# KL 散度学习笔记

郭彦杰15220172202360 , 林志晟15220172202486

October 18, 2019

## Contents



[**摘要**]K-L 散度是信息理论（Information theory）中重要的一个量。信息论是运用概率论与数理统计的方法研究信息、信息熵等问题的应用数学学科。其基本概念包括信息熵，交叉熵，K-L散度，互信息（Mutual information）等等。本文将介绍自信息、熵、KL散度等概念，对KL散度概念做出理解。

[**关键词**]自信息，熵，KL散度

# 1 自信息

由Claude Elwood Shannon提出的自信息（self-information）被用于指代一事件发生时的信息本身。举个例子[1],当一枚色子被随机投出时，如果你告诉我："点数不大于6。"那么这句话的信息量为0，因为这相当于是在重复一个简单的事实（发生概率为1），对我猜出色子的点数没有任何帮助。如果你告诉我："点数小于6"（发生概率为$\frac{5}{6}$），那么这句话虽然信息量大于零，但是对我的帮助不大，因为我仍然要从剩下的5个数字中选一个。假如你告诉我"点数

[1]https://blog.csdn.net/xuejianbest/article/details/80391191

为1”（发生概率为$\frac{1}{6}$），那么这句话给我的信息量就较大。因此从这个角度我们可以说，一个事件发生的概率越小，其包含的信息量越大。

我们通常会看到对于“信息量”$I$的定义是

$$I = log_b(\frac{1}{p_A})$$

其中b取值决定于信息量的单位，例如二进制数当中b取值为2.而$p_A$指的是事件A发生的概率。那么为什么要选择对数函数呢？因为信息量还需要具有“可加性”。仍然以扔色子为例。一共扔两次色子，假设有以下两种情况：

- 情况1：第一次将色子投出后，你告诉我点数为2，第二次将色子投出后，你告诉我点数为4。
- 情况2：第一次将色子投出后，你没有告诉我任何信息。第二次将色子投出后，你告诉我：第一次点数为2，第二次点数为4.

我们可以认为，这两种情况在最后给予我们的信息量是一样的（因为我们确实知道了“投两次色子，第一次点数为2，第二次点数为4”这个信息。）通过以上提到的对于信息量的定义，情况1给我们的信息量为：

$$log_b(\frac{1}{\frac{1}{6}}) + log_b(\frac{1}{\frac{1}{6}}) = log_b(6) + log_b(6) = log_b 36$$

而情况2给我们的信息量为：

$$log_b \frac{1}{\frac{1}{6} \times \frac{1}{6}} = log_b 36$$

从这个例子可以看出，对数函数本身的性质很好的满足了我们对于“信息量”函数的需求：

- 发生事情的概率越小，其包涵的信息量越大。
- 具有“可加性”。

# 2 信息熵

## 2.1 定义

1948年，香农（Shannon）提出了“信息熵”的概念，解决了对信息的量化度量问题。信息熵这个词是从热力学中借用过来的。热力学中的热熵是表示分子状态混乱程度的物理量,而香农用信息熵的概念来描述信源的不确定度。

信息熵可以被理解为信息量的期望值[2]。若有$n$种可能结果的事件$X$，其第$i$种结果发生的可能性（概率）是$p(x_i)$，我们则称这个事件的“信息熵”为

$$H = E_p(-logp_i) = -\sum_i p(x_i)logp(x_i)$$

当然，在连续的情况下，我们有

$$H = E_p(-logp_i) = -\int_x p(x)logp(x)dx$$

### 2.2 信息熵在生物学中的应用

Shannon Wiener index 是生物多样性指数之一，是测量生物群落异质性的重要指标，它其实就是一个信息熵。计算公式为：

$$H = -\sum_i^n p_i logp_i$$

其中，$p_i$表示某物种个体数占总个体数的比例。

## 3 K-L散度

### 3.1 定义

我们在衡量线性模型时，通常会对比预测值与观察到的真实值之间的距离，从而衡量我们模型拟合的程度。但是如何去衡量两个概率分布之间的“距离”？K-L散度能够很好地做到这一点。设P，Q是同一概率空间中的两种分布，则PQ间的K-L散度定义为：

$$D_{\mathrm{KL}}(P\|Q) = \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right)$$

