标题 **title**

作者 *author*

2023 年 7 月 18 日

前言

目录

第一部分

科学的逻辑

第二部分

信息与数据

第一章　信息论基础

信息是什么？不同于真实的物理世界，信息仿佛看不见，摸不着. 然而，任何人都可以体会到信息的存在，信息是我们认识世界的基础. 信息的存在正如同物理世界中的能量、动量一般，抽象而具有一般性. 信息论已经在计算机、AI、认知理论等诸多领域中得到了广泛的应用. 本章探讨信息论的基础，并给出他们在 AI 中的一些应用.

在第 1.1 节，我们讨论熵的概念与性质. 在第 1.2 节，我们讨论 Kullback-Leibler 散度的概念与性质. 在第 1.3 节，我们给出 Shannon 定理证明.

§1.1　熵

1.1.1　概念的导出

我们常说“恐惧来源于未知”，信息似乎代表着某种确定的东西，某种知识，因而和不确定性有相反的关系. 更精确地说，*消除不确定性的东西被称为信息*. 当然，这句话本身似乎是一种循环论证，它并没有真正回答信息或者不确定性到底是什么. 所以我们进一步的问题是，给定一个“对象”，如何定量衡量它不确定性（或信息量）?

然而，单个对象的信息是一个非常难以划定的概念. 同样的内容，对于不同的人来说，信息量是完全不同的. 比如说，已经学过信息论的读者再看这一部分内容，他获得的信息一定比没有学过的读者要少得多. 因而实际上，一种更加容易的办法是我们将世界视为不确定的，因而有多种可能的对象，然后考虑这一堆对象的信息量. 比如说，这本书的读者的背景是不确定的，可能学过信息论，也可能没学过，但是我们可以综合考虑不同读者的背景，然后给出一个信息的概率分析.

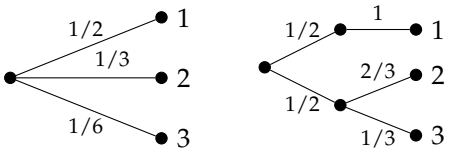
我们可以用数学来表述上面的考虑，假如我们进行一次试验，一共有 n 种可能的结果，第 i 种发生的概率为 p_i. 我们预测试验的结果，如果越能正确地预测，那么就说明我们对这个试验中包含的信息知道的越多. 假如 $p_1 = 1$，那么我们完全确定试验一定会产生结果 1. 如果 $p_i = 1/n$，那么我们完全无法预计试验的结果. 我们对试验结果的预期与

试验结果的概率分布有密切联系. 因此概率分布给我们带来了**信息**，使得我们能够产生不同的判断. 另一方面，概率分布带来了**不确定性**，使我们不能总是确信预言会成真.

我们遵循“信息论之父”Shannon 的思路，为信息提供一个严格的数学模型：熵. 假设随机变量 X 表示了所有可能的结果（编号为 1 到 n），$\Pr(X=i)=p_i$，$p=(p_1,\dots,p_n)$，有时候也把 p_i 写作 $p(i)$. 我们把不确定性度量记为 $H(p)$. Shannon 假设 H 满足以下三个性质：

1. H 是一个连续函数.
2. 事件结局可能数变多则不确定性增大：$p_i=1/n$ 时，$H(p)$ 随 n 单调递增，n 是正整数.
3. 如果一个试验被分成了两个相继的试验，那么原来的 H 应该等于分开之后的 H 的加权和.

注. 第三个假设可以用下图来理解.



假设我们有一个试验，有三种可能的结果，1,2,3，概率分别为 1/2,1/3,1/6. 该试验的不确定性是 $H(1/2,1/3,1/6)$. 我们把试验分成两步相继的试验，第一步试验有两种可能的结果，概率分别都是 1/2. 当第一步试验出现上面的结果时，第二步试验以概率 1 产生结果 1；当第二步试验出现下面的结果时，第二步试验以概率 2/3 产生结果 2，以概率 1/3 产生结果 3. 我们可以看到，分成两步之后，第一步试验的不确定性是 $H(1/2,1/2)$，第二步试验的不确定性有一半概率是 $H(1)$（上面的分支），有一半概率是 $H(2/3,1/3)$（下面的分支），因而加权的不确定性是 $1/2\cdot 0+1/2\cdot H(2/3,1/3)$. 因此第三个假设可以具体表述为

$$H\left(\frac{1}{2},\frac{1}{3},\frac{1}{6}\right)=H\left(\frac{1}{2},\frac{1}{2}\right)+\left[\frac{1}{2}\cdot H(1)+\frac{1}{2}\cdot H\left(\frac{2}{3},\frac{1}{3}\right)\right].$$

这里，我们可以看出 Shannon 的哲学思想：不确定性只来自于概率分布而不是具体对象. 他的考虑具有浓厚的工程意味，正如他自己针对通信的数学理论所说：“消息是具有含义的……然而，通信的语义层面并不是工程问题所关心的.“正是因为抽象掉了具体考虑的对象，信息论的应用才变得如此广泛.

基于上面三个假设，Shannon 证明了如下定理，这一定理直接给出了熵的概念.

定理 1.1 (Shannon 定理) *H 满足三个假设当且仅当*

$$H(p) = -C\sum_i p_i \log p_i,$$

其中 C 是正常数，$0\log 0 = 0$.

这一定理的证明较长并且和后面的讨论关联较小，所以我们在第 1.3 节 中给出证明.

根据对数的的换底公式，可以将 $C\log p_i$ 写为 $\log_b p_i$，这里 $C = 1/\log b$. 于是，Shannon 定理直接给出了熵的如下定义：

定义 1.1 (熵) *分布列 $p = (p_1, \ldots, p_n)$ 的熵定义为*

$$H(p) = -\sum_{i=1}^{n} p_i \log_b p_i.$$

其中 $b = \mathrm{e}$（自然对数底数），$0\log 0 = 0$. 当 $b = 2$ 时，我们记熵为 $H_2(p)$.

通常来说，使用 e 作为底数会使得数学推导简洁，而用 2 为底数则常常是讨论信息量时的习惯. 在后面通信理论中，我们将讨论熵在通信中的含义，以 2 为底的时候熵的实际意义会更清楚些. 如果没有特别强调，我们在讨论时总是假设 $b = \mathrm{e}$.

熵的定义还可以用数学期望的形式写出. 假设 X 的分布列是 p，$p(i) = \Pr(X = i)$，那么我们也可以把熵写成期望的形式：

$$H(p) = -\mathbb{E}[\log p(X)].$$

每一个（离散）随机变量 X 会确定一个分布列 p_X，因此我们也可以定义随机变量的熵：

定义 1.2 (随机变量的熵) *随机变量 X 的熵定义为*

$$H(X) = -\mathbb{E}[\log p_X(X)].$$

其中 p_X 是 X 的分布列，$0\log 0 = 0$.

尽管从信息论的角度我们可以唯一确定熵的定义，但是熵的概念在物理学上早就已经存在. 下面我们给出统计力学中熵的推导过程. 在经典力学中，物理系统的状态由粒子的位置和动量（速度）完全确定，将粒子位置和动量可能的值集合称为相空间，于是物理系统的演化就是相空间中的粒子状态的变化. 将相空间等分成 m 个单元，编号 1 到 m. 假设相空间中有 N 个可区分的粒子，相互独立，没有相互作用，每个粒子等可能出现在每一个单元中. 如果单元 i 中有 N_i 个粒子，那么按照粒子在单元中的分布来看，系统处于某个特定状态的概率为

$$P = \frac{N!}{N_1! \ldots N_m!} \left(\frac{1}{m}\right)^N.$$

这是一个多项分布. 两边取对数，得

$$\log P = \log(N!) - \sum_i \log(N_i!) - N \log m.$$

考虑充分大的 N_i，由 Stirling 公式，有

$$\log(N_i!) \sim \log\left(\sqrt{2\pi N_i}\left(\frac{N_i}{\mathrm{e}}\right)^{N_i}\right) \sim N_i \log N_i.$$

因此，

$$\log P \sim N \log N - \sum_i N_i \log N_i - N \log m \sim N \log N - \sum_i N_i \log N_i. \tag{1.1}$$

假设 N_i 充分大的时候，N_i/N 呈现固定的比例 p_i，那么

$$\begin{aligned} N \log N - \sum_i N_i \log N_i &\sim N \log N - \sum_i N p_i \log(N p_i) \\ &= -N \sum_i p_i \log p_i. \end{aligned}$$

$\log P \sim -N \sum_i p_i \log p_i$. 于是我们证明了：

$$\frac{1}{N} \log P \to H(p_1, \ldots, p_m), \quad N \to \infty.$$

因此，熵刻画了充分多粒子的物理系统某种特定状态出现概率！熵越大的系统越有可能达到. 更进一步，在统计力学中有 Boltzmann H-定理：孤立的粒子系统会向着熵（H）增加的方向演化，并最终达到熵最大的状态. H-定理是热力学第二定律的微观解释，熵越大的系统出现概率越大、越混乱、越接近均衡.

1.1.2 概念与性质

现在，我们将进一步探讨熵的若干拓展定义，并讨论他们的性质.

首先，我们考虑最简单的情形，即分布列为 (p_1, p_2)，此时，我们不妨设 $p_1 = p$，$p_2 = 1 - p$，那么熵就是

$$H(p_1, p_2) = H(p, 1-p) = -p \log p - (1-p) \log(1-p).$$

H 是关于 p 的函数，作图如图 1.1 所示.

