与 Y 相互独立时，等号成立.

推论 5

条件相对熵非负，即 $D(p(y|x)||q(y|x)) \geq 0$. 当且仅当对任意 y 满足 $p(y|x) = q(y|x)$ 时，等号成立。

推论 6

条件互信息非负，即 $I(X;Y|Z) \geq 0$，当且仅当对给定随机变量 Z 时，X 和 Y 是条件独立的，等号成立。

信息不等式

数据处理定理

为了表述数据处理定理, 需要引入三元随机变量 X, Y, Z 的平均条件互信息和平均联合互信息的概念.

定义 12

[平均条件互信息]

$$I(X;\, Y|Z) = E[I(x;\, y|z)] = \sum_x \sum_y \sum_z p(xyz) \log_2 \frac{p(x|yz)}{p(x|z)} \tag{9}$$

它表示随机变量 Z 给定后, 从随机变量 Y 所得到的关于随机变量 X 的信息量.

定义 13

[平均联合互信息]

$$I(X;YZ) = E[I(x;yz)] = \sum_x \sum_y \sum_z p(xyz) \log_2 \frac{p(x|yz)}{p(x)} \tag{10}$$

它表示从二维随机变量 YZ 所得到的关于随机变量 X 的信息量.

可以证明

$$\begin{aligned} I(X;\, YZ) &= \sum_x \sum_y \sum_z p(xyz) \log_2 \frac{p(x|z)p(x|vz)}{p(x)p(x|z)} \\ &= I(X;\, Z) + I(X;\, Y|Z) \end{aligned} \tag{11}$$

同理

$$I(X;\, YZ) = I(X;\, Y) + I(X;\, Z|\, Y) \tag{12}$$

定理 1

[数据处理定理] 如果随机变量 X, Y, Z 构成一个马尔可夫链, 则有以下关系成立:

$$I(X;Z) \leqslant I(X;Y), I(X;Z) \leqslant I(Y;Z) \tag{13}$$

等号成立的条件是对于任意的 x, y, z, 有 $p(x|yz) = p(x|z)$ 和 $p(z|xy) = p(z|x)$。

证明.

当 X, Y, Z 构成一个马尔可夫链时，Y 值给定后，X, Z 可以认为是互相独立的. 所以，

$$I(X; Z|Y) = 0$$

又因为 $I(X; YZ) = I(X; Y) + I(X; Z|Y) = I(X; Z) + I(X; Y|Z)$，并且 $I(X; Y|Z) \geqslant 0$，所以 $I(X; Z) \leqslant I(X; Y)$。

当 $p(x|yz) = p(x|z)$ 时，Z 值给定后，X 和 Y 相互独立，所以

$$I(X; Y|Z) = 0$$

因此

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

这时 $p(x|yz) = p(x|z) = p(x|y)$。$Y$,Z 为确定关系时显然满足该条件。

同理可以证明 $I(X; Z) \leqslant I(Y; Z)$，并且当 $p(z|xy) = p(z|x)$ 时，等号成立。

证毕。 □

$I(X;Z) \leqslant I(X;Y)$ 表明从 Z 所得到的关于 X 的信息量小于等于从 Y 所得到的关于 X 的信息量. 如果把 $Y \to Z$ 看作数据处理系统, 那么通过数据处理后, 虽然可以满足我们的某种具体要求, 但是从信息量来看, 处理后会损失一部分信息, 最多保持原有的信息, 也就是说, 对接收到的数据 Y 进行处理后, 决不会减少关于 X 的不确定性. 这个定理称为数据处理定理. 数据处理定理与日常生活中的经验是一致的. 比如: 通过别人转述一段话或多或少会有一些失真, 通过书本得到的间接经验总不如直接经验来得详实.

