机器学习中有关熵的概念

在具有不确定性研究问题中，才考虑信息量、熵的概念。针对一个问题，可能产生的多种事件，每种事件有不同的发生概率。

# 信息量：

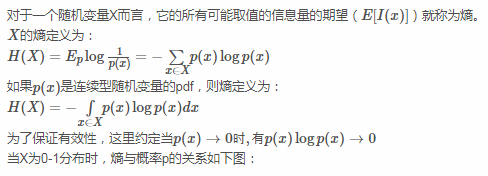
信息量就是以事件的发生概率来为基础来定义的。假设一个离散随机变量X，P(x)=P(X=x)，针对某一事件x0的信息量就定义为：I(x0) = -log(P(x0)).

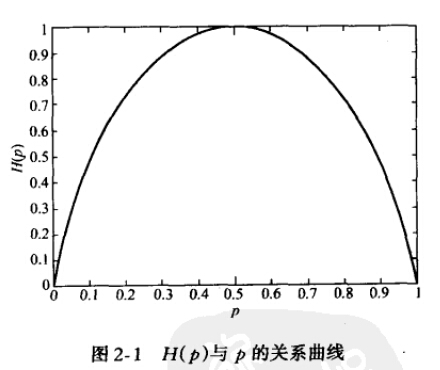
一个事件发生的概率越大，则它所携带的信息量就越小。

# [信息熵](https://baike.baidu.com/item/%E4%BF%A1%E6%81%AF%E7%86%B5)

信息熵是随机变量或整个系统的不确定性。熵越大，随机变量或系统的不确定性就越大。

熵：熵就是以信息量为基础来定义的。对一个随机变量X而言，它所有可能的取值的信息量的期望：E[I(x)]，就称为熵。





可以看出，当两种取值的可能性相等时，不确定度最大（此时没有任何先验知识），这个结论可以推广到多种取值的情况。在图中也可以看出，当p=0或1时，熵为0，即此时X完全确定。

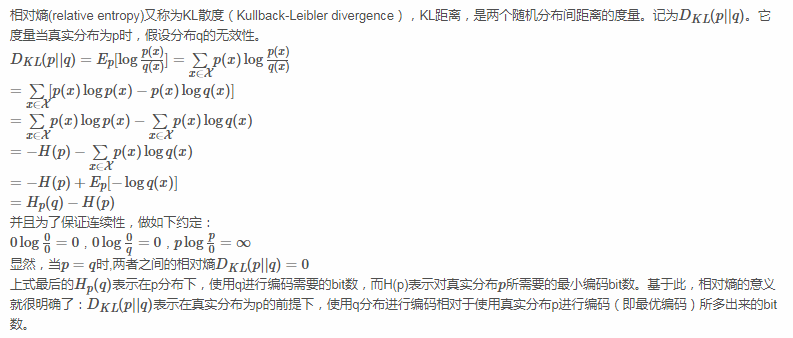
熵的单位随着公式中log运算的底数而变化，当底数为2时，单位为“比特”(bit)，底数为e时，单位为“奈特”。

# 相对熵

相对熵用来衡量两个取值为正的函数或概率分布之间的差异。

相对熵：是对两个随机分布间距离的度量。

当一个为真实分布P, 一个为假设分布Q。那么对随机变量X, 其相对熵定义为：E[log(P(x)/Q(x))] = 。相对熵就表示在真实分布P下，假设分布Q的无效性。



# [交叉熵](https://baike.baidu.com/item/%E4%BA%A4%E5%8F%89%E7%86%B5)

[交叉熵](https://baike.baidu.com/item/%E4%BA%A4%E5%8F%89%E7%86%B5)，用来衡量在给定的真实分布下，使用非真实分布所指定的策略消除系统的不确定性所需要付出的努力的大小。

交叉熵度量两个概率分布间的差异性信息

在信息论中，交叉熵是表示两个概率分布p,q，其中p表示真实分布，q表示非真实分布，在相同的一组事件中，其中，用非真实分布q来表示某个事件发生所需要的平均比特数。

假设现在有一个样本集中两个概率分布p,q，其中p为真实分布，q为非真实分布。假如，按照真实分布p来衡量识别一个样本所需要的编码长度的期望为：

H(p)=  https://gss1.bdstatic.com/9vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D134/sign=f29c291e4136acaf5de092ff48d98d03/ca1349540923dd546f3dcd4fdb09b3de9c82484e.jpg

但是，如果采用错误的分布q来表示来自真实分布p的平均编码长度，则应该是：

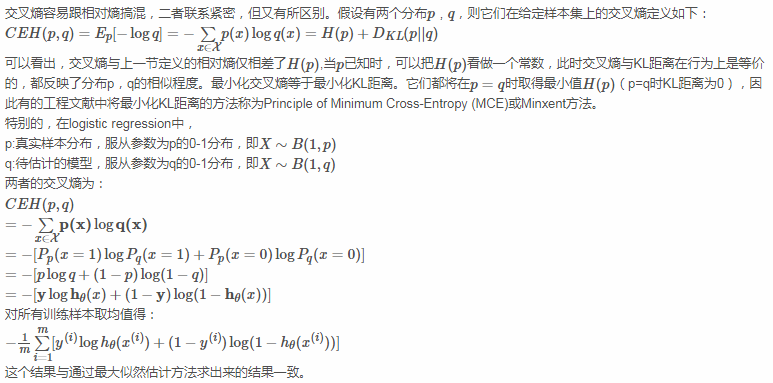
H(p,q)= https://gss3.bdstatic.com/-Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D133/sign=9170a8149b8fa0ec7bc7600e1596594a/37d3d539b6003af3c927b6473f2ac65c1038b67c.jpg

此时就将H(p,q)称之为交叉熵。交叉熵的计算方式如下：

对于离散变量采用以下的方式计算：H(p,q)= https://gss0.bdstatic.com/94o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D138/sign=8fca88ef104c510faac4e61958582528/30adcbef76094b36a4214dcaa9cc7cd98d109d61.jpg

对于连续变量采用以下的方式计算： https://gss0.bdstatic.com/-4o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D263/sign=3dda5d162b3fb80e08d166d105d12ffb/9d82d158ccbf6c81bc397dcfb63eb13533fa408e.jpg

相对熵：当一个为真实分布P, 一个为假设分布Q。那么对随机变量X, 其交叉熵定义为：E[-log(Q(x))] = 。表示：假设分布Q与真实分布P之间的差异。



# 相对熵、交叉熵、信息熵之间的关系

相对熵=交叉熵-信息熵： https://gss2.bdstatic.com/9fo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D181/sign=f263a1203a7adab439d01f4bbad5b36b/21a4462309f790529258553600f3d7ca7bcbd576.jpg