利用导数的方法，很容易证明：

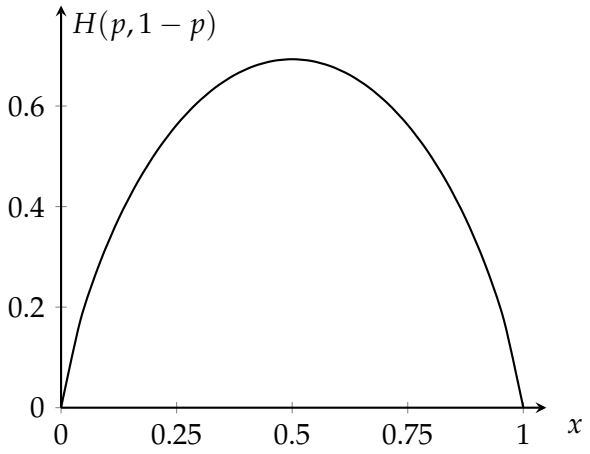


图 1.1: 熵 $H(p)$ 的图像.

命题 1.1 *$H(p)$ 在 $p \in (0, 1/2)$ 严格单调递增，在 $p \in (1/2, 1)$ 严格单调递减. 它的最小值是 0，在 $p \in \{0, 1\}$ 取得；它的最大值是 $\log 2$，在 $p = 1/2$ 取得.*

这与我们对于"不确定性"的直觉是相一致的：当 p 接近 0 或 1 时，我们对于 X 的取值几乎是确定的，因此熵接近 0；当 p 接近 $1/2$ 时，我们对于 X 的取值几乎是完全不确定的，因此熵接近最大值 $\log 2$. 实际上，这样的性质对于一般的分布也是成立的.

考虑一般分布的熵 $H(p) = H(p_1, \ldots, p_n)$. 我们有如下性质：

命题 1.2 *$H(p) \geq 0$，等号成立当且仅当某个 $p_i = 1$.*

证明 这是一个典型的证明，主要的技巧是使用熵的期望形式. 考虑随机变量 X，其分布列为 p. 回忆 Jensen 不等式：如果 f 是一个严格凸函数，那么

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

等号成立当且仅当 X 是常数.

因为 $-\log(\cdot)$ 是严格凸函数，所以根据 Jensen 不等式

$$H(X) = \mathbb{E}[-\log p(X)] \geq -\log \mathbb{E}[p(X)] \geq -\log 1 = 0.$$

等号成立当且仅当 X 是常数，即对某个 i，$p(i) = 1$. □

命题 1.3 *p_i 朝着相等方向改变的时候 H 增加. 也就是说，假设 $p_i < p_j$，再假设 $p_i' > p_i, p_j' < p_j$ 且 $p_i + p_j = p_i' + p + j'$. 用 p_i' 和 p_j' 代替原来的 p_i 和 p_j，那么 H 变大.*

证明 为简化符号，考虑 $i=1$ 和 $j=2$. 利用假设三，第一步试验中，将试验的结果 1 和结果 2 合并，第二步试验再按照 $p_1/(p_1+p_2)$ 和 $p_2/(p_1+p_2)$ 的概率产生结果 1 和结果 2. 于是，

$$\begin{aligned}
&H(p_1,p_2,\dots)\\
=&H(p_1+p_2,p_3,\dots)+(p_1+p_2)H\left(\frac{p_1}{p_1+p_2},\frac{p_2}{p_1+p_2}\right) \quad \text{（假设三）}\\
\le&H(p_1+p_2,p_3,\dots)+(p_1+p_2)<H\left(\frac{p_1'}{p_1'+p_2'},\frac{p_2'}{p_1'+p_2'}\right) \quad \text{（命题 1.1）}\\
=&H(p_1',p_2',p_3,\dots). \quad \text{（假设三）}
\end{aligned}$$

□

命题 1.4 *当 $p_1=\cdots=p_n=1/n$ 时 H 取得最大值 $\log n$.*

证明 若存在 $p_i\neq p_j$，因为 $\sum_i p_i/n=1/n$，根据鸽巢原理，则必有 i,j 满足 $p_i<1/n<p_j$. 根据命题 1.3，我们可以将 p_i 和 p_j 替换为 $1/n$ 和 p_i+p_j-1/n，而 H 增大. 只要还有两个 p_i 不相等，这一过程就可以重复，每一次都会增大 H，直到所有 p_i 都等于 $1/n$. □

至此，命题 1.2 和命题 1.4 证明了一般情形的命题 1.1. 在等可能的时候不确定性最大，熵最大；在确定事件的时候不确定性最小，熵最小. 所以熵是符合直观的定义.

接下来，我们讨论熵的拓展形式.

在一次试验中，我们可以观察多个变量，比如说 X 和 Y. 我们也可以说，我们观察到了一个结果 (X,Y)，服从分布 $p(i,j)$. 因此有对应的熵，这就是**联合分布的熵**：

$$H(X,Y)=-\mathbb{E}[\log p(X,Y)].$$

对应地，我们也可以写成和的形式：

$$H(p)=-\sum_{i,j}p(i,j)\log p(i,j).$$

自然，联合分布也可以引出**边缘分布的熵**：

$$H(X)=-\mathbb{E}[\log p_X(X)]=-\sum_i\sum_j p(i,j)\log\sum_j p(i,j).$$

$$H(Y)=-\mathbb{E}[\log p_Y(Y)]=-\sum_j\sum_i p(i,j)\log\sum_i p(i,j).$$

有了两个随机变量，我们就可以讨论“条件”的概念. 具体来说，我们可以把试验分为两步，第一步观测 X，第二步观测 Y，那么，第二步所产生的熵就是已经知道第一步结果之后的熵，即：

$$H(Y|X=x) = -\mathbb{E}[\log p_{Y|X=x}(Y)|X=x] = -\sum_j p_{Y|X=x}(j)\log p_{Y|X=x}(j),$$

其中 $p_{Y|X=x}(j) = p(x,j)/p_X(x)$. 当我们知道了 $X=x$ 之后，对 Y 的观测就消除了部分的不确定性，因此根据我们对于不确定性和信息关系的讨论，从 $X=x$ 中获得的关于 Y 的信息是

$$I(X=x:Y) = H(Y) - H(Y|X=x).$$

考虑一个特殊情况，$Y=X$，那么刚刚的讨论就变成了自己从自己身上获得的信息，或者说知道 $X=x$ 带来的信息量. 首先有

$$p_{X|X=x}(i) = \begin{cases} 1, & i=x \\ 0, & i\neq x. \end{cases}$$

因此，

$$H(X|X=x) = -\sum_j p_{X|X=x}(j)\log p_{X|X=x}(j) = -1\log 1 = 0.$$

于是，

$$I(X=x:X) = H(X) - H(X|X=x) = H(X).$$

这正是定量版本的“消除不确定性的东西被称之为信息”! 此外，我们之前说过，熵刻画的是一族可能对象的信息，这一点也反映在了这一公式中：只要知道了 X 的值，无论它具体是多少，我们得到的信息量是一样的！

再回到一般情况，还是同样的两步试验，我们定义给定 X 时 Y 的**条件熵**为

$$\begin{aligned} H(Y|X) =& \mathbb{E}[H(Y|X=x)] \\ =& -\mathbb{E}[\log p_{Y|X}(Y)] \\ =& -\sum_x p_X(x)\sum_j p_{Y|X=x}(j)\log p_{Y|X=x}(j) \\ =& -\sum_{x,j} p(x,j)\log p_{Y|X=x}(j). \end{aligned}$$

换言之，我们现在进一步假定 X 也是不知道的，于是 $H(Y|X)$ 就是平均上来说第二步中 Y 的不确定性. 条件熵和熵有着类似的性质：

命题 1.5 $H(Y|X)\geq 0$，等号成立当且仅当 Y 是退化的，即 Y 概率 1 只取一个值.

证明 仿照命题1.2的证明即可. □

类似地，我们可以考虑平均上 Y 中包含的关于 X 的信息量：

$$\mathbb{E}[I(X=x:Y)]=H(Y)-H(Y|X).$$

与之相对应地，平均上 X 中包含的关于 Y 的信息量为

$$\mathbb{E}[I(Y=y:X)]=H(X)-H(X|Y).$$

一个自然的问题是，二者相互包含的信息量是什么关系？根据概率的链式法则，$p(x,y)=p_{X|Y}(x|y)p_Y(y)$，带入 $H(X,Y)$ 的定义得熵的链式法则：

命题 1.6 *对任意离散随机变量 X,Y，$H(X,Y)=H(Y)+H(X|Y)$.*

利用链式法则，我们注意到，$H(X)-H(X|Y)=H(X)-(H(X,Y)-H(Y))=H(X)+H(Y)-H(X,Y)=H(Y)-H(Y|X)$. 所以，$X$ 中包含的 Y 的信息和 Y 中包含的 X 的信息是一样多的！此外，直观上我们还应该觉得，信息量不能是负的，实际上的确如此：

命题 1.7 *$H(X)-H(X|Y)\geq 0$，等号成立当且仅当 X 和 Y 相互独立.*

我们将在第1.2节看到，命题1.7就是K-L散度信息不等式的一个特例，所以我们就不在这里给出证明了. 命题1.7表明知道任何信息都不会增加不确定性，这个原理被称为“Information doesn't hurt.”根据以上讨论，我们可以自然地定义 X 和 Y 的**互信息**为 $I(X;Y)=I(Y;X)=\mathbb{E}[I(X=x:Y)]=\mathbb{E}[I(Y=y:X)]$.

类似联合分布的熵，条件熵和互信息的概念也可以推广到多元情形. 对于三个随机变量 X,Y,Z，我们可以定义条件熵为

$$H(X,Y|Z)=H(X,Y,Z)-H(Z).$$

类似地，我们可以定义互信息为

$$I(X,Y;Z)=H(X,Y)-H(X,Y;Z).$$

他们的含义以及性质和二元情形类似.