23.2.1 连续信源的微分熵

连续信源的数学模型

连续随机变量的取值是连续的，一般用概率密度函数来描述其统计特征。

- 单变量连续信源的数学模型为 $X:\begin{bmatrix}\mathbb{R}\\ p(x)\end{bmatrix}$，并且满足 $\int_R p(x)\mathrm{d}x=1$，$\mathbb{R}$ 是实数域，表示 X 的取值范围。
- 对于取值范围有限的连续信源还可以表示成 $X:\begin{bmatrix}(a,b)\\ p(x)\end{bmatrix}$，并满足 $\int_a^b p(x)\mathrm{d}x=1$，$(a,b)$ 是 X 的取值范围。

连续信源熵的求解思想

- 通过对连续变量的取值进行量化分层，可以将连续随机变量用离散随机变量来逼近。
- 量化间隔越小，离散随机变量与连续随机变量越接近。当量化间隔趋于 0 时，离散随机变量就变成了连续随机变量。
- 通过对离散随机变量的熵取极限，可以推导出连续随机变量熵的计算公式。

连续信源熵的推导

我们把连续随机变量 X 的取值分割成 n 个小区间, 各小区间等宽, 区间宽度 $\Delta=\frac{b-a}{n}$，则变量落在第 i 个小区间的概率为

$$p_i\{a+(i-1)\Delta \leqslant x \leqslant a+i\Delta\}=\int_{a+(i-1)\Delta}^{a+i\Delta} p(x)\mathrm{d}x = p\left(x_i\right)\Delta \tag{14}$$

其中，x_i 是 $a+(i-1)\Delta$ 到 $a+i\Delta$ 之间的某一值。当 $p(x)$ 是连续函数时，由中值定理可知，存在一个 x_i 使(14)式成立，这样，连续变量 X 就可用取值为 $x_i, i=1,2,\cdots,n$ 的离散变量来近似，连续信源就被量化成离散信源, 这 n 个取值对应的概率分布为 $p_i=p\left(x_i\right)\Delta$，这时的离散信源熵是

$$\begin{aligned} H(X) &= -\sum_{i=1}^{n} p\left(x_i\right)\Delta \log_2\left[p\left(x_i\right)\Delta\right] \\ &= -\sum_{i=1}^{n} p\left(x_i\right)\Delta \log_2 p\left(x_i\right) - \sum_{i=1}^{n} p\left(x_i\right)\Delta \log_2 \Delta \end{aligned} \tag{15}$$

当 $n \to \infty$ 时，$\Delta \to 0$，如果(15)极限存在，离散信源熵就变成了连续信源的熵:

$$\begin{aligned}\lim_{\substack{n\to\infty\\ \Delta\to 0}} H(X) &= \lim_{n\to\infty} -\sum_{i=1}^{n} p\left(x_i\right)\Delta \log_2 p\left(x_i\right) - \lim_{n\to\infty}\sum_{i=1}^{n} p\left(x_i\right)\Delta\log_2\Delta \\ &= -\int_a^b p(x)\log_2 p(x)\mathrm{d}x - \lim_{n\to\infty}\log_2\Delta\int_a^b p(x)\mathrm{d}x \\ &= -\int_a^b p(x)\log_2 p(x)\mathrm{d}x - \lim_{\substack{n\to\infty\\ \Delta\to 0}}\log_2\Delta\end{aligned} \tag{16}$$

式(16)第一项一般是定值, 第二项为无穷大量，因此连续信源的熵实际是无穷大量。这一点是可以理解的，因为连续信源的可能取值是无限多的，所以它的不确定性是无限大的，当确知输出为某值后，所获得的信息量也是无限大。

微分熵

在丢掉第二项后, 定义第一项为连续信源的**微分熵**:

$$h(X) = -\int_{\mathbb{R}} p(x)\log_2 p(x)\mathrm{d}x \tag{17}$$

微分熵又称为**差熵**。虽然 $h(X)$ 已不能代表连续信源的平均不确定性，也不能代表连续信源输出的信息量，但是它具有和离散熵相同的形式，也具有离散熵的主要特性，比如可加性，但是不具有非负性。另外，我们在实际问题中常常考虑的是熵差，比如平均互信息，在讨论熵差时，只要两者离散逼近时所取的间隔 Δ 一致，这两个无限大量就将互相抵消，所以熵差具有信息的特性，如非负性。由此可见，连续信源的熵 $h(X)$ 具有相对性。

