机器学习理论—信息论：自信息、熵、交叉熵与KL散度

苗思奇
PhD Student@Purdue CS

1. Self-Information

信息论中最根本的一个问题就是，怎么量化一个事件 x 包含的信息量 $I(x)$ 呢？一个量化事件信息量的思路是，量化观测到该事件发生后，给人带来的惊讶程度。具体来说，如果我们有一个信息量的度量 $I(x)$， 它应当是以事件 x 发生的概率 $P(x)$ 参数化的，即 $I(x) = f(P(x))$， 同时我们会希望 $f(P(x))$ 有以下属性：

1. $f(P(x))$ 是关于 $P(x)$ 单调递减的。即，概率越高的事件发生后，带来的信息量/惊讶程度越低；概率越低的事件发生后，带来的信息量/惊讶程度越高。
2. $f(P(x)) \geq 0$。 即，任何事件包含的信息量，应当是非负的。
3. $f(P(x))$ 对任意 $P(x)$ 均是连续的。
4. $I(x_1, x_2) = I(x_1) + I(x_2)$。 即，多个独立事件带来的总信息量，应当等于各个事件的信息量之和。

事实上，满足上述四个条件的函数形式只有 $I(x) = K\log(P(x)), K < 0$。 证明如下：

设事件 C 为两个独立事件 A 和 B 的交集，即 $C = A \cap B$。 根据属性4，我们有，

$$I(C) = I(A \cap B) = I(A) + I(B).$$

因为事件 A 、 B 相互独立，我们有，

$$P(C) = P(A \cap B) = P(A) \cdot P(B).$$

因此，

$$\begin{aligned} f(P(C)) &= f(P(A)) + f(P(B)) \\ &= f(P(A) \cdot P(B)). \end{aligned}$$

可以得出，函数 $f(P(x))$ 应当满足 $f(ab) = f(a) + f(b)$。 现在我们证明满足该等式的函数 f， 必然有 $f(a) = K\log(a)$， 其中 K 是任意实常数。

令 $g(a) = f(2^a)$， 我们有

$$g(a+b) = f(2^{a+b}) = f(2^a 2^b) = f(2^a) + f(2^b) = g(a) + g(b).$$

显然，我们有 $g(a+b) = g(a) + g(b)$。 因为 f 是连续的，那么 g 也是连续的。因此，根据柯西函数方程，必然有 $g(a) = Ka$。

因此， $g(a) = f(2^a) = Ka \Rightarrow f(a) = K\log_2(a)$。

为了使 $f(P(x))$ 满足上述条件1和2， K 应当取负数。另外，其中 $\log$ 的底数可任取，为了方便，我们后续均取 2 为底数。

关于 K 的取值，Shannon在他1948年的论文A Mathematical Theory of Communication中写到：

> The choice of coefficient K is a matter of convenience and amounts to the choice of a unit of measure.

出于此intuition，Shannon令 $K=-1$， 最终得到 $I(x)=-\log(P(x))$， 用来描述事件 x 中包含的信息量。

2. Entropy

现在我们有了对于一个事件 x 的信息量的度量 $I(x)$， 但是往往我们更感兴趣的是这些事件对应的随机变量 X 的信息量。

一个直观的做法就是对随机变量 X 中的所有可能事件的信息量求均值，来代表这个随机变量 X 的信息量。设随机变量 $X\sim p(X)$， 那么 X 的熵被定义为：

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

当 X 为离散随机变量时，

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

显然，在这个定义下， $H(p)$ 自然可以代表随机变量 X 的信息量。

"熵代表随机变量的平均信息量"，这个说法还是过于抽象了，我们能否把这个定义变得更加的数学化？自然是可以的，我们接下来引入熵的一个更加数学化的理解，即，熵代表编码随机变量所需的最短平均编码长度。换句话说，一个随机变量的平均信息量，等价于编码这个随机变量所需的最短平均编码长度。

那么，什么叫做编码一个随机变量呢？编码长度又指什么呢？我们用下面的例子进行一个直观的理解。

$p(x_i)$	Code 1	Code 2
1/2	000	0
1/4	001	10
1/8	010	110
1/16	011	1110
1/64	100	111100
1/64	101	111101
1/64	110	111110
1/64	111	111111
$\mathbb{E}[l_i]$	3	2