同样，我们可以定义**条件互信息**为 $I(X;Y|Z)=H(X|Z)-H(X|Y,Z)$，表明 Z 已知时候 Y 给 X 带来的平均信息增益. 类似互信息，我们如下性质：

命题 1.8 *条件互信息满足以下性质：*

1. 非负性：$I(X;Y|Z) \geq 0$，等号成立当且仅当 X 和 Y 在给定 Z 的条件下相互独立.

2. 对称性：$I(X;Y|Z) = I(Y;X|Z)$.

3. 链式法则：$I(X,Y;Z) = I(X;Z|Y) + I(Y;Z)$.

4. 条件信息量：$I(X:X|Y) = H(X|Y) - H(X|X,Y) = H(X|Y)$.

最后一条性质说的其实是，在平均的意义下，给定 Y 的时候，知道 X 所能够得到的额外信息量就是 $H(X|Y)$. 这一命题的证明和前面都非常相似，我们留做习题.

最后，我们将各种熵以及信息量的关系总结为图 1.2. 在集合论中，这样的图被称为 Venn 图，所以我们可以用集合论来理解信息与熵. 对应关系可以总结为表 1.1.

信息论	集合论
$H(X)$	A
$H(Y)$	B
$H(X\|Y)$	$A \setminus B$
$H(X,Y)$	$A \cup B$
$I(X;Y)$	$A \cap B$

表 1.1: 信息论和集合论的对应关系.

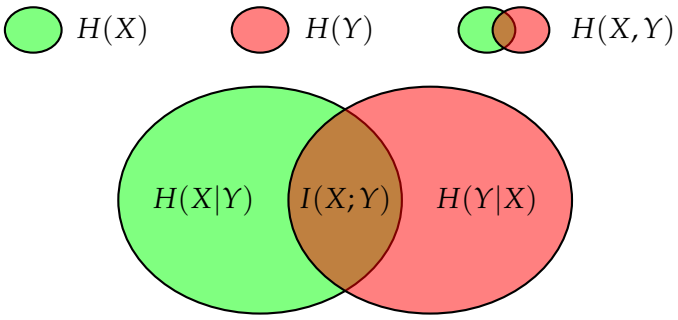


图 1.2: 熵和信息量的关系.

1.1.3 熵与通信理论

最早的时候，Shannon 建立信息论，就是为了给通信理论一个数学基础. 从通信的角度出发，我们可以更本质地理解信息和熵.

通信就是一个发射端和一个接收端，中间有信道传递消息. 将所有可能要传递的消息集合记为 Ω（一个有限集），我们现在考虑 Ω 所蕴含的信息量是多少. 注意到，根据 Shannon 的思想，Ω 里面具体是什么并不重要，重要的是有多少个. 我们可以用自然数 $1, 2, \dots$ 表示集合 Ω 里的元素. 那么，使用二进制编码，我们至少需要 $\log_2 |\Omega|$ 个比特来表示 Ω 里的元素. 于是，假如说随机变量 X 表示收到的消息，那么 X 的熵就定义为 $H(X) = \log_2 |\Omega|$，它衡量了接收端收到的消息的不确定性. 当我们选定了具体的消息 $m \in \Omega$，X 的不确定性被消除了，于是 $X = a$ 的过程产生了（或者说传递了）$\log_2 |\Omega|$ 比特的信息. 比如说，我们发送一个长为 n 的二进制序列，消息的集合大小就是 2^n，发送任何一条具体的消息，我们就传递了 n 比特的信息.

有时候，我们会把消息看成一个序列. 具体来说，我们可以发送独立的 k 条消息，其中第 i 条 X_i 来自消息集合 Ω_i，$|\Omega_i| = n_i$，那么 $(X_1, \dots, X_k)$ 的熵就是

$$H(X_1, \dots, X_k) = \log_2 n_1 + \cdots + \log_2 n_k,$$

它衡量了 k 条消息的不确定性. 在更常见的情况下，每次发送的其实不是一条消息，而是一个字母，所有的字母组成了一个字母表，我们用 $\Sigma = \{x_1, \dots, x_s\}$ 来表示. 于是，X_i 就是消息的第 i 个字母，于是，一条消息可以写作 $X_1 \dots X_k$，其中每一个 X_i 都来自 Σ.

我们现在考虑更加简单的情形，即每一个字母 X_i 其实是同一个随机变量 X 的独立采样. 如果我们具体知道某一个 x_i 出现的次数，那么我们其实可以有更高效的传递信息的方式. 譬如说，在极端情况下，如果只有 x_1 和 x_2 会出现，那么我们其实只需要 $\log_2 2 = 1$ 比特就足够传递所有消息了. 在一般情况下，考虑 Ω 中只包含长为 k 的消息，并且 x_i 在消息中出现 k_i 次，那么所有可能的消息数量为

$$N(k) = \frac{k!}{k_1! \cdot \dots \cdot k_s!}.$$

假定我们需要 $h(\omega)$ 比特来具体确定发的消息是 ω. 首先，无序集合本身需要 $\log_2 |\Omega|$ 比特来编码，其次，我们还需要确定 $(k_1, \dots, k_s)$，确定它的一种方式是按照顺序给出每一个 k_i. 每个 k_i 最多需要 $\log_2 k$ 比特来表示，所以按顺序表示所有的 k_i 至多需要 $s \log_2 k$ 比特. 于是，我们需要的比特数为

$$\log_2 \frac{k!}{k_1! \cdot \dots \cdot k_s!} \le h(m) \le s \log_2 k + \log \frac{k!}{k_1! \cdot \dots \cdot k_s!}.$$

这刚好和我们在统计力学中推导熵的过程是一致的！假设消息足够的长，x_i 出现的频率逐渐接近 p_i，那么同样的推理我们可以知道，

$$h(m) \sim -k \sum_i p_i \log_2 p_i = k H_2(p_1, \dots, p_s).$$

因此，如果知道字母的出现频率，我们传递单位长度的消息至少需要 $H(p_1,\dots,p_s)$ 比特，这完全给出了熵的具体含义，而且，我们现在也不难理解熵的形式为何会出现 log 了：熵就是期望上编码一个字母需要的比特数（即 $\log(1/p(X))$）.

那么，是否有一种编码确实达到了这个理论上的编码长度下界呢？答案是肯定的，它被称为 *Huffmann 编码*. 它的核心思想在于把出现频率高的字母用更短的编码表示. 类似的思想被用在了机器学习的决策树中，作为选择节点非常常用的一种依据.

> **注.** 决策树是一种常用的机器学习分类模型. 假设数据有很多属性 $P_1,\dots,P_k$，这些属性共同决定了某一条数据的类别. 比如，在银行的信用系统中，给定了一个人的性别、是否已婚、是否负债等信息，我们希望给他评估一个信用评级. 决策树的做法是，将决策过程写成一棵树，然后叶节点是决策类别的结果. 比如说，我们会先看这个人是否负债，如果不负债，那么看是否已婚，如果已婚，那么我们信用评级就给 *A*. 那么，如何选择每个节点需要去判断的属性呢？树本身其实就是一种广义的消息，从根节点沿着树走到叶节点得到的就是一条消息. 直观上，如果先选择带来信息增益比较高的属性，那么我们就可以用更少的比特来表示这条消息，或者说，我们决策树结构更加简单. 这样的选择方式叫做 ID3 策略.

我们进一步的问题是，为什么我们知道了每个字母的频次就可以压缩编码？我们接下来将要说明，其实长为 k 的消息中的“典型的消息”数量是远远少于所有 k 长消息的数目，因此我们实际上相当于只是针对一个子集进行编码. 注意到，当 k 充分大的时候，

$$\log_2 N(k) \sim h(m) \sim kH_2(p_1,\dots,p_s).$$

因此，

$$N(k) \approx 2^{kH_2(p_1,\dots,p_s)} = \mathrm{e}^{kH(p_1,\dots,p_s)}.$$

然而，长为 k 的所有消息数目为

$$s^k = \mathrm{e}^{k\log s}.$$

根据命题 1.4，只有当所有 p_i 相等的时候 $N(k)$ 才会达到这一量级. 从这个意义上说，熵所刻画的信息量定量刻画了数据压缩可能的极限.

以上关于信息编码下界以及数据压缩的讨论，再更一般的情况下也成立，此时这样的性质被称为*渐近等分性*. 而这一性质成立对应的结果被称为 *Shannon–McMillan–Breiman* 定理，它的陈述以及证明都需要用到更多随机过程的知识，这里就不再给出了.

> **注.** 现代的主流信息论都是从 *Shannon* 发展起来的. 然而，这一信息论也有很多问题. 首先，信息论使用了概率论进行建模. 但我们已经看到，概率要么是作为频率的近似理论（频率学派），要么反映了人们对未知的信念（主观学派）. 无论哪种解释，都将问题简化了. 正如 *Kolmogorov* 所说：“如果事情没有按照我们的预期发展，那么问题一定出在我们对于概率和真实世界的随机之间关系不清晰的认识上.”其次，这一信息论考虑的是一族对象的信息. 我

们是否能够用这样的方式来衡量单个对象的信息量呢？比如，我们要考虑这本书中包含的信息量，是它放在所有可能的书的集合中去考虑呢，还是把它的每一个章节分开考虑成一个随机序列呢？因此，信息论并不能很好地回答“单个对象”的信息量的问题.

现代概率论的奠基人 *Kolmogorov* 也非常严肃地考虑了这一问题. 他提出了被后世称为 ***Kolmogorov* 复杂度**的概念，旨在刻画一个随机字符串的随机程度. 简单来说，一个字符串的 *Kolmogorov* 复杂度就是描述它所需要的最短的代码长度，越随机的字符串就越需要更复杂的程序去描述它的产生方式. 利用这一概念，我们可以将信息的概念变成一个对象自己的属性，而不再需要把对象放在可能的一堆对象中去考虑. 这是信息论的另一种构建思路.

