信息论的魅力

Magical Girl

August 2-5 2022

1 介绍

信息是人类社会传播的一切内容，人们通过信息得以认识世界并改变世界。当下的时代因为信息的极大丰富而被称作“信息时代”。《信息简史》这一本书具体的描绘了信息的发展故事，从宏观上说明了“信息时代是如何发展而来”，又揭示了“它将走向何处”。信息论则是更为严谨的一门学科，用数学和统计学来解释信息。

信息论和概率论息息相关。信息论的一个核心思想是，传达信息的“价值”取决于其内容发生的概率。如果发生了极有可能发生的事件，则该消息携带的信息非常少。另一方面，如果发生极不可能的事件，则消息的信息量要大得多。

本文将略微探讨信息论的一些来源、性质，并且用之来一些“相关”的现象。

期待这篇文章可以让更多的人喜欢上信息论这门学科。

2 信息熵及其解释

相信“信息熵”这个名词，以及它的形式对所有的计算机学科学生一定并不陌生，其公式也是耳熟能详。

若某一事件 x 有 n 种结果，概率为 $p(a_1), p(a_2)...p(a_n)$，定义 x 的信息熵：

$$\begin{aligned} Entro(x) &= -\sum_{i=1}^{n} p(a_i)\log_2(p(a_i)) \\ p(a_i)log(p(a_i)) &= 0 \; when \; p(a_i) = 0 \end{aligned} \tag{1}$$

可是，为什么信息熵会选择这个公式？我们可以严谨的从数学上证明此公式的一些性质，但在此之前，我们先从两个角度来“解释”这个公式。

2.1 角度 1：信息表示

这个角度最为直观，用一个例子可以描述这个公式的“两端”。

我们首先仅考虑一个是/否的单概率事件 x，即在该事件中我们只有两种选择，其概率分别为 p 和 $(1-p)$。此时由公式知：

$$Entro(x) = -(p\log_2(p) + (1-p)\log_2(1-p)) \tag{2}$$

假如每天晚上我都有 0.5 的概率选择和女朋友通话，且每一次选择是一个独立事件。此时我需要用 1 个 bit 来告诉女朋友我今晚是否通话，因此传递这件事的信息熵 E=1(bit)，代入两个概率（通话和不通话）发现符合公式。

将概率从 0.5 调至 1，即我每天一定会进行通话。于是我便不用发消息，也可以做到传递这个信息，因为这个事件是"必然的"，即传递因此传递这件事的信息熵 E=0(bit)，代入两个概率（通话和不通话）发现也符合公式。

于是从第一个角度，我们理解了单概率公式的两种情况（各半的概率 p=0.5 和确定事件 p=1）时的解释。

2.2 角度 2：哈夫曼编码

仅有角度 1，我们无法表示其余的情况。因此，我们选择从角度 2 展开描述。

假如有一个事件 x，分别有三种结果，50% 的结果是 A，25% 的结果是 B，25% 的结果是 C。现在我们这个事件发生了 n 次，我们需要编码这个结果。我们采用哈夫曼编码来进行编码，哈弗曼编码算法的原理可见WEBSITE。

假设哈夫曼编码的结果如下：0 代表 A，10 代表 B，11 代表 C。根据概率和期望，平均编码的长度为 1*0.5+2*0.25+2*0.25=1.5。同时，同时，我们计算这个事件 x 的熵：$Entro(x) = -(P_a * \log_2(P_a) + P_b * \log_2(P_b) + P_c * \log_2(P_c)) = 1.5$

可以看出，编码的长度和事件的信息熵是一致的，这也符合直觉：编码的长度即"描绘"信息的长度。描绘需要的编码越长，代表事件信息量越大，即"熵"越大。

3 概率学的反直觉：贝叶斯定理

信息论和概率论息息相关，但不必担心，我们需要的知识一些基本的离散概率知识，假设读者已经知道"概率"和"条件概率"这两个概念，那我们就先来介绍另一个重要的内容——贝叶斯定理。

