


(/apps/redirect?utm\_source=side-banner-click)

# 如何理解K-L散度（相对熵）

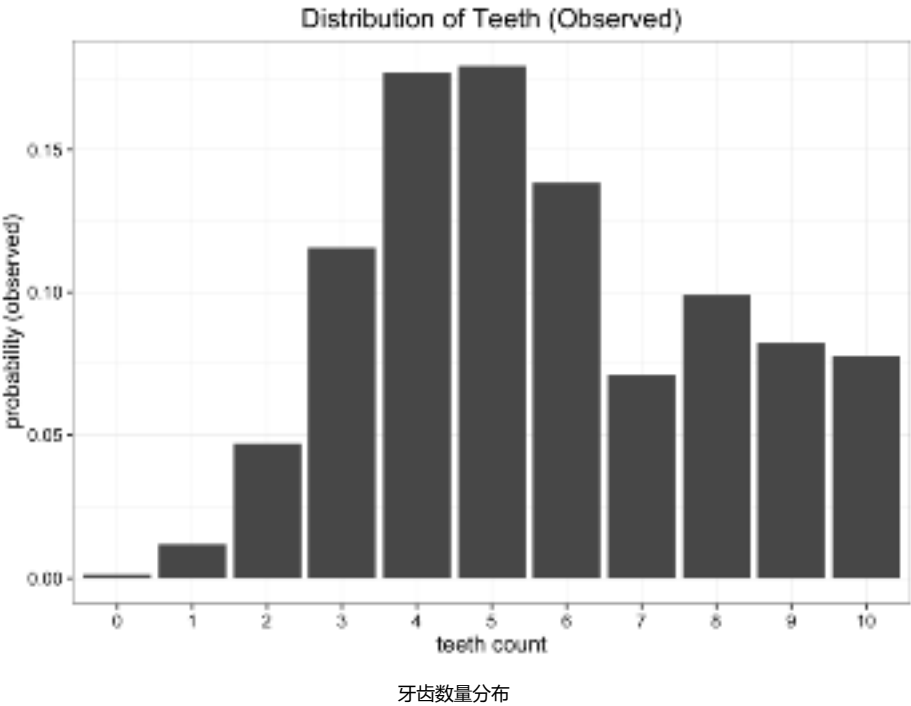
 Aspirinrin (/u/4c432a56a21a) [+关注](#)  
2017.06.28 17:58\* 字数 3267 阅读 7041 评论 3 喜欢 19 赞赏 1  
(/u/4c432a56a21a)

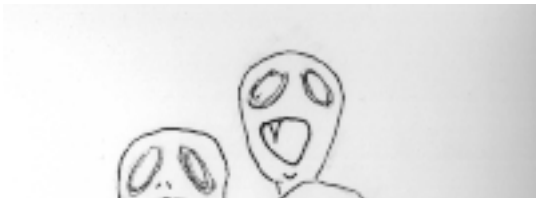
Kullback-Leibler Divergence，即 K-L散度，是一种量化两种概率分布P和Q之间差异的方式，又叫 相对熵。在概率学和统计学上，我们经常会使用一种 更简单的、近似的分布 来替代 观察数据 或 太复杂的分布。K-L散度能帮助我们度量使用一个分布来近似另一个分布时所损失的信息。

K-L散度定义见文末附录1。另外在附录5中解释了为什么在深度学习中，训练模型时使用的是 Cross Entropy 而非 K-L Divergence。

我们从下面这个问题出发思考K-L散度。

假设我们是一群太空科学家，经过遥远的旅行，来到了一颗新发现的星球。在这个星球上，生存着一种长有牙齿的蠕虫，引起了我们的研究兴趣。我们发现这种蠕虫生有10颗牙齿，但是因为不注意口腔卫生，又喜欢嚼东西，许多蠕虫会掉牙。收集大量样本之后，我们得到关于蠕虫牙齿数量的经验分布，如下图所示