§1.2 Kullback-Leibler 散度

1.2.1 定义

为了引入 K-L 散度，我们从互信息出发. 它的定义是：

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= -\sum_x p_X(x)\log p_X(x) + \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p_Y(y)} \\ &= -\sum_{x,y} p(x,y)\log p_X(x) + \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p_Y(y)} \\ &= \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p_X(x)p_Y(y)}. \end{aligned}$$

根据命题 1.7，$I(X;Y)\geq 0$，等号成立当且仅当 X,Y 相互独立，即 $p(x,y)=p_X(x)p_Y(y)$. X,Y 之间的互信息越大，说明他们之间的关联越强，分布越不独立，$p(x,y)$ 越不接近 $p_X(x)p_Y(y)$. 实际上，这样的想法可以被推广到一般分布上.

我们从数理统计的视角出发，考虑两个概率分布的似然函数p_1 和 p_2（也就是他们的分布列）. 抽取一个样本 X，考虑假设检验问题：

$$H_1\text{：样本 } X \text{ 来自 } p_1 \text{ 的分布} \quad \text{vs.} \quad H_2\text{：样本 } X \text{ 来自 } p_2 \text{ 的分布}$$

假设检验中有一种很常用的技巧，称为*似然比检验法*，即考虑两个假设分布的似然比 p_1/p_2. 如果这个比值越大，就越说明 p_1 的值更大，因而更有可能，倾向于接受 H_1，反之则越倾向于接受 H_2. 于是，可以自然定义区分 H_1 和 H_2 的检验量为对数似然比：

$$\log(p_1(x)/p_2(x)).$$

假设 H_1 是真的，那么在 H_1 的世界里，这个检验量的期望为

$$\mathbb{E}_{X\sim p_1}(\log(p_1(X)/p_2(X))) = \sum_i p_1(i)\log\frac{p_1(i)}{p_2(i)}.$$

实际上，上面的期望就是 K-L 散度的定义.

定义 1.3 (Kullback-Leibler 散度) *对于两个概率分布列 p_1, p_2，他们的* ***Kullback-Leibler 散度****或***相对熵***被定义为*

$$D(p_1\|p_2) = \mathbb{E}_{X\sim p_1}(\log(p_1(X)/p_2(X))) = \sum_i p_1(i)\log\frac{p_1(i)}{p_2(i)}.$$

其中规定 $0\log(0/0) = 0$，$0\log(0/a) = 0$，$a\log(a/0) = +\infty$.

我们马上知道，互信息是 K-L 散度的一种特殊情况：

命题 1.9 *对于两个随机变量 X, Y，成立 $I(X;Y) = D(p_{X,Y}\|p_X p_Y)$，其中 $p_{X,Y}$ 是 X, Y 的联合分布列，p_X, p_Y 分别是 X, Y 的边缘分布列.*

K-L 散度可以看成两个分布之间的区分衡量标准，但他不是度量. 一般来说，甚至连对称性都不成立. 例如，设 p_1 和 p_2 都是定义在 $0, 1$ 上的 Bernoulli 分布，参数分别为 $1/2$ 和 $1/4$. 于是

$$D(p_1\|p_2) = \frac{1}{2}\log\frac{1/2}{3/4} + \frac{1}{2}\log\frac{1/2}{1/4} = \frac{1}{2}\log\frac{4}{3}.$$

$$D(p_2\|p_1) = \frac{3}{4}\log\frac{3/4}{1/2} + \frac{1}{4}\log\frac{1/4}{1/2} = \frac{1}{2}\log\frac{3\sqrt{3}}{4}.$$

这两个值是不相等的.

我们在定义中还提到了 K-L 散度的另一个名字——相对熵. 实际上，这可以从编码中看出来. 假设事实上消息中字母的分布是 p_1，那么期望上编码单位长度消息需要的比特数是 $H(p_1) = \mathbb{E}_{X\sim p_1}[\log p_1(X)]$. 如果我们错误地认为消息中字母的分布是 p_2 并使用最优编码，那么实际上期望编码单位长度消息需要的比特数是 $\mathbb{E}_{X\sim p_1}[\log p_2(X)]$. 由于错误的认识所产生的额外编码长度是

$$\mathbb{E}_{X\sim p_1}[\log p_1(X) - \log p_2(X)] = D(p_1\|p_2).$$

根据在第 1.1.3 节 中的讨论，我们知道，额外的编码长度代表的是额外的不确定性，因而这一概念是某种“熵”的概念. 这正是“相对熵”的由来，$D(p_1\|p_2)$ 表示了当我们错误地把 p_1 当成 p_2 时带来的额外的不确定性，或者说额外的信息损失.

在机器学习中，比起讨论 K-L 散度，更加常用的是直接讨论量 $\mathbb{E}_{X\sim p_1}[\log p_2(X)]$. 从机器学习的观点来说，$p_1$ 是真实的分布，而 p_2 是我们所学习到的分布. 根据刚刚的讨论，这个量越小越说明 p_2 接近真实的 p_1，因此这又是一种衡量两个分布之间关系的量，我们称之为*交叉熵*：

定义 1.4 (交叉熵) 给两个随机变量 X,Y，X 的分布为 p_X，Y 的分布为 p_Y，则 X 的分布 p_X 和 Y 的分布 p_Y 的**交叉熵**[1]为

$$CH(p_X,p_Y)=\mathbb{E}_{X\sim p_X}[\log p_Y(X)]=-\sum_i p_X(i)\log p_Y(i).$$

在机器学习的分类问题中，我们希望学习到的分布 p_Y 尽可能地接近真实的分布 p_X，所以我们训练的目标经常是最小化交叉熵 $CH(p_X,p_Y)$. 有趣的是，从数理统计的角度来看，最小化交叉熵等价于进行最大似然估计，因此这为最大似然估计提供了一种信息论意义下的理解. 相关讨论留作练习.

1.2.2 两个关于信息的不等式

利用 K-L 散度，我们可以给出两个关于信息的不等式，它们分别是信息不等式和数据处理不等式.

定理 1.2 (信息不等式) 对于两个概率分布列 p,q，成立 $D(p\|q)\geq 0$，当且仅当 $p=q$ 时取等号.

证明 由于 $\log x$ 是凸函数，所以由 Jensen 不等式，我们有

$$D(p\|q)=-\mathbb{E}_{X\sim p}\left[\log\frac{q(X)}{p(X)}\right]\geq-\log\mathbb{E}_{X\sim p}\left[\frac{q(X)}{p(X)}\right]=-\log\sum_i p(i)\cdot\frac{q(i)}{p(i)}=0.$$

因此，$D(p\|q)\geq 0$，当且仅当 $p=q$ 时取等号. □

信息不等式表明，K-L 散度虽然不是度量，但却是非负的，因而确实可以被作为熵，用来衡量“额外的不确定性”. 此外，命题 1.7 是信息不等式的直接推论. 利用类似的方法，我们可以证明条件互信息的非负性（即命题 1.8 中的第一条）.

接下来我们叙述并证明数据处理不等式.

定理 1.3 (数据处理不等式) 假设随机变量 X,Y,Z 形成了 *Markov* 链，那么 $I(X;Y)\geq I(X;Z)$. 特别地，对任意函数 f，成立 $I(X;Y)\geq I(X;f(Y))$.

证明 根据互信息链式法则，

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

[1]文献中，经常会直接写为 $H(p_X,p_Y)$，但是在本书中为了区分熵，我们使用了符号 CH.

根据 Markov 性，条件在 Y 上，X 和 Z 相互独立. 因此，$I(X;Z|Y)=0$，根据条件互信息的非负性，$I(X;Y|Z)\geq 0$，所以 $I(X;Y)\geq I(X;Z)$.

显然，$X,Y,f(Y)$ 也形成了 Markov 链，所以 $I(X;Y)\geq I(X;f(Y))$. □

数据处理不等式表明，无论我们对随机变量 Y 进行了何种处理，甚至是允许带随机的处理，它的信息量都不会增加.

1.2.3 在机器学习中的应用：语言生成模型

现如今，机器学习中最为瞩目的成果之一就是*大语言模型*，它通过学习人类海量的高质量语料库来形成一个生成式的模型，其中最为典型的例子是 ChatGPT. 从思路上来说，大语言模型的核心思想非常简单：给一段话，将其中一些词掩盖掉，让模型填出这些词来. 例如，给出“我在 [mask] 面条，真好吃“，模型应该能够填出“我在吃面条，真好吃“. 这样的思想，对于更一般的数据也是成立的：用（修改改过的）数据本身作为输入，训练一个*编码器*，然后将编码器的输出送入*解码器*，而解码器的输出具有原始数据的格式，我们希望这一输出能够尽量匹配原始的输入. 在自然语言处理中，一个*生成模型*往往同时有编码器和解码器. 比如说，图 1.3 展示的就是 BART [LLG+19] 的结构.

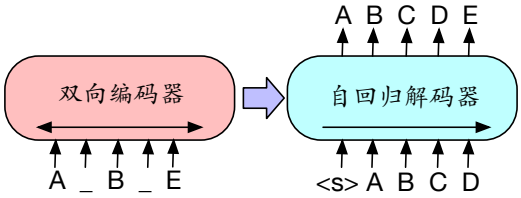


图 1.3: 生成式语言模型 BART 的示意图.

我们已经指出，熵和编码有着密切的联系. 从这个角度出发，我们很容易理解生成模型背后的思想：我们希望通过训练的方式得到一个由神经网络所表示的编码和解码规则，他要尽可能符合真实数据的分布.

我们可以用一种非常简单的模型去理解这一过程. 假设所有的单词的集合为 Σ，单词数为 k 的文本集合为 Ω. 我们希望训练一个生成模型 M，给它输入 $k-1$ 个单词，它可以给出第 k 个单词的概率分布，我们选择出现概率最大的那个词作为预测. 在训练的时候，对于一个句子 ω，我们只保留前 $k-1$ 个词，得到 $\omega[1:k]$，然后将它输入到生成模型 M 中，让它去预测第 k 个词.