贝叶斯定理的内容非常简单：

如果有两个事件 A 和 B，如果知道 P(A)、P(B)，条件概率 P(B|A)、P(A|B)，可以得到公式：

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)} \tag{3}$$

但这也会造成一个反直觉的现象，即"假阳性"现象。假设有某种病理学测试，可以测出某一个人是否患某种病。

若人群中患病的概率 P=0.005，测试的假阳性和假阴性的概率均为 1%，那：

真阳性且被测出的总概率 P(RelPos)=0.005*0.99=0.00495

假阳性的概率为 P(FalPos)=0.995*0.01=0.00995

测试结果为阳性的总概率 P(Pos)=0.005*0.99+0.995*0.01=0.0149。

因此，如果某人检测呈阳性，那么此人患病的概率为 0.0495/0.0149 = 33.2%。即在三个阳性的人中，只有一个人患病，这和”99%” 的概率相差甚远。基于贝叶斯定理的简单假设，我们却得到了反直觉的结论。

目前而言，似乎贝叶斯定理和“信息熵”之间差的甚远，没有联系。我们目前从概率学角度解释其中的“反常”，但我们在之后还会用信息论的角度解释其中的“奥秘”。

4 分布最低点：Gibbs 不等式

接下来我们介绍 Gibbs 不等式，这是信息熵的一个性质。先介绍另一个概念，两个事件的“交叉熵”，这是衡量两个事件概率分布（信息量）之间的差异：假如 x 事件共有 n 种结果，概率分布为 $p_1, p_2 ... p_n$； y 事件共有也 n 种结果，概率分布为 $q_1, q_2 ... q_n$。如果 x 事件和 y 事件的概率分布越相近，交叉熵 H(x,y) 越接近熵 $Entro(x)$ 和 $Entro(y)$。

x 和 y 的交叉熵为：$H(x,y) = -\sum_{i=1}^{n} p_i \log_2(q_i)$。

Gibbs 不等式的含义为：任意事件 x 的信息熵一定小于等于 x 和任意事件 y 的交叉熵。(证明作为习题)

即：$\sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} q_i = 1$，且 $0 \le p_i, q_i \le 1$，则有：

$$\forall x \forall y \; Entro(x) \le H(x,y) \tag{4}$$

那交叉熵有什么作用呢？我们可以举一个机器学习中的例子。

因为对任何事件 x,y，交叉熵 $H(x,y)$ 都大于 $Entro(x)$，因此 $H(x,y)$ 可以被用作一种拟合函数 R。如果 x 的概率为训练问题真实的概率分布，y 的概率为训练结果的概率分布，则 $H(x,y)$ 和 $Entro(x)$ 越接近，x 和 y 越相近，即在某种意义上，说明机器的训练效果越接近真实。

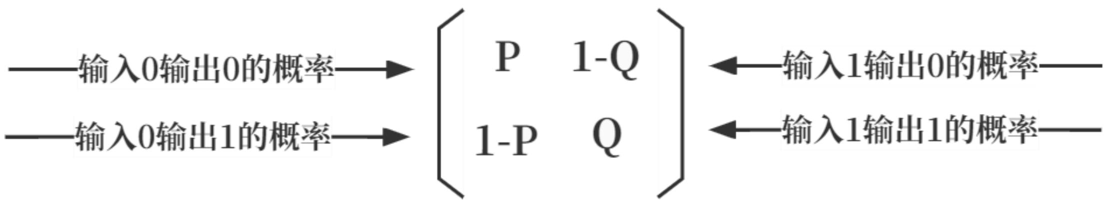
当然这也说明了信息熵的一个特点，即信息熵是交叉熵的意义下能量最低的点。

Gibbs 不等式可以推出很多重要的不等式，如 Kraft 不等式 (Kraft Inequality)，香农信源编码定理 (Shannon’s source coding theorem) 等，这里不多赘述。

5 信息传输：信道模型

随后我们介绍信道模型，这是最简单的信息传输抽象。我们将物理的介质抽象为信道，信道中输入为一串序列（如二进制信道的输入总是一串 01），而输出则是另一个串序列。那么最简单的信息传输模型如下图，输入经过编码成为序列，在信道中传输后被解码成为输出。数字流的最小单位是 1 个数字，如 2 进制信道每次接受 1 个比特，将它输出出去。我们定义传输矩阵这一概念，每个理想信道拥有对应的传输矩阵。传输矩阵的某元素 T_{ab} 是信道**输入 b，输出 a** 的概率。

我们以一个二进制信道的传输矩阵举例，如下图。



当这个矩阵是单位阵时，该信道为“理想信道”，即输入完全传输至输出的信道。

但当 P<1，Q<1 时，该信道就是“有噪音信道”，有概率输出错误的比特。特别的，当 P=Q=0 时，这是一个“反相器”，可以理解为非门。而当 P=Q=0.5 时，这个信道是随机数生成器，它的输出和输入无关，平均的随机输出。

