计算语言学的数学基础

(3) 概率论与信息论

王予沛

2025 年 3 月 10 日

北京师范大学数字人文系

1. 概率论基础

2. 信息论基础

概率论基础

概率论基础

- 概率是从事件空间到区间 $[0,1]$ 的映射。
- 概率的两个视角
  - 频率视角：重复实验，记录结果。例如扔骰子。
  - 贝叶斯视角：根据已知信息，更新对事件发生可能的信心，如天气预报。这时概率可理解为对某件事情的信念度（degree of belief）。
- 概率分布
  - 离散型随机变量 $\rightarrow$ 概率质量函数
  - 连续型随机变量 $\rightarrow$ 概率密度函数
- 归一性：概率质量函数或概率密度函数在所有可能取值上的求和或积分为 1。
- 条件概率：在已知某些事件发生的条件下，事件发生的概率。
- 独立性：两个事件的发生互不影响。
- 期望：随机变量的加权平均值：$E(X)=\sum_i x_i p_i$。
- 方差：随机变量与其期望的差的平方的期望：$Var(X)=E((X-E(X))^2)$。

信息论基础

惊讶度

你扔一次硬币，你的朋友猜测硬币落地后花面朝上。如果你的朋友猜对了，你或许不会太惊讶，因为硬币落地后花面朝上的概率是 0.5。

但如果你投掷一枚骰子，你的朋友猜测骰子落地后点数是 6。如果你的朋友猜对了，你可能会更加惊讶，因为骰子落地后点数是 6 的概率是 1/6。

我们发现，概率越小的事件，发生时带来的惊讶度越大。我们是否可以找到一个衡量惊讶度的函数，来描述事件发生的惊讶程度?

惊讶度

我们想定义的描述惊讶度的函数应该满足以下条件:

- 事件发生的概率越小，惊讶度越大；事件发生的概率越大，惊讶度越小。
- 当一件事情的概率比另一件事情的概率小 n 倍时，它的惊讶度应该比另一件事情大 n 倍。例如你的朋友连续三次猜对了骰子的点数 $\{1,3,6\}$，你应该比猜对了单次惊讶 3 倍。

严格来说，惊讶度应该是一个非负，和事件发生的概率负相关，并且在概率需做乘法的时候，惊讶度做加法。

我们考虑一个事件 x，它的概率为 $p(x)$。我们定义它的惊讶度为:

$$h(x) = \log \frac{1}{p(x)}$$

将其平均，我们将得到对一个概率分布的总体惊讶度:

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

我们将这个值称为概率分布的**熵**（entropy）。

从贝叶斯视角看惊讶度和熵

假设有一枚硬币非常地不均匀，它有 99% 的概率正面朝上，1% 的概率反面朝上。但我们事先不知道这一点，我们先验地认为硬币是均匀的，即正面朝上的概率是 50%。

我们扔 10 次硬币，结果是 10 次正面朝上，这大大出乎我们的意料。

如果这枚硬币是均匀的，我们扔 10 次硬币，结果是 10 次正面朝上的概率是 $0.5^{10} \approx 0.001$，这件事情的惊讶度是 $\log \frac{1}{0.001} \approx 7$；而实际上由于硬币非常不均匀，这个概率是 $0.99^{10} \approx 0.9$，这件事情的惊讶度是 $\log \frac{1}{0.9} \approx 0.1$。

我们如此惊讶，并不是因为发生了什么奇怪的事情，而是因为我们事先的信念是错误的。

交叉熵

现在，我们定义一个指标，当随机变量的真实分布为 P，但我们对该变量的信念是分布 Q 时我们的惊讶程度。

$$H(P,Q)=\sum_{x\in\mathcal{X}}p(x)\log\frac{1}{q(x)} \tag{1}$$

我们将这个值称为**交叉熵**（cross entropy）。交叉熵可用于衡量两个概率分布之间差异的指标。注意到，交叉熵不能关于 P 和 Q 对称。

显然，只要我们用错误的分布来描述真实分布，我们在抽样观测时就一定会对结果感到惊讶。因此交叉熵就一定大于熵。

$$H(P,Q)\geq H(P)$$

等号成立当且仅当 $P=Q$。

例子：均匀和非均匀硬币

当我们认为硬币均匀$P = \{0.5, 0.5\}$但实际硬币不均匀$Q = \{0.99, 0.01\}$时，投掷一次硬币的交叉熵为：

$$H(P, Q) = 0.99 \log\left(\frac{1}{0.5}\right) + 0.01 \log\left(\frac{1}{0.5}\right) \approx 0.7$$