对于这一个具体的句子来说，理想的分布应该是一个 *Dirac* 分布[2] $\delta_{\omega[k]}$，即以概率 1

[2]Dirac 分布是一个数学物理中更加常用的名字. 在概率论中，这也被称为*退化分布*；而在机器学习中，分

取到 $\omega[k]$. 假如说生成模型的输出是一个概率分布 $M(\omega[1:k-1]) = p$，那么，我们可以用 K-L 散度去衡量这两个分布的差异，因为 $H(\delta_{\omega[k]})$ 是固定的，所以我们只考虑交叉熵 $CH(\delta_{\omega[k]}, p)$. 因为一次训练会给多个样本，所以我们的目标是同时最小化这些交叉熵的和. 假如训练集是 T，我们的目标就是

$$\min_M \sum_{\omega \in T} CH(\delta_{\omega[k]}, M(\omega[1:k-1])).$$

实际上，这个例子是有普适性的，所有的监督训练的分类问题都可以用这种方式来建模. 而在 **??** 我们也会看到，此时交叉熵实际上被作为了一种*损失函数*.

§1.3 附录：Shannon 定理的证明

我们在这一部分给出 Shannon 定理（定理 1.1）的证明. 整体上的思路是：

1. 证明如果 f 是单调函数，对正整数 m, n 成立 $f(mn) = f(m) + f(n)$，那么 $f(n) = C \log n$.
2. 求出 $H(1/n, \ldots, 1/n)$ 的表达式.
3. 假设 p_i 是有理数，设 $p_i = n_i / \sum_j n_j$，考虑 $\sum_j n_j$ 个等可能试验结果，利用假设 3 推出 H 的表达式.
4. 利用有理数的稠密性和 H 的连续性推出一般情形.

最后一步是显然的，我们只需要证明前三步即可.

对第一步，我们需要证明的是，如果 f 是单调函数，对正整数 m, n 成立 $f(mn) = f(m) + f(n)$，那么 $f(n) = C \log n$. 首先，利用数学归纳法容易看出，对正整数 n, k，成立

$$f(n^k) = kf(n). \tag{1.2}$$

设 m, n 是任意两个大于 1 的整数，再选任意大的正整数 k，从 m 进制数的性质可以看出，总存在正整数 l 使得

$$m^l \le n^k < m^{l+1}. \tag{1.3}$$

根据 f 的单调性，我们有

$$f(m^l) \le f(n^k) < f(m^{l+1}).$$

利用式 (1.2)，我们有

$$lf(m) \le kf(n) < (l+1)f(m) \iff \frac{l}{k} \le \frac{f(n)}{f(m)} < \frac{l+1}{k}.$$

将式 (1.3) 取对数，得到

$$l \log m \le k \log n < (l+1) \log m \iff \frac{l}{k} \le \frac{\log n}{\log m} < \frac{l+1}{k}.$$

所以

$$\left| \frac{\log n}{\log m} - \frac{f(n)}{f(m)} \right| \le \frac{1}{k}.$$

布经常会表示为一个概率向量，文献中称为*独热向量*.

因为 k 可以是任意大的正整数，取 $k \to \infty$，我们就得到了

$$\frac{\log n}{\log m} = \frac{f(n)}{f(m)}.$$

由 m,n 的任意性，取 $m = 2$，我们就得到了 $f(n) = (f(2)/\log 2) \cdot \log n = C \log n$. 容易检验，$f(1) = 0 = C \log 1$，因此这一等式对所有正整数 n 都成立.

对第二步，我们需要求出 $f(n) = H(1/n, \ldots, 1/n)$ 的表达式. 我们要利用第一步的结果，首先，根据假设二，$f(n)$ 是单调递增的函数. 然后，考虑 mn 个等可能试验，我们可以将它分成两步试验，第一步有 m 中等可能的结果，而在每一种结果之下，第二步有 n 种等可能结果. 根据假设三，

$$f(mn) = f(m) + \frac{1}{n} \cdot nf(n) = f(m) + f(n).$$

所以 $f(n)$ 符合第一步的假设. 第二步就可以直接从第一步推出.

最后，我们证明第三步. 设 $p_1, \ldots, p_n$ 都是有理数，那么，他们可以被写为

$$p_i = \frac{n_i}{\sum_{j=1}^{n} n_j}.$$

其中 n_i 是非负整数. 我们考虑 $\sum_j n_j$ 个等可能试验，这个试验可以被看成两步的试验，第一步有 n 种可能的结果，第 i 种结果出现的概率是 p_i，而在第 i 种结果之下，第二步有 n_i 种等可能的结果. 根据假设三，和证明的第三步，我们有

$$C \log \sum_{j=1}^{n} n_j = H(p_1 + \cdots + p_n) + \sum_{i=1}^{n} p_i \cdot C \log n_i.$$

因此，

$$\begin{aligned}
H(p_1, \ldots, p_n) &= C \left(\log \sum_{j=1}^{n} n_j - \sum_{i=1}^{n} p_i \log n_i \right) \\
&= C \left(\log \sum_{j=1}^{n} n_j - \sum_{i=1}^{n} p_i \log \left(p_i \sum_{j=1}^{n} n_j \right) \right) \\
&= C \left(\log \sum_{j=1}^{n} n_j - \sum_{i=1}^{n} p_i \log p_i - \sum_{i=1}^{n} p_i \log \sum_{j=1}^{n} n_j \right) \\
&= -C \sum_{i=1}^{n} p_i \log p_i.
\end{aligned}$$

这正是我们要证明的. 于是，我们证明了 Shannon 定理.

§1.4 习题

1. 我们在熵以及 K-L 散度的定义中，都规定了一些无定义的量的值，这些值并不是随便规定的，他们实际上反映了熵或者 K-L 散度定义中的连续性.

(1) 证明：对给定的 $a > 0$，$\lim_{x \to 0+} x \log(x/a) = 0$，因此我们规定了 $0 \log 0 = 0$ 以及 $0 \log(0/a) = 0$.

(2) 证明：对给定的 $a > 0$，$\lim_{x\to 0+} x\log(a/x) = +\infty$，因此我们规定了 $0\log(a/0) = +\infty$.

2. 考虑关于 n 的正实数序列 $a_1(n),\dots,a_k(n)$ 以及 $b_1(n),\dots,b_k(n)$，假设对所有 i，都成立 $\lim_{n\to\infty} a_i(n)/b_i(n) = 1$，证明：

$$\lim_{n\to\infty} \frac{a_1(n)+\cdots+a_k(n)}{b_1(n)+\cdots+b_k(n)} = 1.$$

由此证明式 (1.1).

3. 证明命题 1.1.

4. 用 Lagrange 乘子法重新证明命题 1.4.

提示：如果你不知道 *Lagrange* 乘子法，可以参考 **??**.

5. 证明命题 1.8.

6. [Tin62] 仿照集合论的思路，我们可以定义三个随机变量的互信息为：

$$I(X;Y;Z) = I(X;Y) - I(X;Y|Z).$$

(1) 证明对称性：$I(X;Y;Z) = I(Y;X;Z) = I(X;Z;Y)$.

(2) 举一个例子说明，可能会有 $I(X;Y;Z) < 0$，所以这样定义的互信息并不一定真的代表“信息量”.

7. 举一个例子说明，即便 $D(p_1\|p_2)$ 很接近 0，$D(p_2\|p_1)$ 也可能会很大.

8. （单变量数据处理不等式）对任意离散随机变量 X 和函数 f，证明：$H(X) \geq H(f(X))$.

9. 考虑二分类的学习问题，此时对单个样本我们观察到的结果要么是 0 或 1，假设在真实世界中样本总体服从参数为 θ 的 Bernoulli 分布，即 $\Pr(X=1) = 1-\Pr(X=0) = \theta$. 假设我们的数据集是 $(x_1,y_1),\dots,(x_N,y_N)$，他们是从总体中独立采样得到的.

(1) 将问题考虑成一个数理统计问题，估计 θ. 写出似然函数 $L(\theta;y_1,\dots,y_N)$.

(2) 再将问题考虑为一个信息论问题，写出每个样本的真实分布与估计分布之间的交叉熵之和 $CH(\theta;y_1,\dots,y_N)$.

(3) 证明：$\max_\theta L(\theta;y_1,\dots,y_N) = \min_\theta CH(\theta;y_1,\dots,y_N)$，也就是说，最大似然估计等价于最小化交叉熵.

10. 请查找文献回答以下问题：

(1) Fisher 信息量是什么？它与 K-L 散度有什么样的关系？

(2) 列举其他概率分布之间散度的概念，他们是否是度量？

(3) 列举概率分布之间的度量，他们之间是否有关联？

§1.5 章末注记

信息一词的英文是“information”，从动词“inform”来，意思是告知、通知. 早在 15 世纪中叶，“information”一词的出现了义项“在通信中针对特定主题的知识”.[Inf] 这说明在那个时候人类就已经意识到，通信会产生新的东西，被称为知识或信息. 然而，人类对信息的严谨探索起步晚得多. 关于信息的物理学讨论源自统计力学，Boltzmann 提出了著名的熵，证明了 H 定理，以此给出了热力学第二定律的微观解释. 关于 Boltzmann 的工作，参见 [Uff22].