假设一离散随机变量 X 的分布如上，其共对应8个事件，每个事件发生的概率不一。现在我们希望的是：对每个事件进行二进制编码，使得传递该随机变量的取值时，所需的平均编码长度最小。

显然，若令每个事件的编码长度为 l_i， 如果我们利用上表 Code 1 编码随机变量 X 的各个取值，那么平均来看我们传递 X 的值时需要 $\mathbb{E}[l_i]=3$ bits 的编码长度。而利用 Code 2 进行编码的话，平均只需要 2 bits 的编码长度。这是因为 Code 2 对概率更大的事件采用了更短的编码，从而降低了编码所需的平均长度，这一方法同霍夫曼编码的思路一致。

现在的问题是，给定随机变量 X， 我们能否预先求得其最短的平均编码长度？答案就是利用随机变量 $X\sim p(X)$ 的熵 $H(p)$。

不难计算， $H(p)=-\sum p(x_i)\log p(x_i)=2$ bits。 也就是说，熵 $H(p)$ 也可以理解为编码随机变量 X 时，所需的最短平均编码长度。

通过以上定义，显然，一个随机变量的信息量与其所需的最短平均编码长度是等价的。这也是很直觉的，如果一个随机变量最优的平均编码长度更大，那么它应当包含更大的信息量；如果一个随机变量所需的最优平均编码长度很小，那么它包含的信息量也应当是很小的。

举个例子，编码掷硬币的结果所需的最优平均编码长度为 $H=-2\cdot\frac{1}{2}\log(\frac{1}{2})=1$ bit。 而编码掷骰

掷骰子需要更大的

平均编码长度，其对应的随机变量也有着更大的信息量。

我们刚刚只是陈述了结论：随机变量的熵即等价于编码该随机变量所需的最短平均编码长度，接下来我们提供证明。

假设编码的字符集的大小为 D， 若采用二进制编码，则 $D=2$。 另外我们假设存在 m 个事件需要编码，每个事件的编码长度为 l_i。 根据编码理论中的Kraft–McMillan Inequality，在给定的码字字长 $l_1,\dots,l_m$ 下能够成功编码，当且仅当

$$\sum_{i=1}^{m} D^{-l_i} \leq 1.$$

至此，寻找最优平均编码长度的问题可以写成如下优化问题：

$$\begin{aligned} \min_{l_i} & \sum_{i=1}^{m} p(x_i) l_i \\ \text{s.t.} & \sum_{i=1}^{m} D^{-l_i} \leq 1. \end{aligned}$$

我们利用Lagrangian multiplier进一步求解带约束的优化问题，即

$$J = \sum_{i=1}^{m} p(x_i) l_i + \lambda \left(\sum_{i=1}^{m} D^{-l_i} - 1 \right).$$

易得 $l_i^* = -\log_D p(x_i)$。 若采用二进制编码，显然，

$$\sum p(x_i) l_i^* = -\sum p(x_i) \log p_i = H(p),$$

其中，熵的单位为 bit， 若采用 e 为底数，则熵的单位为 nat。

因此，随机变量 $X \sim p(X)$ 的熵 $H(p)$ 即是编码随机变量 X 的最优平均编码长度。

3. Cross-Entropy

在说交叉熵之前，我们再回顾一下熵的定义：

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

设 p 为真实分布， q 为 p 的近似分布，交叉熵被定义为：

$$H(p,q) = \mathbb{E}_{X \sim p(X)} \left[-\log q(X) \right].$$

交叉熵和熵的定义长的很像，它们之间的区别可以这样理解：

1. 因为 X 的实际分布为 p， 所以计算期望编码长度时，尽管我们可能并不知道 p， 但理论上总是基于真实分布 $X \sim p(X)$ 计算期望。
2. 当我们利用正确的分布 $p(X)$ 进行编码时， $\log$ 里面的真数是 $p(X)$。 最终算出来的就是随机变量 X 的最优期望编码长度，即熵。
3. 当我们利用错误的分布 $q(X)$ 进行编码时， $\log$ 里面的真数是 $q(X)$。 最终算出来的自然不再是熵，而是我们用错误的分布 $q(X)$ 进行编码后，算出来的随机变量 X 的期望编码长度。