联合熵和条件熵

同样，可以定义两个连续随机变量的联合熵：

$$h(XY) = -\iint_{\mathbb{R}^2} p(xy)\log_2 p(xy)\mathrm{d}x\mathrm{d}y \tag{18}$$

以及条件熵

$$h(X|Y) = -\iint_{\mathbb{R}^2} p(xy)\log_2 p(y|x)\mathrm{d}x\mathrm{d}y \tag{19}$$

$$h(Y|X) = -\iint_{\mathbb{R}^2} p(xy)\log_2 p(x|y)\mathrm{d}x\mathrm{d}y \tag{20}$$

微分熵、联合熵和条件熵三者的关系

并且它们之间也有与离散随机变量一样的相互关系:

$$h(XY) = h(X) + h(Y|X) = h(Y) + h(X|Y) \tag{21}$$

$$h(X|Y) \leqslant h(X) \tag{22}$$

$$h(Y|X) \leqslant h(Y) \tag{23}$$

相对熵

我们也可以定义连续信源的相对熵。

定义 14

同一随机变量 X 的两个概率密度函数 $p(x),q(x)$ 的相对熵，或称 KL 散度，定义为：

$$D(p(x)||q(x)) = \int p(x)\log\frac{p(x)}{q(x)}dx = -H(X) - \int p(x)\log q(x)dx$$

连续信源的相对熵也是非负的，因为

$$-D(p(x)||q(x)) = \int p(x)\log\frac{q(x)}{p(x)}dx \le \int p(x)(\frac{q(x)}{p(x)} - 1)dx = \int q(x)dx - \int p(x)dx = 0$$

其中用到了不等式 $\ln t \le t-1$。因此 $D(p(x)||q(x)) \ge 0$，并且当且仅当 $p(x) = q(x)$ 时，取等。

例 6

X 是在区间 (a,b) 内服从均匀分布的连续随机变量，求微分熵.

$$p(x)=\begin{cases}\frac{1}{b-a} & x\in(a,b)\\ 0 & x\notin(a,b)\end{cases}$$

解

$$h(X)=-\int_a^b p(x)\log_2 p(x)\mathrm{d}x=-\int_a^b \frac{1}{b-a}\log_2\frac{1}{b-a}\mathrm{d}x=\log_2(b-a)$$

当 $(b-a)>1$ 时，$h(X)>0$；

当 $(b-a)=1$ 时，$h(X)=0$；

当 $(b-a)<1$ 时，$h(X)<0$，这说明连续熵不具有非负性，失去了信息的部分含义和性质(但是熵差具有信息的特性)。

例 7

求均值为 m, 方差为 σ^2 的高斯分布的随机变量的微分熵.

解

高斯随机变量的概率密度为 $p(x)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-m)^2}{2\sigma^2}}$，微分熵为

$$
\begin{aligned}
h(X) &= -\int_{-\infty}^{+\infty} p(x)\log_2 p(x)\mathrm{d}x = -\int_{-\infty}^{+\infty} p(x)\log_2\left[\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-m)^2}{2\sigma^2}}\right]\mathrm{d}x \\
&= -\int_{-\infty}^{+\infty} p(x)\log_2\frac{1}{\sqrt{2\pi}\sigma}\mathrm{d}x - \log_2 e\int_{-\infty}^{+\infty} p(x)\left[-\frac{(x-m)^2}{2\sigma^2}\right]\mathrm{d}x \\
&= \log_2\sqrt{2\pi}\sigma + \log_2 e\int_{-\infty}^{+\infty} p(x)\frac{(x-m)^2}{2\sigma^2}\mathrm{d}x \\
&= \log_2\sqrt{2\pi}\sigma + \frac{1}{2}\log_2\mathrm{e} = \log_2\sqrt{2\pi e}\sigma
\end{aligned}
$$

我们看到, 正态分布的连续信源的微分熵与数学期望 m 无关, 只与方差 σ^2 有关.