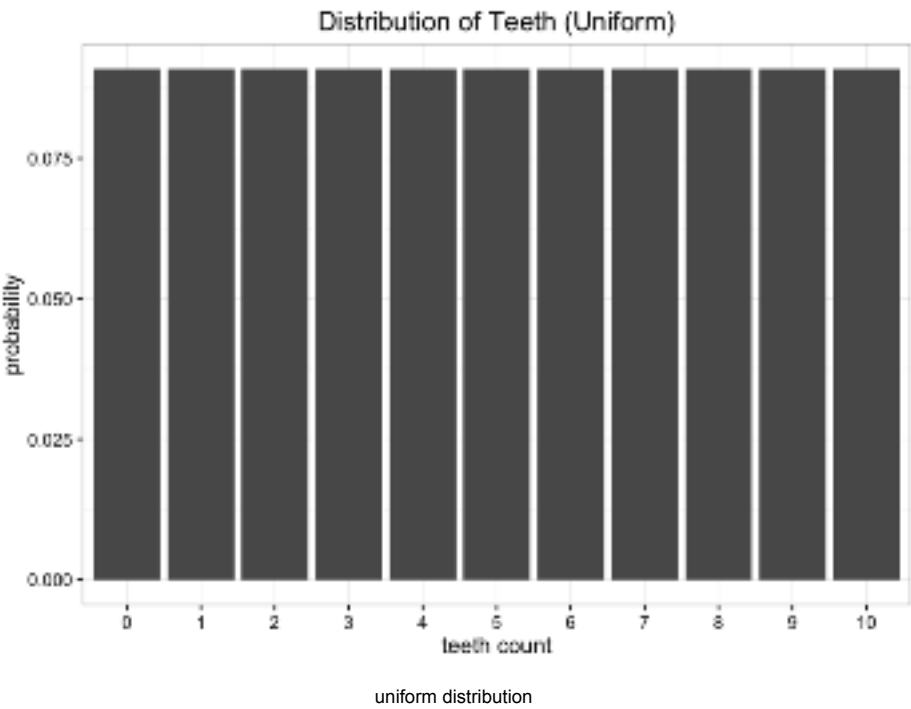




会掉牙的外星蠕虫

(/apps/redirect?utm\_source=side-banner-click)

这些数据很有价值，但是也有点问题。我们距离地球太远了，把这些概率分布数据发送回地球过于昂贵。还好我们是一群聪明的科学家，用一个只有一两个参数的简单模型来近似原始数据会减小数据传送量。最简单的近似模型是 均分布，因为蠕虫牙齿不会超过 10颗，所以有11个可能值，那蠕虫的牙齿数量概率都为  $1/11$ 。分布图如下：



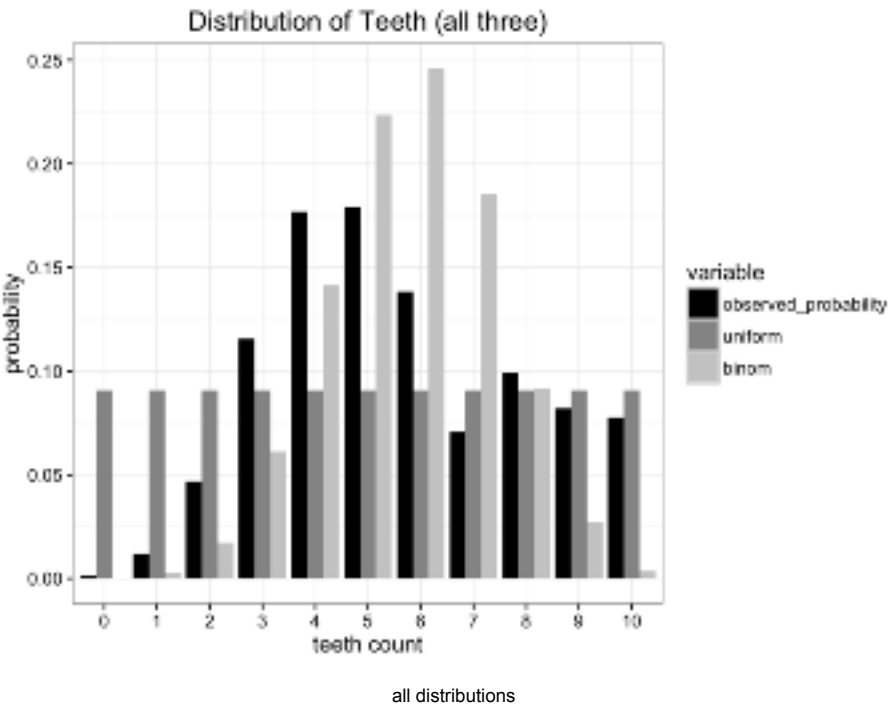
显然我们的原始数据并非均分布的，但也不是我们已知的分布，至少不是常见的分布。作为备选，我们想到的另一种简单模型是 二项式分布binomial distribution。蠕虫嘴里面共有  $n=10$  个牙槽，每个牙槽出现牙齿与否为独立事件，且概率均为  $p$ 。则蠕虫牙齿数量即为期望值  $E[x]=n \cdot p$ ，真实期望值即为观察数据的平均值，比如说 5.7，则  $p=0.57$ ，得到如下图所示的二项式分布：



binomial

对比一下原始数据，可以看出均分布和二项分布都不能完全描述原始分布。

(/apps/redirect?utm\_source=side-banner-click)