一般认为，现代信息论的起源是 Shannon 的论文 [Sha48]，他在论文中提出了信息的数学定义，以及信息的基本性质. 但是，Shannon 的工作并不是孤立的，他的工作是在统计力学的基础上发展起来的. 事实上，Shannon 在论文中也提到了 Boltzmann 的熵. 这篇工作也被视为通信理论以及编码理论的奠基性工作. Shannon 在这篇论文中还给出了渐近意义下达到理论下界的最优编码，并且独立地被 Fano [Rob49] 以一种不同的形式发现，因此后世称为 Shannon-Fano 编码. 但是 Shannon-Fano 编码并不是精确地达到下界，实际上，最优编码是 Huffman [Huf52] 给出的. Shannon 在这篇论文中还讨论了渐近等分性，后来 MciMillan 的工作 [McM53] 和 Breiman 的工作 [Bre57] 拓展了这一结果，因此后世称为 Shannon-McMillan-Breiman 定理.

关于信息论与集合论的关系工作，可以参见 Hu Kuo Ting 的工作 [Tin62]. 他的工作还给出了多个随机变量互信息的定义，在这一章习题中有涉及.

相对熵的概念依然是从 Shannon 的奠基性论文 [Sha48] 中提出的，但他只局限于通信的问题. 更加一般的讨论是由 Kullback 和 Leibler 在 [KL51] 给出，他们的是一种数理统计的思路，但是他们也具体地讨论了这一概念与信息的关系. 他们的论文中也讨论了交叉熵这一概念.

机器学习中编码器和解码器的思路，最早是由 Rumelhart，Hinton 和 Williams 在 [RHW86] 中提出，他们将编码器和解码器的整体称作*自编码器*. 这篇工作几乎可以被视为深度学习的开山之作，它还提出了训练神经网络最常用的反向传播算法.

关于信息论的经典教科书，可以参见 [CT12] ，此外，概率论的教材中也有很多很好的讨论，比如 [Jay02]，[Shi96] 以及 [李 10].

第二章　Johnson-Lindenstrauss 引理

我们已经在上一章看到，使用概率分布建模的信息论在机器学习中起到了举足轻重的作用。基于概率论的信息论总是考虑一个集合的对象的信息量，因此**数据**成为了这种方法论的核心前提：数据表征了一个集合的对象的某一特征。在这一章，我们将探讨机器学习中数据的特性，以及一种重要的数据压缩的原理：Johnson-Lindenstrauss 引理。证明这一引理所用到的概率论技术是**矩法**，这是*机器学习理论*中最为核心的几个技术之一。因此本章也可以看做机器学习理论的一个引论。

§2.1　机器学习中的数据

从编码的角度来说，数据最简单的表示方法是使用固定长度的字符串。比如说，人的生理性别有男或者女两种，于是我们可以用字符串 0 表示男，1 表示女。这样，我们就可以用一个长度为 1 的字符串来表示人的生理性别。人的属性还有很多，比如说年龄、身高、体重、学历、职业等等，这些属性都可以分别用固定长度的字符串来表示。于是，一个人就被抽象为了一个固定长度的字符串。

然而，这种表示方式必须要假定数据只取有限个值。有时候，为了简化建模和计算，我们还会考虑可以取无限个值的数据。我们看一个具体的例子，人的身高。从现代物理的角度来说，身高的变化是离散的，它有一个最小变化的单位。从生物学的角度来说，身高是有上界的，比如说所有人的身高都不会超过十米。因此，从理论上说，身高也只能取有限个值，所以也可以用字符串来表示。然而，更加方便的方式是假定身高是一个*非负实数*，因此用一个数而不是一个字符串来表示。

因此，更加常见的情况下，我们会用*实数*或者*整数*来编码数据。此时，将对象的多种属性按顺序排在一起，我们就得到了一个*向量*。总而言之，在机器学习的框架，*数据被表示成数值向量*. 例如，要表示一个人的年龄、身高、体重、学历、职业，我们需要

1. 身高使用厘米作为单位；体重用千克作为单位；给学历编一个编号，比如 0 是高中，1 是本科，2 是硕士，3 是博士，-1 是其他；给职业也编号，例如 1 表示提示词工程师。

2. 用一个五维向量来表示年龄、身高、体重、学历、职业。例如，$(20, 180, 70, 2, 1)$ 表示一个年龄为 20 岁，身高 180 厘米，体重 70 千克，学历为硕士，职业为提示词工程师的人。

注. *[lhy: 介绍一下计算机中对数值的编码]*。

机器学习中，如此表示数据具备了独特的性质，一言以蔽之：**维数高，但是稀疏**。[lhy: 介绍一下高维高斯分布的特点，以及图像处理中数据稀疏的特点。]

§2.2 矩法与集中不等式

我们先引入示性函数的概念.

定义 2.1 (示性函数) 对事件 A，定义 A 的示性函数为一个从样本空间 Ω 到 $\mathbb{R}$ 的随机变量：

$$I(A)(\omega) := \begin{cases} 1, & \omega \in A. \\ 0, & \omega \notin A. \end{cases}$$

从定义就可以得到如下基本性质：

命题 2.1 设 A, B 是两个事件，则

1. $I(AB) = I(A)I(B)$.
2. $I(A)^2 = I(A)$.
3. $I(A \cup B) = I(A) + I(B) - I(AB)$.

证明 这里只作为一个示意，证明第三点，其他都类似。我们需要证明，对任意样本点 $\omega \in \Omega$，我们有

$$I(A \cup B)(\omega) = I(A)(\omega) + I(B)(\omega) - I(AB)(\omega).$$

假设 $\omega \in A \cup B$，那么左边等于 1。我们分类讨论：

- 如果 $\omega \in A$，那么右边第一项为 1.
  - 如果 $\omega \in B$，那么右边第二项为 1。此时自然也有 $\omega \in AB$，所以右边第三项为 1，因此右边等于 1，等于左边。
  - 如果 $\omega \notin B$，那么右边第二项为 0。此时自然也有 $\omega \notin AB$，所以右边第三项为 0，因此右边等于 1，等于左边。
- 如果 $\omega \notin A$，那么右边第一项为 0。此时必须有 $\omega \in B$，所以右边第二项为 1。但是此时自然也有 $\omega \notin AB$，所以右边第三项为 0，因此右边等于 1，等于左边。

如果 $\omega \notin A \cup B$，讨论类似，这里不再赘述。 □

示性函数之所以重要，是因为它联系了期望与概率。我们先来看一个显然的命题：

命题 2.2 *设 A 是一个事件，则*

$$\mathbb{E}[I(A)] = \Pr(A).$$

示性函数可以把对概率的计算变成对期望的计算。回忆期望的线性性：设 $a, b \in \mathbb{R}$，X, Y 是有期望的随机变量，那么成立

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

利用期望的线性性，示性函数可以导出很多概率恒等式与不等式. 例如：容斥公式

$$\begin{aligned}\Pr(A \cup B) &= \mathbb{E}[I(A \cup B)] = \mathbb{E}[I(A) + I(B) - I(AB)] \\ &= \mathbb{E}[I(A)] + \mathbb{E}[I(B)] - \mathbb{E}[I(AB)] \\ &= \Pr(A) + \Pr(B) - \Pr(AB).\end{aligned}$$

对于概率论以及机器学习理论来说，下面的这个不等式非常重要：

定理 2.1 (Markov 不等式) *如果 X 是非负有期望的随机变量，$a > 0$，那么*

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

证明 直接利用示性函数，我们有：

$$\begin{aligned}\mathbb{E}[X] =& \mathbb{E}[XI(X \geq a) + XI(X < a)] \\ =& \mathbb{E}[\underbrace{XI(X \geq a)}_{\geq aI(X \geq a)}] + \mathbb{E}[\underbrace{XI(X < a)}_{\geq 0}] \\ \geq& a\mathbb{E}[I(X \geq a)] = a\Pr(X \geq a).\end{aligned}$$

□

注. 为了使得证明有效，我们必须要假设上面的推导中出现的期望都是存在的，当然这实际上很容易验证。为了避免不必要的技术细节，在后面的所有证明以及推导中，我们都会默认写出来的期望是存在的，不再赘述。

我们利用 Markov 不等式可以直接得到以下结果.

推论 2.1 (Chebyshev 不等式) 设 X 是任意有方差的随机变量，那么对任意 $a > 0$，成立

$$\Pr(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathsf{Var}(X)}{a^2}.$$

证明 设 $Y = (X - \mathbb{E}[X])^2$，$t = a^2$，那么 Y 是非负随机变量，且 $\mathbb{E}[Y] = \mathsf{Var}(X)$，于是由 Markov 不等式，我们有

$$\begin{aligned}\Pr(|X - \mathbb{E}[X]| \geq a) &= \Pr(|X - \mathbb{E}[X]|^2 \geq a^2) \\ &= \Pr(Y \geq t) \\ &\leq \frac{\mathbb{E}[Y]}{t} = \frac{\mathsf{Var}(X)}{a^2}.\end{aligned}$$

□

Chebyshev 不等式告诉我们采样到偏离其期望的概率有一个上界. 像这样利用矩（即 $\mathbb{E}[f(X)]$）来估计概率上界的方法被称为矩法.

实际上，很多情况下，偏离期望是非常小概率的事件，远小于上面的估计值. 为了得到更精确的上界，我们需要一些技巧. 考虑任意随机变量 X，对 $\lambda > 0$，

$$X \geq a \iff \lambda X \geq \lambda a \iff e^{\lambda X} \geq e^{\lambda a}.$$

由 Markov 不等式（如何得到?），

$$\Pr(X \geq a) = \Pr\left(e^{\lambda X} \geq e^{\lambda a}\right) \leq e^{-\lambda a} \cdot \mathbb{E}\left[e^{\lambda X}\right].$$

注意到这个不等式应该对任意 $\lambda > 0$ 成立，所以

$$\Pr(X \geq a) \leq \inf_{\lambda > 0} e^{-\lambda a} \cdot \mathbb{E}\left[e^{\lambda X}\right].$$

以上方法可以得到概率更精确的上界。这样用指数进行推导的方法称为指数矩或 *Cramér-Chernoff 方法*.

