維基百科

相对熵

维基百科，自由的百科全书

KL散度（**Kullback-Leibler divergence**，简称**KLD**）[1]，在讯息系统中称为相对熵（relative entropy），在连续时间序列中称为随机性（randomness），在统计模型推断中称为讯息增益（information gain）。也称讯息散度（information divergence）。

KL散度是两个概率分布P和Q差别的非对称性的度量。 KL散度是用来度量使用基于Q的分布来编码服从P的分布的样本所需的额外的平均比特数。典型情况下，P表示数据的真实分布，Q表示数据的理论分布、估计的模型分布、或P的近似分布。[2]

目录

定义

对于离散随机变量，其概率分布P 和 Q的KL散度可按下式定义为

$$D_{\mathrm{KL}}(P\|Q) = -\sum_i P(i) \ln \frac{Q(i)}{P(i)}.$$

等价于

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}.$$

即按概率P求得的P和Q的对数商的平均值。KL散度仅当概率P和Q各自总和均为1，且对于任何i皆满足$Q(i)>0$及$P(i)>0$时，才有定义。式中出现$0 \ln 0$的情况，其值按0处理。

对于连续随机变量，其概率分布P和Q可按积分方式定义为 [3]

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)}\, \mathrm{d}x,$$

其中p和q分别表示分布P和Q的密度。

更一般的，若P和Q为集合X的概率测度，且P关于Q绝对连续，则从P到Q的KL散度定义为

$$D_{\mathrm{KL}}(P\|Q) = \int_X \ln \frac{\mathrm{d}P}{\mathrm{d}Q}\, \mathrm{d}P,$$

其中，假定右侧的表达形式存在，则$\frac{\mathrm{d}Q}{\mathrm{d}P}$为$Q$关于$P$的R–N导数。

相应的，若P关于Q绝对连续，则

$$D_{\mathrm{KL}}(P\|Q)=\int_X \ln\frac{\mathrm{d}P}{\mathrm{d}Q}\,\mathrm{d}P=\int_X \frac{\mathrm{d}P}{\mathrm{d}Q}\ln\frac{\mathrm{d}P}{\mathrm{d}Q}\,\mathrm{d}Q,$$

即为P关于Q的相对熵。

特性

相对熵的值为非负数：

$$D_{\mathrm{KL}}(P\|Q)\geq 0,$$

由吉布斯不等式可知，当且仅当$P=Q$时$D_{KL}(P\|Q)$为零。

尽管从直觉上KL散度是个度量或距离函数, 但是它实际上并不是一个真正的度量或距离。因为KL散度不具有对称性：从分布P到Q的距离通常并不等于从Q到P的距离。

$$D_{\mathrm{KL}}(P\|Q)\neq D_{\mathrm{KL}}(Q\|P)$$

KL散度和其它量的关系

自信息和KL散度

$$I(m)=D_{\mathrm{KL}}(\delta_{im}\|\{p_i\}),$$

互信息和KL散度

$$\begin{aligned}I(X;Y)&=D_{\mathrm{KL}}(P(X,Y)\|P(X)P(Y))\\&=\mathbb{E}_X\{D_{\mathrm{KL}}(P(Y|X)\|P(Y))\}\\&=\mathbb{E}_Y\{D_{\mathrm{KL}}(P(X|Y)\|P(X))\}\end{aligned}$$

信息熵和KL散度

$$\begin{aligned}H(X)&=\mathrm{(i)}\,\mathbb{E}_x\{I(x)\}\\&=\mathrm{(ii)}\log N-D_{\mathrm{KL}}(P(X)\|P_U(X))\end{aligned}$$

条件熵和KL散度

$$\begin{aligned}H(X|Y)&=\log N-D_{\mathrm{KL}}(P(X,Y)\|P_U(X)P(Y))\\&=\mathrm{(i)}\ \log N-D_{\mathrm{KL}}(P(X,Y)\|P(X)P(Y))-D_{\mathrm{KL}}(P(X)\|P_U(X))\\&=H(X)-I(X;Y)\\&=\mathrm{(ii)}\ \log N-\mathbb{E}_Y\{D_{\mathrm{KL}}(P(X|Y)\|P_U(X))\}\end{aligned}$$

交叉熵和KL散度

$$\mathrm{H}(p,q)=\mathrm{E}_p[-\log q]=\mathrm{H}(p)+D_{\mathrm{KL}}(p\|q).$$

参考文献

1. Kullback, S.; Leibler, R.A. On Information and Sufficiency. Annals of Mathematical Statistics. 1951, **22** (1): 79–86. MR 39968. doi:10.1214/aoms/1177729694.
2. Kullback, S.; Leibler, R.A. On information and sufficiency. Annals of Mathematical Statistics. 1951, **22** (1): 79–86. MR 39968. doi:10.1214/aoms/1177729694.
3. C. Bishop (2006). Pattern Recognition and Machine Learning. p. 55.