可是，我们不禁要问，哪一种分布更加接近原始分布呢？  
已经有许多度量误差的方式存在，但是我们所要考虑的是减小发送的信息量。上面讨论的均分布和二项式分布都把问题规约到只需要两个参数，牙齿数量和概率值（均分布只需要牙齿数量即可）。那么哪个分布保留了更多的原始数据分布的信息呢？这个时候就需要K-L散度登场了。

数据的熵

K-L散度源于信息论。信息论主要研究如何量化数据中的信息。最重要的信息度量单位是熵 Entropy，一般用  $H$  表示。分布的熵的公式如下：

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

Entropy

上面对数没有确定底数，可以是 2、e 或 10，等等。如果我们使用以 2 为底的对数计算 H 值的话，可以把这个值看作是编码信息所需要的最少二进制位数 bits。上面空间蠕虫的例子中，信息指的是根据观察所得的经验分布给出的蠕虫牙齿数量。计算可以得到原始数据概率分布的熵值为 3.12 bits。这个值只是告诉我们编码蠕虫牙齿数量概率的信息需要的二进制位 bit 的位数。

可是熵值并没有给出压缩数据到最小熵值的方法，即如何编码数据才能达到最优（存储空间最优）。优化信息编码是一个非常有意思的主题，但并不是理解K-L散度所必须的。熵的主要作用是告诉我们最优编码信息方案的理论下界（存储空间），以及度量数据的信息量的一种方式。理解了熵，我们就知道有多少信息蕴含在数据之中，现在我们就可以计算当我们用一个带参数的概率分布来近似替代原始数据分布的时候，到底损失了多少信息。请继续看下节内容。↓↓↓



## K-L散度量信息损失

只需要稍加修改 熵H 的计算公式就能得到 K-L散度 的计算公式。设  $p$  为观察得到的概率分布， $q$  为另一分布来近似  $p$ ，则  $p$ 、 $q$  的 K-L散度 为：

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

entropy-p-q

显然，根据上面的公式，K-L散度其实是数据的原始分布 $p$ 和近似分布 $q$ 之间的对数差值的期望。如果继续用 2 为底的对数计算，则**K-L散度值表示信息损失的二进制位数**。下面公式以期表达K-L散度：

$$D_{KL}(p||q) = E[\log p(x) - \log q(x)]$$

DKL1

一般，K-L散度以下面的书写方式更常见：

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

DKL2

注： $\log a - \log b = \log (a/b)$

OK，现在我们知道当用一个分布来近似另一个分布时如何计算信息损失量了。接下来，让我们重新回到最开始的蠕虫牙齿数量概率分布的问题。

## 对比两种分布

首先是用均分布来近似原始分布的K-L散度：

$$D_{kl}(\text{Observed} || \text{Uniform}) = 0.338$$

DKL-uniform

接下来计算用二项式分布近似原始分布的K-L散度：

$$D_{kl}(\text{Observed} || \text{Binomial}) = 0.477$$

DKL-binomial

通过上面的计算可以看出，使用均分布近似原始分布的信息损失要比用二项式分布近似小。所以，如果要从均分布和二项式分布中选择一个的话，均分布更好些。

## 散度并非距离

很自然地，一些同学把K-L散度看作是不同分布之间距离的度量。这是不对的，因为从K-L散度的计算公式就可以看出它不符合对称性（距离度量应该满足对称性）。如果用上文观察的数据分布来近似二项式分布，得到如下结果：

$$D_{kl}(\text{Binomial} || \text{Observed}) = 0.330$$

(/apps/redirect?  
utm\_source=side-  
banner-click)

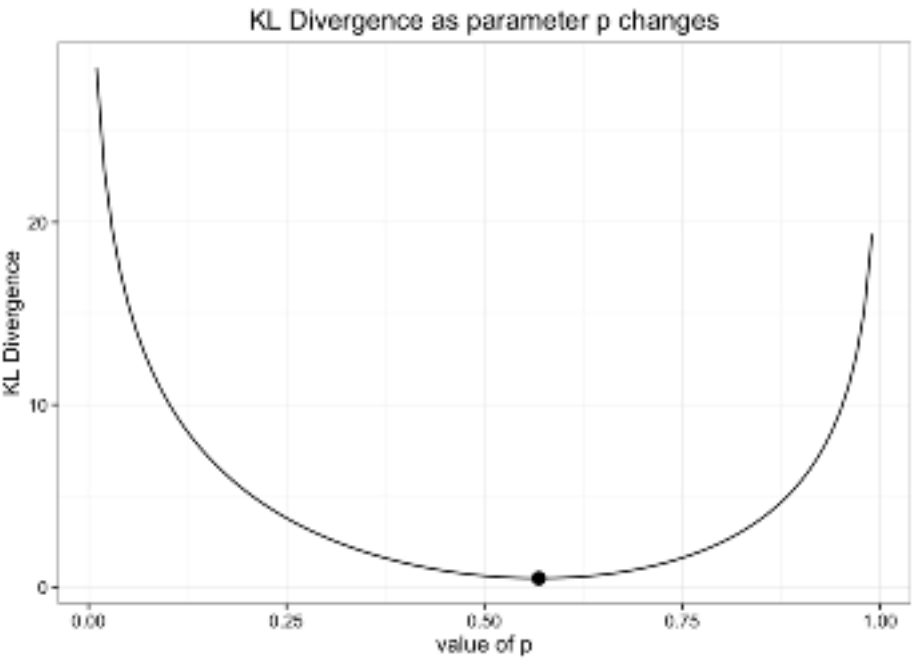
(/apps/redirect?utm\_source=side-banner-click)