利用指数矩，我们可以更加精确地研究 Chebyshev 不等式中随机变量所表现出来的性质，这种性质被称为概率的集中性. 我们可以用集中不等式来刻画这样的性质。这样的不等式描述随机变量 X 有多大概率偏离某个值 μ 多少值（t），它表现为

$$\Pr(|X - \mu| \geq t) \leq \text{小量}.$$

通常来说，μ 是随机变量的期望或者中位数，在这本书中，只会讨论关于期望的集中性. 我们可以看到 Chebyshev 不等式就是一种特殊的集中不等式，但是它的界太松. 利用指数矩，我们将证明界更紧的 Hoeffding 不等式和 Chernoff 不等式.

定理 2.2 (Hoeffding 不等式) *设 $X_1, \dots, X_n$ 相互独立且服从对称 Bernoulli 分布，即 X_i 满足 $\Pr(X_i = 1) = 1 - \Pr(X_i = -1) = 1/2$. 考虑向量 $a = (a_1, \dots, a_n) \in \mathbb{R}^n$，对任意 $t \geq 0$，我们有*

$$\Pr\left(\sum_{i=1}^{n} a_i X_i \geq t\right) \leq \exp\left(-\frac{t^2}{2\|a\|_2^2}\right).$$

证明 由指数矩，我们有

$$\begin{aligned}
\Pr\left(\sum_{i=1}^{n} a_i X_i \geq t\right) &= \Pr\left(\exp\left(\lambda \sum_{i=1}^{n} a_i X_i\right) \geq \exp(\lambda t)\right) \\
&\leq e^{-\lambda t} \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} a_i X_i\right)\right] \\
&= e^{-\lambda t} \prod_{i=1}^{n} \mathbb{E}\left[\exp(\lambda a_i X_i)\right].
\end{aligned}$$

这个不等式对任意 $\lambda > 0$ 都成立. 利用 $X_1, \dots, X_n$ 服从对称 Bernoulli 分布，得到（习题）[lhy: 习题]

$$e^{-\lambda t} \prod_i \mathbb{E}\left[\exp(\lambda a_i X_i)\right] \leq \exp\left(-\lambda t + \frac{\lambda^2}{2} \sum_i a_i^2\right). \tag{2.1}$$

由于这一不等式对任意 $\lambda > 0$ 都成立，根据二次函数的性质，取 $\lambda = t / \sum_i a_i^2$，可得

$$\begin{aligned}
\inf_{\lambda > 0} \exp\left(-\lambda t + \frac{\lambda^2}{2} \sum_i a_i^2\right) &= \exp\left(-\frac{t}{\sum_i a_i^2} t + \frac{1}{2}\left(\frac{t}{\sum_i a_i^2}\right)^2 \sum_i a_i^2\right) \\
&= \exp\left(-\frac{t^2}{2\|a\|_2^2}\right).
\end{aligned}$$

□

利用相同的证明技巧，我们可以证明一般形式的 Hoeffding 不等式，我们把证明留作习题.[lhy: 习题]

定理 2.3 (Hoeffding 不等式，一般情形) *设 $X_1, \dots, X_n$ 是相互独立的随机变量，对任意 i 都成立 $X_i \in [m_i, M_i]$。那么对任意 $t \geq 0$，我们有*

$$\Pr\left(\sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} (M_i - m_i)^2}\right).$$

下面我们介绍 Chernoff 不等式.

定理 2.4 (Chernoff 不等式) 设 $X_1,\dots,X_n$ 是相互独立的随机变量，分别服从于参数为 $p_1,\dots,p_n$ 的 *Bernoulli* 分布. 记 $\sum_{i=1}^n X_i$ 的期望为 $\mu = \sum_{i=1}^n p_i$，对于任意 $t > \mu$，我们有

$$\Pr\left(\sum_{i=1}^n X_i \geq t\right) \leq e^{-\mu}\left(\frac{e\mu}{t}\right)^t.$$

这里 e 是自然对数的底数.

证明 和证明 Hoeffding 不等式的第一步相同，我们先利用指数矩，对任意 $\lambda > 0$ 有

$$\Pr\left(\sum_{i=1}^n X_i \geq t\right) \leq e^{-\lambda t}\prod_{i=1}^n \mathbb{E}\left[\exp(\lambda X_i)\right].$$

然后，将 $\prod_{i=1}^n \mathbb{E}[\exp(\lambda X_i)]$ 进一步放缩：

$$\begin{aligned}\prod_{i=1}^n \mathbb{E}\left[\exp(\lambda X_i)\right] &= \prod_{i=1}^n \left(e^{\lambda} p_i + (1-p_i)\right)\\ &\leq \prod_{i=1}^n \exp\left((e^{\lambda}-1)p_i\right).\end{aligned}$$

因此

$$\begin{aligned}\Pr\left(\sum_{i=1}^n X_i \geq t\right) &\leq e^{-\lambda t}\prod_{i=1}^n \exp\left((e^{\lambda}-1)p_i\right)\\ &= e^{-\lambda t}\exp\left((e^{\lambda}-1)\sum_{i=1}^n p_i\right)\\ &= \exp\left(\mu e^{\lambda} - t\lambda - \mu\right).\end{aligned}$$

右边的最小值在 $\lambda = \log(t/\mu)$ 取得，代入得到：

$$\Pr\left(\sum_{i=1}^n X_i \geq t\right) \leq e^{-\mu}\left(\frac{e\mu}{t}\right)^t.$$

□

§2.3 J-L 引理的陈述与证明

有了上面矩法的准备，我们可以陈述并证明 J-L 引理了.

定理 2.5 (Johnson-Lindenstrauss 引理) 给定 N 个单位向量 $v_1,\dots,v_N \in \mathbb{R}^m$ 和 $n > 24\log N/\epsilon^2$，随机矩阵 $A \in \mathbb{R}^{n\times m}$ 每个元素独立重复采样自 $\mathcal{N}(0,1/n)$，$\epsilon \in (0,1)$ 是给定常数，那么至少有 $(N-1)/N$ 的概率，使得对所有的 $i \neq j$，都成立

$$(1-\epsilon)\left\|v_i - v_j\right\|_2^2 < \left\|Av_i - Av_j\right\|_2^2 < (1+\epsilon)\left\|v_i - v_j\right\|_2^2.$$

我们可以把 n 理解成降维后的维度，Av_i 是降维后的向量. 这个引理告诉我们只要 $n > 24\log N/\epsilon^2$，我们就可以用变换 A 把原本 m 维的向量映射到 n 维空间，并且保证它们相对距离的偏离不超过 ϵ，因此我们可以把 A 看成一个损失率很低的压缩变换。不严格地说，塞下 N 个向量，只需要 $\mathcal{O}(\log N)$ 维空间.

下面我们开始证明 J-L 引理。为了看出来证明的思路，我们第一个任务是算出压缩后 Av_i 的分布. 我们首先回忆一些正态向量的基本性质。关于正态向量的讨论，可以参考**??**。

命题 2.3 假设 $u \sim \mathcal{N}(\mu,\Sigma)$ 是一个 n 维正态向量，M 是一个 $m\times n$ 矩阵，那么 Mu 是一个 m 维正态向量，并且 $Au \sim \mathcal{N}(M\mu, M\Sigma M^T)$.

利用这一个命题，很容易可以得到 Av_i 的分布：

引理 2.1 假设 $u \in \mathbb{R}^m$ 是一个单位向量，那么 $Au \sim \mathcal{N}(0, n^{-1}I_n)$.

证明 将 A 视作一个 mn 维的正态向量，注意到，$(Au)_i = \sum_{j=1}^m A_{ij}u_j$，所以 Au 是一个从向量 A 线性变换得到的向量。根据命题 2.3，Au 是一个正态向量，只需计算它的期望和协方差矩阵。

注意到，对不同的 i，向量 $(A_{ij})_j$ 相互是独立的，所以分量 $(Au)_i$ 相互也是独立的，因此只需要计算正态变量 $(Au)_i$ 的期望与方差。其期望为 $\sum_{j=1}^m 0\cdot u_j = 0$，方差为

$$\sum_{j=1}^m \left(\frac{1}{n}\cdot u_j^2\right) = \frac{1}{n}.$$

所以 Au 的期望是 0，协方差矩阵是 $n^{-1}I_n$. □

然而，我们关心的其实不单单是 Av_i 的分布，更重要的其实是 $Av_i - Av_j$ 的分布，即压缩后的向量之间的相对距离，幸运的是，我们并不需要做额外的什么计算，我们直接有如下结果：

引理 2.2 向量 $u = \frac{v_i - v_j}{\left\|v_i - v_j\right\|_2}$ 是一个单位向量，因此 $Au \sim \mathcal{N}(0, n^{-1}I_n)$.

J-L 引理实际上在说，$\|Au\|_2$ 偏离 1 的一定程度的概率是非常小的。于是，为了证明 J-L 引理，我们最重要的任务是给出 Au 这样向量模长的集中不等式：

引理 2.3 (单位模引理) *设 $u \sim \mathcal{N}(0, n^{-1}I_n)$，$\epsilon \in (0,1)$ 是给定的常数，那么我们有*

$$\Pr(|\,\|u\|_2^2 - 1| \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2 n}{8}\right).$$

注意到 $\mathbb{E}[\|u\|_2^2] = n \cdot (1/n) = 1$，所以这个引理在说高维空间中，如果正态向量具有单位模长平方期望，那么它的模长就会集中在单位长度附近，因此称为单位模引理.