因为熵 $H(p)$ 是随机变量 X 的最优期望编码长度，因此从其定义中我们可以直接得到 $\mathbb{E}_{X \sim p(X)} \left[-\log p(X) \right] \leq \mathbb{E}_{X \sim p(X)} \left[-\log q(X) \right]$。 但我们这里依然证明一下。

对两个离散随机变量的分布 p 、 q， 我们总有

令上述右式减左式，我们只需要证明

$$\sum_{i=1}^{n} p(x_i)\left[\log(p(x_i)) - \log(q(x_i))\right] \geq 0 \tag{1}$$

因为对任意 $x > 0$, $\ln x \leq x - 1$, 所以 $-\log_2 x \geq (1-x)/\ln 2$。不难证明，

$$\begin{aligned}
&\sum_{i=1}^{n} p(x_i)\left[\log(p(x_i)) - \log(q(x_i))\right] \\
=&\sum_{i=1}^{n} p(x_i)\left[\log\left(\frac{p(x_i)}{q(x_i)}\right)\right] \\
\geq&\frac{1}{\ln 2}\sum_{i=1}^{n} p(x_i)\left(1 - \frac{q(x_i)}{p(x_i)}\right) \\
=&\frac{1}{\ln 2}\left(\sum_{i=1}^{n} p(x_i) - \sum_{i=1}^{n} q(x_i)\right) \\
=&0
\end{aligned}$$

显然，我们确实有 $\mathbb{E}_{X\sim p(X)}\left[-\log p(X)\right] \leq \mathbb{E}_{X\sim p(X)}\left[-\log q(X)\right]$。

在了解了交叉熵和熵的关系后，我们就可以从信息论的角度理解，为什么交叉熵可以在机器学习中作为损失函数。我们在最小化交叉熵的时候，事实上是在逼近最优期望编码长度，即利用 $q(x)$ 逼近 $p(x)$， 使得交叉熵尽可能的小，以接近熵的值。

4. KL Divergence

对于离散随机变量，分布 p 和 q 的KL散度的定义如下:

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

对KL散度在信息论中的一个直观的理解是将其写开，即

$$\begin{aligned}
D_{KL}(p\|q) &= -\sum_{i=1}^{n} p(x_i)\cdot\log\frac{q(x_i)}{p(x_i)} \\
&= -\sum_{i=1}^{n} p(x_i)\cdot\log q(x_i) + \sum_{i=1}^{n} p(x_i)\cdot\log p(x_i) \\
&= H(p,q) - H(p).
\end{aligned}$$

通过上节我们知道，交叉熵 $H(p,q)$ 指利用分布 q 编码随机变量 X 所需的期望编码长度，而熵 $H(p)$ 指编码随机变量 X 所需的最优期望编码长度。

既然 $D_{KL}(p\|q) = H(p,q) - H(p)$， 那么显然，其意味着利用 q 编码 X 所带来的额外编码长度。事实上，上一节中的式(1)左侧等价于KL散度，因此KL散度恒大于等于零。

倘若我们优化KL散度，即是希望减小所需的额外编码数，使得分布 p 和 q 变得接近。这里有两种情况:

1. 若真实分布 p 恒定，那么优化KL散度等价于优化交叉熵，其目的是令交叉熵逼近最优期望编码长度，使得 q 尽可能接近 p。 在训练辨别模型时，往往是这种情况。为了简化计算，人们往往直接对交叉熵进行优化。
2. 若真实分布 p 不恒定，那么优化KL散度会同时改变交叉熵和熵的值，使得 p 与 q **相互**接近。在训练生成模型时，往往是这种情况，为了使分布 p 与 q 相互接近，我们必须直接对KL散度进行优化。

5. References

[1] Link1

[3] Link3

[4] Link4

[5] Link5

[6] Link6

[7] Link7

编辑于 01-20

机器学习 深度学习（Deep Learning） 信息论