所以， $D_{kl}(\text{Observed} || \text{Binomial}) \neq D_{kl}(\text{Binomial} || \text{Observed})$ 。  
也就是说，用  $p$  近似  $q$  和用  $q$  近似  $p$ ，二者所损失的信息并不是一样的。

使用K-L散度优化模型

前面使用的二项式分布的参数是概率  $p=0.57$ ，是原始数据的均值。 $p$  的值域在  $[0, 1]$  之间，我们要选择一个  $p$  值，建立二项式分布，目的是最小化近似误差，即K-L散度。那么  $0.57$  是最优的吗？

下图是原始数据分布和二项式分布的K-L散度变化随二项式分布参数  $p$  变化情况：



二项分布K-L值变化曲线

通过上面的曲线图可以看出，K-L散度值在圆点处最小，即  $p=0.57$ 。所以我们之前的二项式分布模型已经是最优的二项式模型了。注意，我已经说了，是而像是模型，这里只限定在二项式模型范围内。

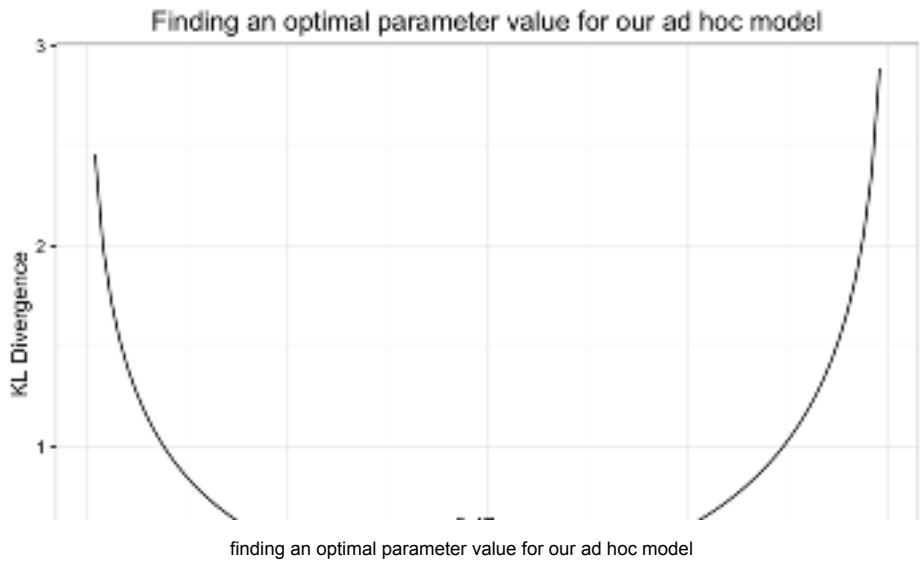
前面只考虑了均分布模型和二项式分布模型，接下来我们考虑另外一种模型来近似原始数据。首先把原始数据分成两部分，1) 0-5颗牙齿的概率和 2) 6-10颗牙齿的概率。概率值如下：