23.2.2 连续信源的最大熵

离散信源当信源符号为等概分布时有最大熵. 连续信源微分熵也有极大值, 但是与约束条件有关, 当约束条件不同时, 信源的最大熵不同. 我们一般关心的是下面两种约束下的最大熵.

定理 2

在均值一定的情况下, 服从均匀分布的随机变量 X 具有最大熵.

即

$$p(x)=\begin{cases}\frac{1}{b-a} & a\leqslant x\leqslant b\\ 0 & \text{其他}\end{cases}$$

$$h(X)=-\int_a^b p(x)\log_2 p(x)\mathrm{d}x=-\int_a^b \frac{1}{b-a}\log_2\frac{1}{b-a}\mathrm{d}x=\log_2(b-a)$$

这个结论与离散信源在等概分布时达到最大熵的结论类似.

定理 3

对于固定均值为 μ 和方差为 σ^2 的连续随机变量，当服从高斯分布 $N(\mu,\sigma^2)$ 时具有最大熵.

证明.

对给定的 $p(x)$，利用相对熵非负，有

$$H(p) \leq -\int_{-\infty}^{+\infty} p(x)\log q(x)\mathrm{d}x$$

取 $q(x)=N(\mu,\sigma^2)$，有

$$H(p) \leq -\int_{-\infty}^{+\infty} p(x)\log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\right)\mathrm{d}x = \int_{-\infty}^{+\infty} p(x)\left\{\frac{(x-\mu)^2}{2\sigma^2}+\log\sqrt{2\pi}\sigma\right\}\mathrm{d}x$$

$$= \frac{1}{2\sigma^2}\int_{-\infty}^{+\infty} p(x)(x-\mu)^2\mathrm{d}x+\log\sqrt{2\pi}\sigma = \frac{1}{2}+\log\sqrt{2\pi}\sigma$$

当 $p(x)=N(\mu,\sigma^2)$ 时，可以取等，证毕。 □

这说明，当均值和方差一定时，高斯分布的连续信源的熵最大。

23.3.1 机器学习中常用的信息量度量：信息熵和互信息

回顾

- 信息熵： $H(X) = \mathbb{E}[I(x_i)] = -\sum_{i=1}^{q} p(x_i) \log p(x_i)$
  信息熵是对信息不确定性的度量，也可以这样理解，数据信息熵越小，数据就越纯。
- 互信息：假设带标签 X 的数据集有若干属性 $Y_1, Y_2, \cdots, Y_n$，我们想通过选择数据集的某个属性判断数据集的标签，那么我们就要根据哪一个属性对标签的信息量最大以选择属性。这时我们就要利用互信息

$$\arg\max_i I(X; Y_i) = \arg\max_i (H(X) - H(X|Y_i))$$

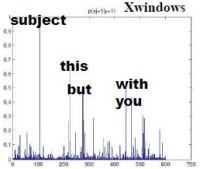
  属性对标签的互信息可以理解成选择特定属性后，不确定性的下降量，也就是数据纯度的提升量。在机器学习中，这个量等价于信息增益。

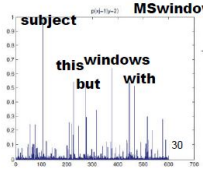
信息熵和互信息（信息增益）在机器学习的决策树算法中具有重要应用！

互信息的应用：特征选择

在特征选择时，可以通过计算特征与目标之间的互信息，选择与目标互信息最大的那些特征，抛弃与目标关系不大的特征。

- 给定文档分类任务，将文档分成 class 1 (X windows) and class 2 (MS windows)，特征为 600 个二维特征 (600 个词语分别是否在文档中出现), 令 $p(x_i)$ 为词语在文档中出现的概率，$p(x_i|y_j)$ 为在 y_j 分类下词语在文档中出现的概率。则可计算 $I(X;Y) = H(X) - H(X|Y)$，互信息高的词语 (windows, microsoft) 更有判别性。