但绝大多数的信道的 P,Q 都接近 1，这样的信道就很大的概率原样输出，满足信道的职责；也有很小的概率输出错误的值，这个值或许和一些干扰干扰有关。

6 创立事件：信道的性质

前排提示：这一节是作者自己的一家之言。

理想信道的一个本质是什么？我们又如何理解传输这一概念？在介绍传输相关的概念之前，我想先解释信道的一大本质。

信道的一大本质是创立一个新的随机事件 $(Input, Output)$**，借由这个事件，传递信息的熵总是不变或增加**。这句话很难理解，同样的，我们用一个例子来解释。

我们假设这是个随机数生成器，即对于任意输入，信道均匀的输出 0 和 1。那么我们假设输入是 X，输出是 Y，那么这个信道的传输是一个概率事件 (X,Y)，它存在四个结果：$P(Y=1|X=0), P(Y=0|X=0), P(Y=1|X=1), P(Y=1|X=1)$。

概率如下：

$$\begin{aligned}
&P(X=0)=P(X=1)=0.5 \\
&P(Y=1|X=0)=P(Y=0|X=0)=0.25 \\
&P(Y=1|X=1)=P(Y=1|X=1)=0.25
\end{aligned} \tag{5}$$

那这个“信道事件”的熵是多少？首先的，我们会想到用传统的信息熵定义这个事件的熵。在这个条件下，信道事件 $Entro(S)=2(bit)$。但，我们需要知道的是，信道最短的编码是 1 个 bit，在 1 个 bit 中，这四个事件不可能同时出现。

当 X=0 时，只存在两个事件 $P(Y=1|X=0), P(Y=0|X=0)$。

当 X=1 时，只存在两个事件 $P(Y=1|X=1), P(Y=0|X=1)$。

我们分开考虑 $X=0$ 和 $X=1$ 两个事件的熵 $Entro(X=0)$ 和 $Entro(X=1)$，当固定 $X=0$ 或 $X=1$ 时，事件的概率为 $P(Y=1)=P(Y=0)=0.5$，此时 $Entro(X=0)=Entro(X=1)=1(bit)$。

因此，我们根据概率可以求出该加权的平均信息熵：

$$\overline{Entro(X)} = p(X=0) * Entro(X=0) + p(X=1) * Entro(X=1) \tag{6}$$

将这个公式扩展，如果信道可以传输的信号值为 0 至 n，输出为 0 至 m，可以得到信道“建立的事件”的平均熵 Entro(X,Y)：

$$\begin{aligned} Entro(X,Y) &= \sum_{i=0}^{n} p(X=i) * Entro(X=i,Y) \\ Entro(X=i,Y) &= -\sum_{j=0}^{m} p(Y=j|X=i) * log_2(p(Y=j|X=i)) \end{aligned} \tag{7}$$

此外，我们需要意识到，Y 是这个信道的输出，因此输出的熵 $Entro(Y)$ 是各个输出的概率 $P(Y=i)$ 组成的。

$$\begin{aligned} Entro(Y) &= -\sum_{i=0}^{m} P(Y=i) * log_2(P(Y=i)) \\ P(Y=i) &= \sum_{j=0}^{n} p(Y=i|X=j) \end{aligned} \tag{8}$$

Entro(X)、Entro(X,Y)、Entro(Y) 之间有密切的关系, 具体的关系我们会在后一节阐述。借助 Gibbs 不等式，我们可以证明（证明留作习题）：

$$Entro(X) \leq Entro(Y) \tag{9}$$

即这就是本节思想的核心：

信道的一大本质是创立一个新的随机事件 $(Intput, Output)$**，借由这个事件，传递信息的熵总是不变或增加**。

7 再谈贝叶斯：噪声、损失、共同信息

接下来我们谈这三个概念：噪声 (Noise)、损失 (Loss)、共同信息 (Mutual Information)。

“信息的损失” 这个词非常常用，但严格意义上说，信息很难能真正被称得上是” 损失了”。事实上，信息和能量类似，不会” 损失”。正如热力学的原理所述的，能量不会随着熵的增加而损失，只是会渐渐” 无法被利用”。