对于连续型随机变量，K-L散度定义为：

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x)\log\left(\frac{p(x)}{q(x)}\right)dx$$

虽然K-L散度想要衡量的是两个概率分布间的“距离”，但是和欧式空间中的“距离”又有所不同——我们平时说的距离，A点到B点的距离是等于B点到A点的距离的。这就是欧式空间中的“距离”的对称性。

$$D(A,B) = D(B,A)$$

不过在K-L散度中，$D_{\mathrm{KL}}(P\|Q)$却不一定等于$D_{\mathrm{KL}}(Q\|P)$

[2]https://zhuanlan.zhihu.com/p/74075915

因为K-L散度是我们想要衡量的“距离”，因此，K-L具有非负性。我们将在下文证明其非负性。事实上，测量两个概率分布之间距离所用的“散度”有很多，如下表：

| Divergence | $I_\phi(p,q)$ |
|---|---|
| Kullback-Leibler | $\Sigma p_\omega \log\left(\frac{p_\omega}{q_\omega}\right)$ |
| Burg entropy | $\Sigma q_\omega \log\left(\frac{q_\omega}{p_\omega}\right)$ |
| J-divergence | $\Sigma\left(p_\omega - q_\omega\right)\log\left(\frac{p_\omega}{q_\omega}\right)$ |
| $\chi^2$distance | $\Sigma\frac{(p_\omega-q_\omega)^2}{p_\omega}$ |
| Modified $\chi^2$distance | $\sum\frac{(p_\omega-q_\omega)^2}{q_\omega}$ |
| Variation distance | $\Sigma\lvert p_\omega - q_d\rvert$ |
| Hellinger distance | $\Sigma(\sqrt{p_\omega}-\sqrt{q_d})^2$ |

## 3.2 散度与交叉熵

在离散情况下，我们对K-L散度定义的公式进行变形，得到：

$$\begin{aligned} D_{\mathrm{KL}}(P\|Q) &= -\sum_{x\in\mathcal{X}} p(x)\log q(x) + \sum_{x\in\mathcal{X}} p(x)\log p(x) \\ &= \mathrm{H}(p,q) - \mathrm{H}(p) \end{aligned} \tag{1}$$

其中，H(P,Q)称为交叉熵，计算式如下：

$$H(p,q) = \mathrm{E}_p(-\log q) \tag{2}$$

类似的，对于连续的情况，我们有：

$$\begin{aligned} D_{\mathrm{KL}}(P\|Q) &= -\int_{x\in\mathcal{X}} p(x)\log q(x) + \int_{x\in\mathcal{X}} p(x)\log p(x) \\ &= \mathrm{H}(p,q) - \mathrm{H}(p) \end{aligned} \tag{3}$$

注意$H(p,q)$与$H(q,p)$是不一样的，两者包含不同的信息。

## 3.3 K-L散度与模型选择

在我们学习的课程中，K-L散度用来度量两个分布之间的差异性。可以通过Jensen不等式证明，K-L散度是非负的。证明[3] 如下：

设$f(x)$为非负函数，且有

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

[3]https://zhuanlan.zhihu.com/p/39682125

若$g$是任意实可测函数且函数$\varphi$图像是凸的（即其本身是凹函数），那么有Jensen不等式如下：

$$\varphi\left(\int_{-\infty}^{\infty} g(x) f(x) dx\right) \leq \int_{-\infty}^{\infty} \varphi(g(x)) f(x) dx$$

注意到，$-lnx$是严格的凹函数（它的图像是凸的，即它本身是凹函数）且$\int q(x)dx = 1$，将$\varphi(x) = -\ln(x), g(x) = \frac{q(x)}{p(x)}, f(x) = p(x)$代入上述Jensen不等式当中，则有

$$KL(p\|q) = \int p(x)\left\{-\ln\left[\frac{q(x)}{p(x)}\right]\right\} dx \geq -\ln\left[\int q(x) dx\right] = 0$$