highest MI	MI
windows	0.215
microsoft	0.095
dos	0.092
motif	0.078
window	0.067

KL 散度

回顾

相对熵或称 KL 散度可以衡量同一个随机变量 x 的概率分布 $p(x)$ 和 $q(x)$ 的差异：

$$D_{KL}(p||q) = \mathbb{E}_{x\sim p}[\log p(x) - \log q(x)]$$

- 在机器学习中，p 往往用来表示样本的真实分布，q 用来表示模型所预测的分布，那么 KL 散度就可以计算两个分布的差异，也就是 Loss 损失值。从 KL 散度公式中可以看到 q 的分布越接近 p（q 分布越拟合 p），那么散度值越小，即损失值越小。
- 相对熵可以衡量两个随机分布之间的距离，当两个随机分布相同时，它们的相对熵为零，当两个随机分布的差别增大时，它们的相对熵也会增大。所以相对熵可以用于比较文本的相似度，先统计出词的频率，然后计算相对熵。
- 相对熵是一些优化算法，例如最大期望算法（EM 算法）的损失函数。此时参与计算的一个概率分布为真实分布，另一个为理论（拟合）分布，相对熵表示使用理论分布拟合真实分布时产生的信息损耗。

交叉熵

一个和 KL 散度密切联系的量是**交叉熵** (cross-entropy)。将 KL 散度公式进行变形

$$D_{KL}(p\|q) = \sum_{i=1}^{n} p(x_i) \log \frac{p(x_i)}{q(x_i)} = \sum_{i=1}^{n} p(x_i) \log p(x_i) - \sum_{i=1}^{n} p(x_i) \log q(x_i)$$

$$= -H(p(x)) + (-\sum_{i=1}^{n} p(x_i) \log q(x_i))$$

等式的前一部分恰巧就是 p 的熵，等式的后一部分就是交叉熵。

定义 15

设关于随机变量 x 的两个分布 $p(x), q(x)$，关于这两个分布的**交叉熵**定义为：

$$H(p, q) = -\mathbb{E}_{x\sim p} \log q(x) = H(p) + D_{KL}(p||q) = -\sum_{i=1}^{n} p(x_i) \log q(x_i) \tag{24}$$

- 针对 q 最小化交叉熵等价于最小化 KL 散度，因为 q 并不参与被省略的那一项。
- 在给定 p 的情况下，如果 q 和 p 越接近交叉熵越小；如果 q 和 p 越远交叉熵就越大。

交叉熵的应用

交叉熵是 Shannon 信息论中一个重要概念，也是用于度量两个概率分布间的差异性信息。

- 在机器学习中，我们需要评估 label 和 predicts 之间的差距，可以使用 KL 散度，即 $D_{KL}(y||\tilde{y})$ ，由于 KL 散度中的前一部分 $-H(y)$ 不变，故在优化过程中，只需要关注交叉熵就可以了。所以一般在机器学习中直接用交叉熵做 Loss，评估模型。
- 交叉熵也可在神经网络中作为损失函数，p 表示真实标记的分布，q 则为训练后的模型的预测标记分布，交叉熵损失函数可以衡量 p 与 q 的相似性。交叉熵作为损失函数还有一个好处是使用 sigmoid 函数在梯度下降时能避免均方误差损失函数学习速率降低的问题，因为学习速率可以被输出的误差所控制。
- 在特征工程中，可以用来衡量两个随机变量之间的相似度。在自然语言模型中，由于真实的分布 p 是未知的，模型是通过训练集得到的，交叉熵就是衡量这个模型在测试集上的正确率。

JS 散度

JS 散度（Jensen-Shannon Divergence）是一种对称的衡量两个分布相似度的度量方式，其取值在 0 到 1 之间。

定义 16

设关于随机变量 x 的两个分布 $p(x), q(x)$，关于这两个分布的 ***JS* 散度**定义为：

$$D_{JS}(p,q) = \frac{1}{2}D_{KL}(p,M) + \frac{1}{2}D_{KL}(q,M)$$

其中 $M = \frac{1}{2}(p+q)$

JS 散度是 KL 散度一种改进，解决了 KL 散度非对称的问题。但这两种散度都存在一个问题，即如果两个分布 p, q 没有重叠或者重叠非常少时，那么 KL 散度值是没有意义的，而 JS 散度值是一个常数，这在学习算法中是比较致命的，这就意味着这一点的梯度为 0，也即梯度消失了，此时 KL 散度和 JS 散度都很难衡量两个分布的距离。