同样的，信息不会随着熵的增加而损失，只是会渐渐的”无法被识别”。

那为什么信息看上去在噪声的影响下就消失了？噪声会使得信息量减少吗？看上去是的，因为，在前面我们提到，当 P=Q=0.5 时就创造了一个完全随机的信道，这样的信道看上去会让信息”消失”。

但事实上，噪音的干扰是让信息”无法被识别”，从而增加信息熵（不确定性）。

我们知道，输入的熵是 Entro(X)，输出事件的熵是 Entro(Y)，且 Entro(X)、Entro(Y)、Entro(X,Y) 是密切相关的，但这仅是公式上的相关性。我们从更直觉的方面开始谈起输入和输出的关系。我们知道所有的事件概率 $P(X=i)$ 条件概率 $P(Y=j|X=i)$，根据贝叶斯公式，我们可以求出 $P(B)$ 和 $P(X=i|Y=j)$。

贝叶斯定理告诉我们，知道更多会“改变”事件发生的概率。同样的，在这里告诉我，我们知道的事件结果可以让我们推断发生事件的概率，让我们摆脱一无所知的状态。概率中有“先验和后验”，那信息熵我们也定义“先验和后验”。

也就是说，已知某个输出 $Y=j$，我们可以推断输入 $X=i$ 的概率，即我们知道逆过来的条件概率 $P(X=i|Y=j)$。也就是说，在这个事件下，我们可以知道一个关于输入的后验分布 P'(X)，即此时的后验信息熵 $Entro(X'|Y=j)$，这个熵较大概率是和先验信息熵 $Entro(X)$ 不同的。

后验信息熵 $Entro(X'|Y=j)$ 的概率如下：

$$Entro(X'|Y=j) = -\sum_{i=0}^{n} P(X=i|Y=j) * log_2(P(X=i|Y=j)) \tag{10}$$

将后验信息熵进行加权求和，我们可以得到后验的平均信息熵 $\overline{Entro(X')}$。

$$\overline{Entro(X')} = \sum_{j=0}^{m} P(Y=j) * (Entro(X'|Y=j) \tag{11}$$

此外，共同信息 (Mutual Information) 的概念就是传递前后的信息熵之差，下图中也展现了 M 的位置：

$$M = Entro(X) - \overline{Entro(X')} \tag{12}$$

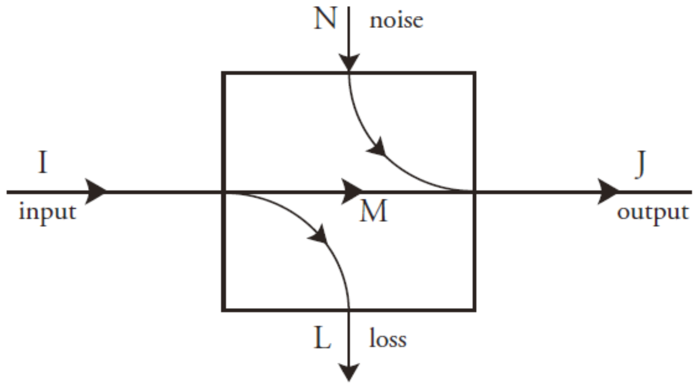
同时，因为任取 X 和 Y 我们知条件概率 $P(X=i|Y=j) \leq P(x=i)$, 因此有：

$$Entro(X'|Y=j) \leq -\sum_{i=0}^{n} P(X=i|Y=j) * log_2(P(X=i)) \tag{13}$$

借助这个不等式，我们可以证明不等式（留作习题）：

$$\overline{Entro(X')} \leq Entro(X)\ ,\ M \geq 0 \tag{14}$$

因此，我们从信息熵的角度重新介绍信道，我们几乎已经介绍了看懂这张图的全部知识。这是信道的另一个角度。



最后，我们重新谈起假阳性的“贝叶斯之怪”。虽然我们没有完全介绍 Noise 和 Loss 的数学表达式，没有完全介绍下面这个表格中 Input、Output、Mutual Information、Noise、Loss 每一个值的计算方法，但我们将阐释结果中的奥秘。

我们是不知道最初的信息量 I 的。借助信道，我们可以得到最终的信息量 J。这张表显示，先验概率的不同让我们最终能知道的不同。若真实事件的信息量 I 比较少（符合疾病的特性：人群中总体得病和健康的概率是悬殊的），经过通道得到的信息量 J 也比较少，即我们通过“阳性”得到的信息仍然是较少的。

