<https://blog.csdn.net/mieleizhi0522/article/details/80200126>

<https://blog.csdn.net/tsyccnh/article/details/79163834>

<http://www.cnblogs.com/yjmyzz/p/7822990.html>

<https://www.cnblogs.com/ljy2013/p/6432269.html>

信息量：下面的例子：

事件A：巴西队进入世界杯决赛圈

事件B：中国队进入世界杯决赛圈

可以看出事件B的信息量更大，因为其发生的概率小，概率越小的事件发生，其信息量就越大。

假设是一个离散型随机变量，其概率分布函数为，则的信息量为：概率对数的负数，概率越小，信息量越大



信息熵：对于某个事件，有n种可能性，每一种可能性的概率为，则信息熵为所有信息量的期望：



单分类（二项分布，0/1）的信息熵为：



交叉熵：其中为真实分布，为预测分布，n为可能性



交叉熵越小，则表示预测分布与真实分布越接近。

对于二项分布：n=2，交叉熵为：



如果有多个样本：



优达学城的例子：有3（样本数）道门，绿->红->蓝，是否有礼物，真实预测为(1, 1, 0)，预测的概率为(0.8, 0.7, 0.1)

交叉熵为：-ln0.8-ln0.7-ln0.9=0.69

如果预测的概率为(0.2, 0.3, 0.9)，则交叉熵为：-ln0.2-ln0.3-ln0.1=5.12

可以看出，预测结果与真实越接近，交叉熵越小，相反，交叉熵越大。

如果有n个类别，m个样本，交叉熵为：



其中，表示每个样本，各种类别的概率相加为1

交叉熵与梯度下降：

在机器学习中，可以把交叉熵作为误差函数，并进行梯度下降操作。

Logistic回归：采用Logistic函数作为激励函数



单分类情况下，误差函数为：为真实分布，为预测分布



其中：



公式推导：

首先求：



求导法则：

<http://netedu.xauat.edu.cn/jpkc/netedu/jpkc/gdsx/homepage/5jxsd/51/513/5302/530202.htm>



从而推出：



同理可以推出：



进一步推导：





梯度下降：