当我们认为硬币不均匀$P = \{0.99, 0.01\}$但实际硬币均匀$Q = \{0.5, 0.5\}$时，投掷一次硬币的交叉熵为：

$$H(P, Q) = 0.5 \log\left(\frac{1}{0.99}\right) + 0.5 \log\left(\frac{1}{0.01}\right) \approx 2.3$$

从本例可见，我们对某一个不符合真实分布的信念越自信，实验结果越出乎意料，我们就会越惊讶，或者说，错误的惩罚越重。

例子：垃圾邮件分类

假设我们想开发一个简单的邮件分类系统，用于判断收到的邮件是" 正常邮件" 还是" 垃圾邮件"。这是一个典型的二分类问题。我们用 0 表示正常邮件，用 1 表示垃圾邮件。

假设我们有一个已训练好的模型，对 5 封测试邮件进行了预测，得到了以下结果：

邮件编号	真实标签	模型预测
邮件 1	0(正常)	0.2 (20% 可能是垃圾邮件)
邮件 2	0(正常)	0.1 (10% 可能是垃圾邮件)
邮件 3	1(垃圾)	0.7 (70% 可能是垃圾邮件)
邮件 4	1(垃圾)	0.9 (90% 可能是垃圾邮件)
邮件 5	0(正常)	0.4 (40% 可能是垃圾邮件)

例子：垃圾邮件分类

在实际应用时，我们有时会对所有样本的交叉熵取平均，得到一个平均交叉熵。在本例中，我们这样计算[1]：

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]$$

其中，N 是样本数量，y_i 是真实标签（代表真实分布），p_i 是模型预测的概率（代表模型对垃圾邮件分类的“信念”分布）。

- 邮件 1（真实标签 $=0$，预测概率 $=0.2$）：
  模型预测正常邮件的概率为 0.8，代入公式：

$$-[0 \cdot \log(0.2) + (1 - 0) \cdot \log(0.8)] = -\log(0.8) = 0.223$$

- 邮件 2（真实标签 $=0$，预测概率 $=0.1$）：

$$-[0 \cdot \log(0.1) + (1 - 0) \cdot \log(0.9)] = -\log(0.9) = 0.105$$

[1]前面出现负号是因为 $\log(\frac{1}{p}) = -\log(p)$. 由于是二分类，所以概率只有 p 和 $1 - p$ 两种情况.

例子：垃圾邮件分类

- 邮件 3（真实标签 $=1$，预测概率 $=0.7$）：

$$-[1 \cdot \log(0.7) + (1-1) \cdot \log(0.3)] = -\log(0.7) = 0.357$$

- 邮件 4（真实标签 $=1$，预测概率 $=0.9$）：

$$-[1 \cdot \log(0.9) + (1-1) \cdot \log(0.1)] = -\log(0.9) = 0.105$$

- 邮件 5（真实标签 $=0$，预测概率 $=0.4$）：

$$-[0 \cdot \log(0.4) + (1-0) \cdot \log(0.6)] = -\log(0.6) = 0.511$$

最后，计算平均损失：

$$\text{平均损失} = \frac{0.223 + 0.105 + 0.357 + 0.105 + 0.511}{5} = 0.260$$

例子：垃圾邮件分类

交叉熵损失值越小，表示模型预测效果越好：

- 当模型对正常邮件预测的"正常概率"接近 1 时，损失值接近 0
- 当模型对垃圾邮件预测的"垃圾概率"接近 1 时，损失值接近 0
- 当模型预测错误（信念和真实分布不一致）时，损失值会增大

在我们的例子中：

- 邮件 2 和邮件 4 的损失值最小 (0.105)，说明模型对这两封邮件的预测最准确
- 邮件 5 的损失值最大 (0.511)，说明模型对该邮件的预测信心不足

总的平均损失为 0.260，这个数值可以与其他模型的损失进行比较，帮助我们选择更好的模型。

KL 散度

我们发现，交叉熵带来的惊讶，有两个成分：

- 真实分布 P 本身的不确定性；
- 我们用 Q 描述 P 时带来的惊讶。

我们希望将二者剥离开，特别是希望仅仅考虑我们用 Q 描述 P 时带来的惊讶，由于真实分布 P 本身的不确定性已经可以用熵来描述，我们只需要考虑交叉熵与熵的差值。

$$D_{KL}(P \parallel Q) = H(P, Q) - H(P)$$

我们将这个值称为**KL 散度**（KL divergence）。同样地，KL 散度不能关于 P 和 Q 对称。