火星十一郎

海纳百川，有容乃大；壁立千仞，无欲则刚...

为什么我的眼中常含泪水，因为我有一个算法不会...

浅谈KL散度

一、第一种理解

相对熵（relative entropy）又称为KL散度（Kullback-Leibler divergence，简称KLD），信息散度（information divergence），信息增益（information gain）。

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的编码来编码来自P的样本平均所需的额外的比特个数。 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，模型分布，或P的近似分布。

根据shannon的信息论，给定一个字符集的概率分布，我们可以设计一种编码，使得表示该字符集组成的字符串平均需要的比特数最少。假设这个字符集是X，对$x \in X$，其出现概率为$P(x)$，那么其最优编码平均需要的比特数等于这个字符集的熵：

$$H(X)=\sum_{x \in X}P(x)\log[1/P(x)]$$

在同样的字符集上，假设存在另一个概率分布$Q(X)$。如果用概率分布$P(X)$的最优编码（即字符x的编码长度等于$\log[1/P(x)]$），来为符合分布$Q(X)$的字符编码，那么表示这些字符就会比理想情况多用一些比特数。KL-divergence就是用来衡量这种情况下平均每个字符多用的比特数，因此可以用来衡量两个分布的距离。即：

$$D_{KL}(Q||P)=\sum_{x \in X}Q(x)[\log(1/P(x))] - \sum_{x \in X}Q(x)[\log[1/Q(x)]]=\sum_{x \in X}Q(x)\log[Q(x)/P(x)]$$

由于$-\log(u)$是凸函数，因此有下面的不等式

$$D_{KL}(Q||P) = -\sum_{x \in X}Q(x)\log[P(x)/Q(x)] = E[-\log P(x)/Q(x)] \geqslant -\log E[P(x)/Q(x)] = -\log\sum_{x \in X}Q(x)P(x)/Q(x) = 0$$

昵称：火星十一郎
园龄：6年4个月
粉丝：640
关注：16
+加关注

即KL-divergence始终是大于等于0的。当且仅当两分布相同时，KL-divergence等于0。

==========

举一个实际的例子吧：比如有四个类别，一个方法A得到四个类别的概率分别是0.1,0.2,0.3,0.4。另一种方法B（或者说是事实情况）是得到四个类别的概率分别是0.4,0.3,0.2,0.1,那么这两个分布的KL-Distance(A,B)=0.1*log(0.1/0.4)+0.2*log(0.2/0.3)+0.3*log(0.3/0.2)+0.4*log(0.4/0.1)

这个里面有正的，有负的，可以证明KL-Distance()>=0.

从上面可以看出， KL散度是不对称的。即KL-Distance(A,B)!=KL-Distance(B,A)

KL散度是不对称的，当然，如果希望把它变对称，

Ds(p1, p2) = [D(p1, p2) + D(p2, p1)] / 2.

二、第二种理解

今天开始来讲相对熵，我们知道信息熵反应了一个系统的有序化程度，一个系统越是有序，那么它的信息熵就越低，反之就越高。下面是熵的定义

如果一个随机变量X的可能取值为$X = \{x_1, x_2, ..., x_n\}$，对应的概率为$p(X = x_i) \quad (i = 1, 2, ..., n)$，则随机变量$X$的熵定义为

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

有了信息熵的定义，接下来开始学习相对熵。

1. 相对熵的认识

相对熵又称互熵，交叉熵，鉴别信息，Kullback熵，Kullback-Leible散度（即KL散度）等。设$p(x)$和$q(x)$是X取值的两个概率概率分布，则p对q的相对熵为

$$D(p||q) = \sum_{i=1}^{n} p(x) \log \frac{p(x)}{q(x)}$$

在一定程度上，熵可以度量两个随机变量的距离。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

用来度量使用基于Q的编码来编码来自P的样本平均所需的额外的位元数。 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，模型分布，或P的近似分布。

2. 相对熵的性质

相对熵（KL散度）有两个主要的性质。如下

（1）尽管KL散度从直观上是个度量或距离函数，但它并不是一个真正的度量或者距离，因为它不具有对称性，即

$$D(p||q) \neq D(q||p)$$

（2）相对熵的值为非负值，即

$$D(p||q) \geq 0$$

在证明之前，需要认识一个重要的不等式，叫做吉布斯不等式。内容如下

若 $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = 1$，且 $p_i, q_i \in (0,1]$，则有：

$$-\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log q_i$$，等号成立当且仅当 $p_i = q_i \forall i$

在信息论和概率论，它能应用在Fano不等式和讯号源编码定理的证明。

约西亚·吉布斯在19世纪提出它。

证明

吉布斯不等式等价于：

$$0 \geq \sum_{i=1}^{n} p_i \log q_i - \sum_{i=1}^{n} p_i \log p_i = \sum_{i=1}^{n} p_i \log(q_i/p_i) = -D_{\mathrm{KL}}(P\|Q)$$（见相对熵）

证明最右的项小于或等于0的方法有几种，

- 已知 $\ln(x) \leq x - 1$，等号成立当且仅当 x=1。

$$\sum_{i=1}^{n} p_i \log(q_i/p_i) \leq \sum_{i=1}^{n} p_i(q_i/p_i - 1) = \sum_{i=1}^{n}(q_i - p_i) = \sum_{i=1}^{n} q_i - \sum_{i=1}^{n} p_i = 0$$

- 根据对数和不等式或延森不等式：

$$\sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \geq (\sum_{i=1}^{n} p_i) \log \frac{\sum_{i=1}^{n} p_i}{\sum_{i=1}^{n} q_i} = 0$$

3. 相对熵的应用

相对熵可以衡量两个随机分布之间的距离，当两个随机分布相同时，它们的相对熵为零，当两个随机分布的差别增

大时，它们的相对熵也会增大。所以相对熵（KL散度）可以用于比较文本的相似度，先统计出词的频率，然后