$$[6, 11] = \frac{p}{5}; [0, 5] = \frac{1 - p}{6}$$

ad hoc model

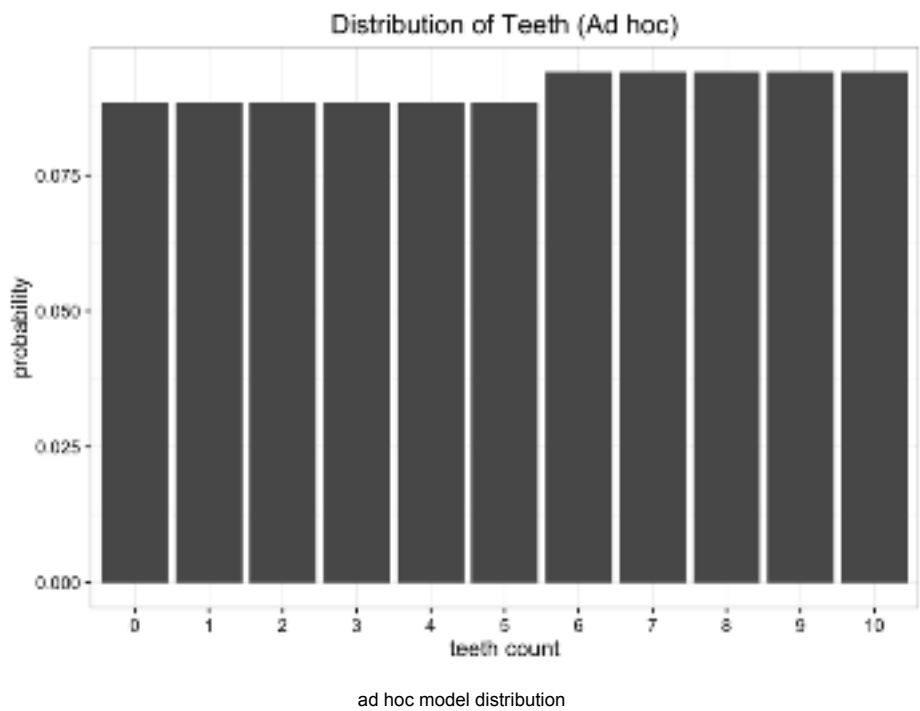
即，一只蠕虫的牙齿数量  $x=i$  的概率为  $p/5$ ； $x=j$  的概率为  $(1-p)/6$ ， $i=0,1,2,3,4,5$ ； $j=6,7,8,9,10$ 。  
Aha，我们自己建立了一个新的（奇怪的）模型来近似原始的分布，模型只有一个参数  $p$ ，像前面那样优化二项式分布的时候所做的一样，让我们画出K-L散度值随  $p$  变化的情况：





(/apps/redirect?  
utm\_source=side-  
banner-click)

当  $p=0.47$  时，K-L值取最小值  $0.338$ 。似曾相识吗？对，这个值和使用均分布的K-L散度值是一样的（这并不能说明什么）！下面我们继续画出这个奇怪模型的概率分布图，看起来确实和均分布的概率分布图相似：



我们自己都说了，这是个奇怪的模型，在K-L值相同的情况下，更倾向于使用更常见的、更简单的均分布模型。

回头看，我们在这一小节中使用K-L散度作为目标方程，分别找到了二项式分布模型的参数  $p=0.57$  和上面这个随手建立的模型的参数  $p=0.47$ 。是的，这就是本节的重点：**使用K-L散度作为目标方程来优化模型**。当然，本节中的模型都只有一个参数，也可以拓展到有更多参数的高维模型中。

### 变分自编码器VAEs和变分贝叶斯法

如果你熟悉神经网络，你肯能已经猜到我们接下来要学习的内容。除去神经网络结构的细节信息不谈，整个神经网络模型其实是在构造一个参数数量巨大的函数（百万级，甚至更多），不妨记为  $f(x)$ ，通过设定目标函数，可以训练神经网络逼近非常复杂的真实函数  $g(x)$ 。训练的关键是要设定目标函数，反馈给神经网络当前的表现如何。训练过程就是不断减小目标函数值的过程。



我们已经知道K-L散度用来度量在逼近一个分布时的信息损失量。K-L散度能够赋予神经网络近似表达非常复杂数据分布的能力。变分自编码器（Variational Autoencoders, VAEs）是一种能够学习最佳近似数据集中信息的常用方法，Tutorial on Variational Autoencoders 2016 (<https://link.jianshu.com?t=https://arxiv.org/abs/1606.05908>)是一篇关于VAEs的非常不错的教程，里面讲述了如何构建VAE的细节。What are Variational Autoencoders? A simple explanation (<https://link.jianshu.com?t=https://medium.com/@dmonn/what-are-variational-autoencoders-a-simple-explanation-ea7dcca0e3>)简单介绍了VAEs，Building Autoencoders in Keras (<https://link.jianshu.com?t=https://blog.keras.io/building-autoencoders-in-keras.html>)介绍了如何利用Keras库实现几种自编码器。

(/apps/redirect?utm\_source=side-banner-click)