证明 $|\,\|u\|_2^2 - 1| \geq \epsilon$ 发生有两种可能，$\|u\|_2^2 - 1 \geq \epsilon$ 和 $1 - \|u\|_2^2 \geq \epsilon$. 我们先来计算 $\|u\|_2^2 - 1 \geq \epsilon$ 的概率，根据指数矩，

$$\Pr\left(\|u\|_2^2 - 1 \geq \epsilon\right) \leq \inf_{\lambda>0}\left\{e^{-\lambda(\epsilon+1)}\mathbb{E}\left[e^{\lambda\|u\|_2^2}\right]\right\}.$$

因为 u 的各个分量是相互独立的，所以我们可以把 $\|u\|_2^2$ 展开

$$\mathbb{E}\left[e^{\lambda\|u\|_2^2}\right] = \mathbb{E}\left[e^{\lambda\sum_i u_i^2}\right] = \mathbb{E}\left[\prod_i e^{\lambda u_i^2}\right] = \prod_i \mathbb{E}\left[e^{\lambda u_i^2}\right].$$

可以算得 $\mathbb{E}\left[e^{\lambda u_i^2}\right] = \sqrt{n/(n-2\lambda)}$[lhy: 习题]，所以

$$\Pr\left(\|u\|_2^2 - 1 \geq \epsilon\right) \leq \inf_{\lambda>0}\left\{e^{-\lambda(\epsilon+1)}\left(\frac{n}{n-2\lambda}\right)^{n/2}\right\}.$$

可以验证最小值在 $\lambda = n\epsilon/(2(1+\epsilon))$ 处取到，代入可得

$$\Pr\left(\|u\|_2^2 - 1 \geq \epsilon\right) \leq e^{n(\log(1+\epsilon)-\epsilon)/2} \leq e^{-n\epsilon^2/8}.$$

这里最后一个不等号使用了不等式 $\log(1+\epsilon) \leq \epsilon - \epsilon^2/4$.

计算 $1 - \|u\|_2^2 \geq \epsilon$ 的概率的过程和 $\|u\|_2^2 - 1 \geq \epsilon$ 几乎完全相同的，可以得到

$$\Pr\left(1 - \|u\|_2^2 \geq \epsilon\right) \leq e^{n(\log(1-\epsilon)+\epsilon)/2} \leq e^{-n\epsilon^2/8}.$$

$$\begin{aligned}\Pr\left(|\,\|u\|_2^2 - 1| \geq \epsilon\right) &\leq \Pr\left(\|u\|_2^2 - 1 \geq \epsilon\right) + \Pr\left(1 - \|u\|_2^2 \geq \epsilon\right) \\ &\leq 2e^{-n\epsilon^2/8}.\end{aligned}$$

□

有了单位模引理，我们就可以很容易证明 J-L 引理了。将引理 2.2 中的 u 带入单位模引理，得到

$$\Pr\left(\left|\left\|\frac{A(v_i-v_j)}{\|v_i-v_j\|_2}\right\|_2^2-1\right|\geq\epsilon\right)\leq 2\exp\left(-\frac{\epsilon^2 n}{8}\right).$$

这个结论对任意 $i\neq j$ 成立，因此遍历所有 i,j 对，可得

$$\begin{aligned}\Pr\left(\exists(i,j):\left|\left\|\frac{A(v_i-v_j)}{\|v_i-v_j\|_2}\right\|_2^2-1\right|\geq\epsilon\right)&\leq 2\sum_{i\neq j}\exp\left(-\frac{\epsilon^2 n}{8}\right)\\&=2\binom{N}{2}\exp\left(-\frac{\epsilon^2 n}{8}\right).\end{aligned}$$

换言之，对任意 i,j，$\left|\left\|\frac{A(v_i-v_j)}{\|v_i-v_j\|_2}\right\|_2^2-1\right|<\epsilon$ 都成立的概率不小于

$$1-2\binom{N}{2}\exp\left(-\frac{\epsilon^2 n}{8}\right)=1-N(N-1)\exp\left(-\frac{\epsilon^2 n}{8}\right).$$

代入 $n>\frac{24\log N}{\epsilon^2}$，可得这一概率

$$1-N(N-1)\exp\left(-\frac{\epsilon^2 n}{8}\right)\geq 1-N(N-1)N^{-3}\geq 1-N^{-1}=\frac{N-1}{N}.$$

很多时候，我们关心的并不是向量间的距离，而是向量的内积（比如使用余弦度量的时候），这时候我们可以使用内积版本的 J-L 的引理：

定理 2.6 (J-L 引理，内积形式) 给定 N 个单位向量 $v_1,\dots,v_N\in\mathbb{R}^m$ 和 $n>24\log N/\epsilon^2$，随机矩阵 $A\in\mathbb{R}^{n\times m}$ 每一个元素都独立重复采样自 $\mathcal{N}(0,1/n)$，$\epsilon\in(0,1)$ 是给定常数，那么至少有 $(N-1)/N$ 的概率，使得对所有的 $i\neq j$，都成立

$$|\langle Av_i,Av_j\rangle-\langle v_i,v_j\rangle|<\epsilon.$$

证明 由原始 J-L 引理可知，至少有 $\frac{N-1}{N}$ 的概率满足对于任意 $i\neq j$ 有：

$$\begin{aligned}(1-\epsilon)\|v_i-v_j\|_2^2<\|Av_i-Av_j\|_2^2<(1+\epsilon)\|v_i-v_j\|_2^2,\\(1-\epsilon)\|v_i+v_j\|_2^2<\|Av_i+Av_j\|_2^2<(1+\epsilon)\|v_i+v_j\|_2^2.\end{aligned}$$

我们将第一行乘 -1 加到第二行可以得到

$$4\langle v_i,v_j\rangle-2\epsilon(\|v_i\|_2^2+\|v_j\|_2^2)<4\langle Av_i,Av_j\rangle<4\langle v_i,v_j\rangle+2\epsilon(\|v_i\|_2^2+\|v_j\|_2^2).$$

因为 v_i,v_j 是单位向量，所以上式等价于 $|\langle Av_i,Av_j\rangle-\langle v_i,v_j\rangle|<\epsilon$. □

§2.4 J-L 引理的应用

回顾：J-L 引理描述的是对于 N 个向量，我们可以将它们降到 $\mathcal{O}(\log N)$ 维空间，并将相对距离的误差控制在一定范围内. 它的内容本身就和降维相关，所以最基本的应用就是直接作为降维方法. 许多其它算法例如局部敏感哈希（LSH）、随机 SVD，本质上也都依赖 J-L 引理. 除此之外，J-L 引理对机器学习模型中维度的选择提供了一些理论解释. 下面我们将介绍两个具体的应用案例.

例 2.1 (词向量维度) *在 NLP 的发展中产生了像 Word2Vec、GloVe 这样经典的词向量模型和基于注意力机制的各种大语言模型。这里一个问题自然的问题是，当我们对 N 个单词进行建模，词向量的维度选择多少比较合适？如果维度过高，会使得后续的计算变得更加复杂. 如果维度过低，会无法完全表达出这些单词本身的信息. 对于这一问题，J-L 引理给出了一个比较直接的结论，$\mathcal{O}(\log N)$ 空间足以容纳下 N 个单词. 但是要注意，这一结论成立的前提是正态随机矩阵，然而单词的空间是否符合正态分布是不知道的，所以这一结果只是从理论上给了一个直观，选择什么样的 n 还是由具体的实验效果来决定.*

例 2.2 (多头注意力) *在注意力机制中，我们往往会先把 `head_size` 降低到 64 再做内积. 那么一个很自然的问题是，`head_size` 为 64 的注意力机制是否足以拟合任何概率分布？具体来说，注意力的计算公式为*

$$a_{ij} = \frac{e^{\langle q_i, k_j \rangle}}{\sum_{j=1}^{L} e^{\langle q_i, k_j \rangle}}.$$

其中 $q_i, k_j \in \mathbb{R}^d$。

我们希望能够做到：给定任意的概率矩阵 (p_{ij})，上述 (a_{ij}) 都能够很好的逼近 (p_{ij}). 换言之，给定 (p_{ij}) 和维度 d，我们是否能找到一组 $q_1, \ldots, q_L, k_1, \ldots, k_L \in \mathbb{R}^d$，使得对应项 a_{ij} 与 p_{ij} 足够接近. 其实这就和词向量模型的维度选择问题是等价的. 词向量的维度变成了`head_size`，词表大小变成了序列长度. J-L 引理告诉我们的答案依然是只需要 $\mathcal{O}(\log N)$ 的空间就足以容纳下 N 个单词，一个很粗糙的计算，

[lhy: 这两个例子写得更详细一些。]

§2.5 习题

§2.6 章末注记

第三部分

决策与优化

第四部分

逻辑与博弈

第五部分

认知逻辑

参考文献

[Bre57] Leo Breiman. The Individual Ergodic Theorem of Information Theory. *The Annals of Mathematical Statistics*, 28(3):809–811, 1957.

[CT12] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[Huf52] David A. Huffman. A Method for the Construction of Minimum-Redundancy Codes. *Proceedings of the IRE*, 40(9):1098–1101, September 1952.

[Inf] Information | Etymology, origin and meaning of information by etymonline. https://www.etymonline.com/word/information.

[Jay02] Edwin T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2002.

[KL51] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[LLG+19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, October 2019.

[McM53] Brockway McMillan. The Basic Theorems of Information Theory. *The Annals of Mathematical Statistics*, 24(2):196–219, June 1953.

[RHW86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, pages 318–362. MIT Press, Cambridge, MA, USA, January 1986.

[Rob49] Robert M. Fano. *The Transmission of Information*. March 1949.

[Sha48] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948.

[Shi96] A. N. Shiryaev. *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer, New York, NY, 1996.

[Tin62] Hu Kuo Ting. On the Amount of Information. *Theory of Probability & Its Applications*, 7(4):439–447, January 1962.

[Uff22] Jos Uffink. Boltzmann's Work in Statistical Physics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2022 edition, 2022.

[李 10] 李贤平. 概率论基础. 高等教育出版社, 2010.

索引