	$p(A)$	$p(B)$	I	L	M	N	J
Family history	0.5	0.5	1.00000	0.11119	0.88881	0.11112	0.99993
Unknown Family history	0.9995	0.0005	0.00620	0.00346	0.00274	0.14141	0.14416

信息论的一个核心概念就是：信息是减少不确定性的工具，而自然的一大本质就是让事件重新回归不确定的。

8 奥卡姆剃刀：最少假设和最大熵估计

最大熵估计基于奥卡姆剃刀假设：当我们根据已有信息做出估计时，需要保留最多的未知信息做出推断。这是奥卡姆剃刀原则的引申：若无必要，勿增信息。

最大熵估计的“目的”很简单。给定约束和条件，关于某一个事件的概率无法确定，选择其中信息熵最大的一个。

我们知道，信息熵代表的是“不确定性的度量”，我们将答案（事件概率分布）的信息熵最大化，其实是在预设“最大的不确定性”，以达到最大的客观我们再举一个例子。假设 Lycoris 咖啡厅中有三种糖：胡桃糖/Kurumi-Candy（1 元）、泷波糖/Takina-Candy（2 元）、千束糖/Chisato-Candy（3 元）。

Uni 每天都会随机买一种糖，他每天买每种糖的概率是一个定值。已知 Uni 长期买糖的平均成本为 1.75 元，请求出 Uni 买糖的概率 P(K-C)、P(T-C)、P(C-C)。

我们得知两个约束条件：

Cons.1：$P(K-C)+P(T-C)+P(C-C)=1$（总概率为 1 约束）

Cons.2：$P(K-C)+2P(T-C)+3P(C-C)=1.75$（平均成本约束）

两个方程，三个未知数，我们自然没办法求出精确解。

如果我们知道一个新的信息，比如：Uni 很喜欢千束，会尽可能的买千束糖。那么在满足条件的同时，答案就是尽可能最大化 $P(C-C)$ 的那个解。此时，结果是 P(K-C)=0.625，P(T-C)=0，P(C-C)=0.375。

这样我们可以看出，为了和千束贴贴，Uni 需要完全放弃泷奈，这是**不可能的！明明千束和泷奈都是我的老婆捏！**

但事实上我们不知道这个信息，要得到这个解需要一个新推测（虽然喜欢千束是真的）。这时，我们可以从信息熵这方面考虑。

Uni 买糖是一个概率事件，有事件就有对应的熵。买糖事件的熵是：

$$\begin{aligned} &Entro(X) = -(F(P(K-C)) + F(P(T-C)) + F(P(C-C))) \\ &Define\ \ F(x) = x * \log_2 x \end{aligned} \tag{15}$$

我们的目的就是最大化 $Entro(X)$。有约束的极值问题的解法有很多，比如拉格朗日乘数法。在这里就不过多赘述。在这个例子中，结果为：P(K-C)= 0.466，P(T-C)=0.318，P(C-C)=0.216。

虽然 Uni 真的很喜欢千束，但是因为穷穷，只能在约 20% 的时间中和千束贴贴。事实上，这个概率还不错，因为 Lycoris 每周只更新一集。

最大熵估计的原理虽然不难，但许多细节因为篇幅缘故无法讲清楚，可以查看原论文WEBSITE。

9 尾声：信息论的天地

香农于 1948 年 10 月发表论文《通信的数学理论》，现代信息论研究也因此开始。借助“信息熵”的概念，我们得以描述信息的含量、描述信息传输、描述所进行的估计的“客观性”，描述其他的许多。

我们知道，信息是主观的，我们许多时候并不知道不确定性的分布。科学是不是一种“创立信道”的过程，将原本我们不知道的信息告诉我们呢？如果是这样的话，我们又应该如何看待科学信道中的噪声呢？

还有些问题我也没有思考过。比如，通讯中 0 和 1 出现的比例直接决定了信息传输的效率。在当前的 Unicode 编码中，就整体的通讯而言，0 和 1 的出现比例是否接近呢？如果只考虑 ASCII 码，不同字母出现的频率不同，将其换算为 0 和 1，当前使用连续的编码在传输效率和先验分布上是否优秀？

MIT6.050 就是信息论的一个非常优秀的课程，在最后的最后，我还是想安利一下：**看看 6.050 喵！看看 6.050 谢谢喵！** MIT 6.050J