变分贝叶斯方法（Variational Bayesian Methods）是一种更常见的方法。这篇文章 (<https://link.jianshu.com?t=https://www.countbayesie.com/blog/2015/3/3/6-amazing-trick-with-monte-carlo-simulations>)介绍了强大的蒙特卡洛模拟方法能够解决很多概率问题。蒙特卡洛模拟能够帮助解决许多贝叶斯推理问题中的棘手积分问题，尽管计算开销很大。包括VAE在内的变分贝叶斯方法，都能用K-L散度生成优化的近似分布，这种方法对棘手积分问题能进行更高效的推理。更多变分推理（Variational Inference）的知识可以访问Edward library for python (<https://link.jianshu.com?t=http://edwardlib.org/>)。

**因为本人没有学习过VAE和变分推理，所以本节内容质量无法得到保证，我会联系这方面的朋友来改善本节内容，也欢迎大家在评论区给出建议**

译自：Kullback-Leibler Divergence Explained (<https://link.jianshu.com?t=https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained>)

作者：Will Kurt

If you enjoyed this post please subscribe (<https://link.jianshu.com?t=http://countbayesie.com/subscribe>) to keep up to date and follow @willkurt (<https://link.jianshu.com?t=https://twitter.com/willkurt>)!

If you enjoyed this writing and also like programming languages, you might like the book on Haskell (<https://link.jianshu.com?t=https://www.manning.com/books/learn-haskell>) I just finished due in print July 2017 (though nearly all the content is available online today).

## 附录

### 1. K-L 散度的定义

To measure the difference between two probability distributions over the same variable  $x$ , a measure, called the *Kullback-Leibler divergence*, or simply, the *KL divergence*, has been popularly used in the data mining literature. The concept was originated in probability theory and information theory.

The KL divergence, which is closely related to *relative entropy*, *information divergence*, and *information for discrimination*, is a non-symmetric measure of the difference between two probability distributions  $p(x)$  and  $q(x)$ . Specifically, the Kullback-Leibler (KL) divergence of  $q(x)$  from  $p(x)$ , denoted  $D_{KL}(p(x), q(x))$ , is a measure of the information lost when  $q(x)$  is used to approximate  $p(x)$ .

Let  $p(x)$  and  $q(x)$  are two probability distributions of a discrete random variable  $x$ . That is, both  $p(x)$  and  $q(x)$  sum up to 1, and  $p(x) > 0$  and  $q(x) > 0$  for any  $x$  in  $X$ .  $D_{KL}(p(x), q(x))$  is defined in Equation (2.1).

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)} \quad (2.1)$$

define K-L divergence

### 2. 计算K-L的注意事项

Notice that attention should be paid when computing the KL divergence. We know  $\lim_{p \rightarrow 0} p \log p = 0$ . However, when  $p \neq 0$  but  $q = 0$ ,  $D_{KL}(p||q)$  is defined as  $\infty$ . This means that if one event  $e$  is possible (i.e.,  $p(e) > 0$ ), and the other predicts it is absolutely impossible (i.e.,  $q(e) = 0$ ), then the two distributions are absolutely different. However, in practice, two distributions  $P$  and  $Q$  are derived from observations and sample counting, that is, from frequency distributions. It is unreasonable to predict in the derived probability distribution that an event is

(/apps/redirect?utm\_source=side-banner-click)

notice

### 3. 遇到 $\log 0$ 时怎么办

**Example 2.24. Computing the KL Divergence by Smoothing.** Suppose there are two sample distributions  $P$  and  $Q$  as follows:  $P : (a : 3/5, b : 1/5, c : 1/5)$  and  $Q : (a : 5/9, b : 3/9, d : 1/9)$ . To compute the KL divergence  $D_{KL}(P||Q)$ , we introduce a small constant  $\epsilon$ , for example  $\epsilon = 10^{-3}$ , and define a smoothed version of  $P$  and  $Q$ ,  $P'$  and  $Q'$ , as follows.

The sample set observed in  $P$ ,  $SP = \{a, b, c\}$ . Similarly,  $SQ = \{a, b, d\}$ . The union set is  $SU = \{a, b, c, d\}$ . By smoothing, the missing symbols can be added to each distribution accordingly, with the small probability  $\epsilon$ . Thus, we have  $P' : (a : 3/5 - \epsilon/3, b : 1/5 - \epsilon/3, c : 1/5 - \epsilon/3, d : \epsilon)$  and  $Q' : (a : 5/9 - \epsilon/3, b : 3/9 - \epsilon/3, c : \epsilon, d : 1/9 - \epsilon/3)$ .  $D_{KL}(P', Q')$  can be computed easily.