当且仅当$p(x) = q(x)$时，K-L散度为0。

在机器学习理论中，我们假设P(x)是x的实际分布，Q(x)是根据观测到的数据对x分布进行的预测，当P(x)和Q(x)越接近时，K-L散度越小，也就是说，预测越准确。当P(x)和Q(x)差距越大时，K-L散度越大，预测越不准确。因此，我们可以将K-L散度看成一个损失函数，定义这种损失函数下的样本误差（In sample error）和实际误差（Out of sample error）[4]，这时，

$$E_{out}(q) = \int \log\left(\frac{p(x)}{q(x)}\right) p(x) dx$$

$$E_{in}(q) = \frac{1}{N} \sum_{i=1}^{N}\left(\log p\left(x_{i}\right) - \log q\left(x_{i}\right)\right)$$

我们要使样本误差最小化。由于$logp(xi)$是定值，要使$E_in$最小化，就是要使

$$E_{in}(q) = -\frac{1}{N} \sum_{i=1}^{N} \log q\left(x_{i}\right)$$

最小化.假如我们用最大似然估计法（MLE）来估计q，我们得到的对数似然函数是：

$$\log \mathcal{L}(q) = \sum_{i=1}^{N} \log q\left(x_{i}\right)$$

于是我们发现，选择模型时，使用K-L散度估计和使用MLE估计在本质上是相同的。

## 3.4 K-L散度在生物学中的应用

在Donaldson-Matasci, Bergstrom, Lachmann（2019）的论文中，他们建立了不确定性与最优增长的模型。假设有$n$个环境，每个环境有一个最佳状态，环境e中最佳状态下的效益是$d_e$，其他状态下的效益为0，环境$e$出现的概率

[4] Prof. Mao's slides

为$p(e)$，出现最佳状态的概率为$x(e)$，那么在一系列观测（N次）中，最佳状态出现的次数为Np(e)，在这段时间内，种群的增长率为：

$$\prod_e (d_e x(e))^{Np(e)}$$

可以使它的对数最大化，除以常数N,得到：

$$g(x) = \sum_e p(e) log(d_e x(e)) = \sum_e p(e) log d_e + \sum_e p(e) log x(e)$$

可以将它化为：

$$\begin{aligned}
\mathrm{g(x)} = & \sum_{\mathrm{c}} \mathrm{p(e)} \log \mathrm{d_c} \\
& + \left( \sum_{\mathrm{c}} \mathrm{p(e)} \log \mathrm{p(e)} - \sum_{\mathrm{c}} \mathrm{p(e)} \log \mathrm{p(e)} \right) \\
& + \sum_{e}^{e} \mathrm{p(e)} \log \mathrm{d_e} + \sum_{\mathrm{e}} \mathrm{p(e)} \log \mathrm{p(e)} \\
& - \sum_{e} \mathrm{p(e)} \log \frac{\mathrm{p(e)}}{\mathrm{x(e)}} \\
= & \sum_{\mathrm{P}}^{e} \mathrm{(e)} \log \mathrm{d_e} - \mathrm{H(E)} - \mathrm{D_{KL}(p \| x)}
\end{aligned}$$

同时，作者还提到，假设一个生态系统中物种$s$的实际频率为$p(s)$，观测到的频率$q(s) = n(s)/N$，可以指定一个0-1函数，抽取一个个体，如果是物种$s$，则为1，否则为0。那么当N足够大时，在实际频率为$p$的条件下，观测到频率为$q$的概率大约是：$2^{-D(p||q)N}$。

K-L散度现在已经被广泛应用在基础科学、工程科学等领域。在如今海量数据涌现的前提之下，我们相信K-L散度将会成为经济研究中不可或缺的好工具。

# 参考文献


[1] Matina C. Donaldson-Matasci, Carl T. Bergstrom and Michael Lachmann. The fitness value of information[J]. Oikos 119 (2010) 219 – 230

[2] Shannon, C. E. Shannon C E . A mathematical theory of communication[J]. Bell Labs Technical Journal, 1948, 27(4):379-423.

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

[4]王克峰. 基于子空间辨识与K-L散度分析的风电机组故障诊断研究[D].沈阳工业大学,2019.