Wasserstein 距离

一个改进的距离是 Wasserstein 距离（Wasserstein Distance），它也是用于衡量两个分布之间的距离。

定义 17

对于两个分布 q_1, q_2，*pth-Wasserstein* 距离定义为

$$W_p(q_1, q_2) = (\inf_{\gamma(x,y)\in\Gamma(q_1,q_2)} E_{(x,y)\sim\gamma(x,y)}[d(x,y)^p])^{\frac{1}{p}}$$

其中 $\Gamma(q_1, q_2)$ 是边际分布为 q_1 和 q_2 的所有可能的联合分布集合，$d(x, y)$ 为 x 和 y 的距离，比如 l_p 距离等。

Wasserstein 距离相比 KL 散度和 JS 散度的优势在于：即使两个分布没有重叠或者重叠非常少，Wasserstein 距离仍然能反映两个分布的远近。

例 8

对于 $\mathbb{R}^n$ 空间中的两个高斯分布 $p = N(\mu_1, \Sigma_1)$ 和 $q = N(\mu_2, \Sigma_2)$，它们的 *2nd-Wasserstein* 距离为

$$W_2(p, q) = \|\mu_1 - \mu_2\|_2^2 + Tr(\Sigma_1 + \Sigma_2 - 2(\Sigma_2^{\frac{1}{2}}\Sigma_1\Sigma_2^{\frac{1}{2}})^{\frac{1}{2}})$$

当两个分布的方差为 *0* 时，*2nd-Wasserstein* 距离等价于欧氏距离。

Wasserstein 距离在深度学习中的 GAN（生成对抗网络）模型中具有重要的应用。

23.3.2 其他概率相关的度量：马氏距离

在前面，我们介绍了一些关于度量两个向量相似度的一些方法。
并且我们提到了闵氏距离 (包括曼哈顿距离、欧氏距离和切比雪夫距离) 存在明显的缺点，并通过下例进行了说明。

例 9

给定二维样本 (身高, 体重)，其中身高范围是 $150 \sim 190$，体重范围是 $50 \sim 60$，有三个样本：$a(180, 50), b(190, 50), c(180, 60)$。

- 通过计算可以得出 ab 之间的闵氏距离等于 ac 之间的闵氏距离，但是身高的 $10cm$ 不等价于体重的 $10kg$。

现在我们就来介绍解决这个问题的一种度量相似度的方式。

定义 18

马氏距离：表示点与一个分布之间的距离。有 m 个样本向量 $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m$，协方差矩阵记为 $\boldsymbol{S}$，均值记为向量 $\boldsymbol{\mu}$，则其中样本向量 $\boldsymbol{x}$ 到 $\boldsymbol{\mu}$ 的马氏距离表示为：

$$dist(\boldsymbol{x}) = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{S}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$$

而其中向量 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 之间的马氏距离定义为：

$$dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i-\boldsymbol{x}_j)^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i-\boldsymbol{x}_j)}$$

若协方差矩阵是单位矩阵（各个样本向量之间独立同分布），则公式就成了：

$$dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i-\boldsymbol{x}_j)^T (\boldsymbol{x}_i-\boldsymbol{x}_j)}$$

也就是欧氏距离了。

马氏距离的优点：量纲无关，排除变量之间的相关性的干扰。

皮尔逊相关系数

相关系数是衡量随机变量 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 相关程度的一种方法，一般用 r 表示。r 的取值范围是 [-1,1]。r 的绝对值越大，则表明 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 相关度越高。当 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 线性相关时，相关系数取值为 1（正线性相关）或 -1（负线性相关）。