example for K-L smoothing

### 4. 信息熵、交叉熵、相对熵

- 信息熵，即熵，香浓熵。编码方案完美时，最短平均编码长度。
- 交叉熵，cross-entropy。编码方案不一定完美时（由于对概率分布的估计不一定正确），平均编码长度。是神经网络常用的损失函数。
- 相对熵，即K-L散度，relative entropy。编码方案不一定完美时，平均编码长度相对于最小值的增加值。

更详细对比，见知乎如何通俗的解释交叉熵与相对熵？(https://link.jianshu.com/?t=https://www.zhihu.com/question/41252833)

### 5. 为什么在神经网络中使用交叉熵损失函数，而不是K-L散度？

K-L散度=交叉熵-熵，即  $D_{KL}(p||q) = H(p, q) - H(p)$ 。

在神经网络所涉及到的范围内， $H(p)$  不变，则  $D_{KL}(p||q)$  等价  $H(p, q)$ 。

更多讨论见Why do we use Kullback-Leibler divergence rather than cross entropy in the t-SNE objective function? (https://link.jianshu.com/?t=https://stats.stackexchange.com/questions/265966/why-do-we-use-kullback-leibler-divergence-rather-than-cross-entropy-in-the-t-sne)和Why train with cross-entropy instead of KL divergence in classification? (https://link.jianshu.com/?t=https://www.reddit.com/r/MachineLearning/comments/4mebv/why\_train\_with\_cross\_entropy\_instead\_of\_kl/)

https://stats.stackexchange.com/questions/265966/why-do-we-use-kullback-leibler-divergence-rather-than-cross-entropy-in-the-t-sne)和Why train with cross-entropy instead of KL divergence in classification? (https://link.jianshu.com/?t=https://www.reddit.com/r/MachineLearning/comments/4mebv/why\_train\_with\_cross\_entropy\_instead\_of\_kl/)

打赏作者e(2.71)元钱，请他喝杯咖啡，继续写作。

赞赏支持



Data Scientist (/nb/7072011)

举报文章 © 著作权归作者所有



Aspirinrin (/u/4c432a56a21a)

写了 39937 字，被 123 人关注，获得了 164 个喜欢  
(/u/4c432a56a21a)

+ 关注

但行好事，莫问前程





喜欢 | 19








更多分享

(http://cwb.assets.jianshu.io/notes/images/1397206  
(/apps/redirect?  
utm\_source=side-  
banner-click)



(/apps/redirect?utm\_source=note-bottom-click)



登录 (/sign-in?utm\_source=desktop&utm\_medium=not-signed-in-comment-form)

3条评论 只看作者

按时间倒序 按时间正序



十日立 (/u/a08f88f9ed9d)  
3楼 · 2018.07.12 16:29  
(/u/a08f88f9ed9d)  
请问有代码吗？

赞 回复



followStep (/u/331157c43a4f)  
2楼 · 2018.03.21 22:26  
(/u/331157c43a4f)  
写的不错，继续加油！

赞 回复

Aspirinn (/u/4c432a56a21a)：@followStep (/users/331157c43a4f) 谢谢支持  
2018.03.23 10:11 回复

添加新评论

被以下专题收入，发现更多相似内容



ML (/c/24fd0e28ca14?utm\_source=desktop&utm\_medium=notes-included-collection)



机器学习 (/c/12685433b6a7?utm\_source=desktop&utm\_medium=notes-included-collection)

高考3500 (/p/0bda5d804ee3?utm\_campaign=maleskine&utm\_content=...

A a (an) [ə, eɪ(ə)] art. 一（个、件……） abandon [əˈbændən] v.抛弃，舍弃，放弃 ability [əˈbɪlɪti] n. 能力；才能 able [ˈeɪb(ə)] a. 能够；有能力的 abnormal [æbˈnɔːm...



0涂桃子 (/u/02eb49244585?  
utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation)

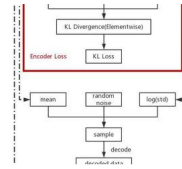
交叉熵 相对熵（KL散度/互熵） (/p/514e871cf230?utm\_campaign=males...

香农熵 熵考察（香农熵）的是单个的信息（分布）的期望：反映了一个系统的无序化（有序化）程度，一个系统越有序，信息熵就越低，反之就越高。交叉熵 交叉熵考察的是两个的信息（分布）的期望：交叉熵...



Arya鑫 (/u/1d11532897bb?  
utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation)

(/p/05608799c3e1?

(/apps/redirect?  
utm\_source=side-  
banner-click)

utm\_campaign=maleskine&amp;utm\_content=note&amp;utm\_medium=seo\_notes&amp;utm\_source=recommendation)

**VAE、GAN、Info-GAN：全解深度学习三大生成模型 (/p/05608799c3e1?u...**

摘要：在深度学习之前已经有很多生成模型，但苦于生成模型难以描述难以建模，科研人员遇到了很多挑战，而深度学习的出现帮助他们解决了不少问题。本章介绍基于深度学习思想的生成模型——VAE和GAN...

肆虐的悲傷 (/u/9c38c7ddde43?

utm\_campaign=maleskine&amp;utm\_content=user&amp;utm\_medium=seo\_notes&amp;utm\_source=recommendation)

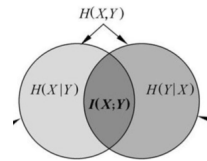
**自信息, 信息熵, 互信息和K-L散度 (/p/44744271d455?utm\_campaign=mal...**

香农-信息论领域的牛顿 香农一生发表的文章并不多，但是篇篇都是精品。A mathematical theory of communication通信的数学理论 第一篇文章中提出了比特（bit）的概念。比特究竟测量的是什么呢？香农...

HappyHorizon (/u/223b9e663367?

utm\_campaign=maleskine&amp;utm\_content=user&amp;utm\_medium=seo\_notes&amp;utm\_source=recommendation)

(/p/c123c7534500?



utm\_campaign=maleskine&amp;utm\_content=note&amp;utm\_medium=seo\_notes&amp;utm\_source=recommendation)

**浅谈自然语言处理基础（上） (/p/c123c7534500?utm\_campaign=maleski...**

本系列第三篇，承接前面的《浅谈机器学习基础》和《浅谈深度学习基础》。自然语言处理绪论 什么是自然语言处理？自然语言处理是研究人与人交际中以及人与计算机交际中的语言问题的一门学科。自然语言处...

我偏笑\_NSNirvana (/u/2293f85dc197?

utm\_campaign=maleskine&amp;utm\_content=user&amp;utm\_medium=seo\_notes&amp;utm\_source=recommendation)

**如何玩转微信小程序？ (/p/f745f9bbf042?utm\_campaign=maleskine&ut...**

小程序怎么玩？（1）场景货架：场景货架与小程序接口的打通。消费者对某一款爆品，热卖品有兴趣，都可将精准产品内容直接推给朋友或社群；当时间、地点、人、需求，在同一纬度发生，即是场景；消费都...

畅移员圈 (/u/e6bceb37fd62?

utm\_campaign=maleskine&amp;utm\_content=user&amp;utm\_medium=seo\_notes&amp;utm\_source=recommendation)

**【316】虚假的自信与虚假的无力 (/p/23c62e266998?utm\_campaign=mal...**

股市一片红牛的时候，所有人言必称股票，每天看着红色数字往上涨就是开心；股市一片惨绿的时候，大家对股市关注度就下降了，每天打开都是跌，也就不爱看、不爱讨论了。但其实，牛市的时候大盘整体都在...

踢球刘 (/u/df96c86ef06c?

utm\_campaign=maleskine&amp;utm\_content=user&amp;utm\_medium=seo\_notes&amp;utm\_source=recommendation)

**执笔人03 (/p/5e2234f67760?utm\_campaign=maleskine&utm\_content=n...**

第三章：初次相遇（一）广凡在街上走啊走，越走心里越没底，他不知道为什么父亲今天会这么晚，难道是推车坏了？毕竟车子用了那么久，而且现在在很多地方铁锈斑斑的。可是如果这样的话，可以通知我让我帮...

游离的某一人格 (/u/4317f5afe2d4?

utm\_campaign=maleskine&amp;utm\_content=user&amp;utm\_medium=seo\_notes&amp;utm\_source=recommendation)

**乡镇行 (/p/5df3486ec7e3?utm\_campaign=maleskine&utm\_content=not...**

日前又到乡镇走了走，看了看，一路走来有很多感慨，有很多惊喜，有很多无奈。还记得小时候出去的时候，都是步行，还记得那时，去亲戚家要过条河，会坐一段路的船，度到河的对岸去，每当这个时候就会...

秋若静美 (/u/298f9b786266?

utm\_campaign=maleskine&amp;utm\_content=user&amp;utm\_medium=seo\_notes&amp;utm\_source=recommendation)



(/p/6b166a5cf84c?




(/apps/redirect?  
utm\_source=side-  
banner-click)

utm\_campaign=maleskine&utm\_content=note&utm\_medium=seo\_notes&utm\_source=recommendation)

8. String to Integer (atoi) (/p/6b166a5cf84c?utm\_campaign=maleskine...

Implement atoi to convert a string to an integer. Hint: Carefully consider all possible input cases. If you want a challenge, please do n...

 YoungDayo (/u/4ee7fbb7d0a9?

utm\_campaign=maleskine&utm\_content=user&utm\_medium=seo\_notes&utm\_source=recommendation)