定义 19

设随机变量 $\boldsymbol{x}, \boldsymbol{y}$，皮尔逊相关系数定义为：

$$r(\boldsymbol{x}, \boldsymbol{y}) = \frac{Cov(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{D(\boldsymbol{x})}\sqrt{D(\boldsymbol{y})}} = \frac{E((\boldsymbol{x} - E\boldsymbol{x})(\boldsymbol{y} - E\boldsymbol{y}))}{\sqrt{D(\boldsymbol{x})}\sqrt{D(\boldsymbol{y})}}$$

其中，$Cov(\boldsymbol{x}, \boldsymbol{y})$ 为 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 的协方差，$\sqrt{D(\boldsymbol{x})}$ 为 $\boldsymbol{x}$ 的标准差，$\sqrt{D(\boldsymbol{y})}$ 为 $\boldsymbol{y}$ 的标准差。

Jaccard 系数

Jaccard 系数又称为 Jaccard 相似系数，用于比较有限样本集之间的相似性与差异性。
Jaccard 系数值越大，样本相似度越高。

定义 20

两个集合 $\mathbb{A}$ 和 $\mathbb{B}$ 的交集元素在 $\mathbb{A}$，$\mathbb{B}$ 的并集中所占的比例，称为两个集合的杰卡德相似系数，用符号 $J(\mathbb{A},\mathbb{B})$ 表示。

$$J(\mathbb{A},\mathbb{B}) = \frac{|\mathbb{A} \bigcap \mathbb{B}|}{|\mathbb{A} \bigcup \mathbb{B}|}$$

当集合 $\mathbb{A}$，$\mathbb{B}$ 都为空时，$J(\mathbb{A},\mathbb{B})$ 定义为 *1*。

杰卡德相似系数是衡量两个集合的相似度一种指标。
对于等概率的随机排列，两个集合的 minHash 值相同的概率等于两个集合的 Jaccard 相似度。关于 minHash 算法，可以参考有关数据科学算法的教材如 *Mining of Massive Datasets*。

杰卡德距离

杰卡德距离：与 Jaccard 系数相反的概念是杰卡德距离。

定义 21

杰卡德距离可用如下公式表示：

$$J_D(\mathbb{A},\mathbb{B}) = 1 - J(\mathbb{A},\mathbb{B}) = \frac{|\mathbb{A} \bigcup \mathbb{B}| - |\mathbb{A} \bigcap \mathbb{B}|}{|\mathbb{A} \bigcup \mathbb{B}|}$$

杰卡德距离用两个集合中不同元素占所有元素的比例来衡量两个集合的区分度。

杰卡德相似系数与杰卡德距离的应用

杰卡德相似系数用在衡量样本的相似度上。

样本 A 与样本 B 是两个 n 维向量，而且所有维度的取值都是 0 或 1。例如：$A = (0111)$ 和 $B = (1011)$。我们将样本看成是一个集合，1 表示集合包含该元素，0 表示集合不包含该元素。

p：样本 A 与 B 都是 1 的维度的个数

q：样本 A 是 1，样本 B 是 0 的维度的个数

r：样本 A 是 0，样本 B 是 1 的维度的个数

s：样本 A 与 B 都是 0 的维度的个数

那么样本 A 与 B 的杰卡德相似系数可以表示为：

$$J = \frac{p}{p+q+r}$$

这里 $p+q+r$ 可理解为 A 与 B 的并集的元素个数，而 p 是 A 与 B 的交集的元素个数。

而样本 A 与 B 的杰卡德距离表示为：

$$J_D = \frac{q+r}{p+q+r}$$

本讲小结

熵、相对熵和互信息

- 自信息和熵
- 联合熵和条件熵
- 互信息和相对熵
- 链式法则和信息不等式

连续信源和数据科学中的信息论

- 微分熵
- 联合熵和条件熵
- 相对熵
- 最大熵

上述概念以及相应扩展的概念如交叉熵、JS 散度、Wasserstein 距离和马氏距离等，在机器学习中具有重要应用，比如用于分类任务中各种相似性或相异性度量，损失的度量等等！