既然分类问题不用mse作为损失函数，是因为函数非凸，那问什么回归要用呢？

**交叉熵/对数似然/逻辑斯蒂损失补充：**

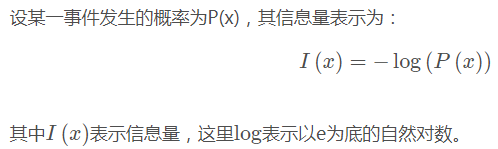
交叉熵是信息论中的一个重要概念，主要用于度量两个概率分布间的差异性，要理解交叉熵，需要先了解下面几个概念。

**信息量：**信息奠基人香农（Shannon）认为“信息是用来**消除**随机**不确定性**的东西”，也就是说衡量信息量的大小就是看这个信息消除不确定性的程度。

“太阳从东边升起”，这条信息并没有减少不确定性，因为太阳肯定是从东边升起的，这是一句废话，信息量为0。

”2018年中国队成功进入世界杯“，从直觉上来看，这句话具有很大的信息量。因为中国队进入世界杯的不确定性因素很大，而这句话消除了进入世界杯的不确定性，所以按照定义，这句话的信息量很大。

根据上述可总结如下：**信息量的大小与信息发生的概率成反比**。概率越大，信息量越小。概率越小，信息量越大。



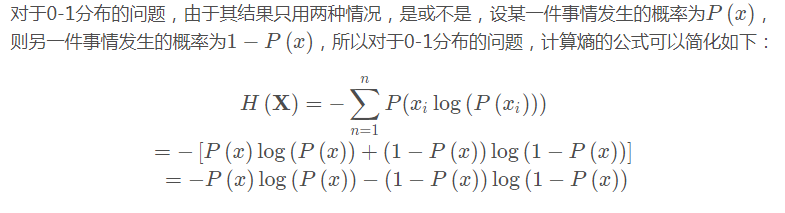
### 信息熵

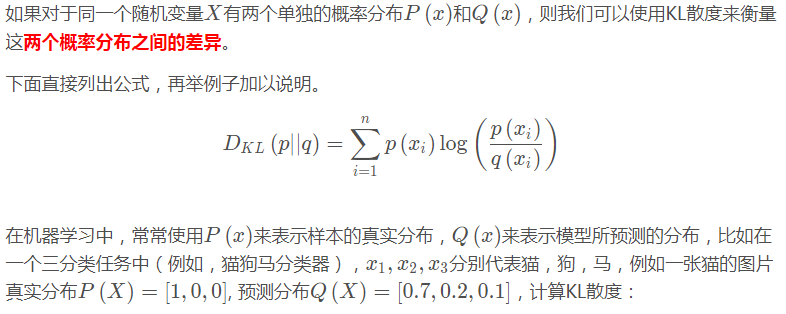
信息熵也被称为熵，用来表示所有信息量的期望。

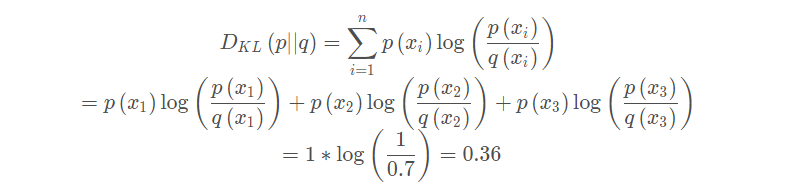
期望是试验中每次可能结果的概率乘以其结果的总和。

使用明天的天气概率来计算其信息熵：

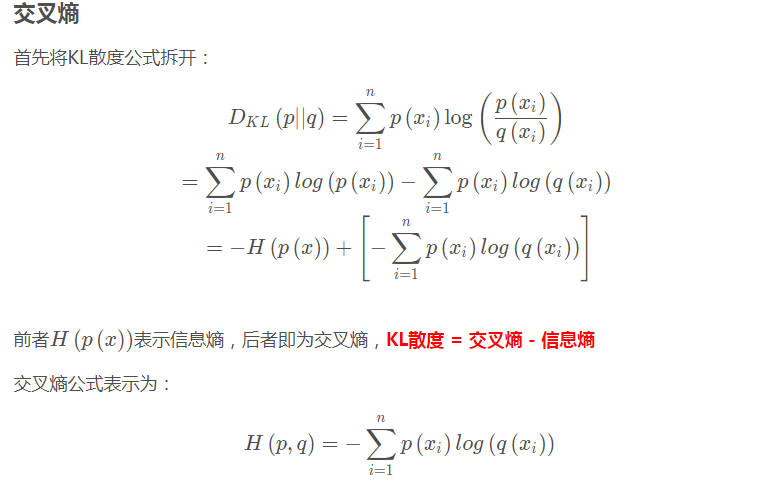


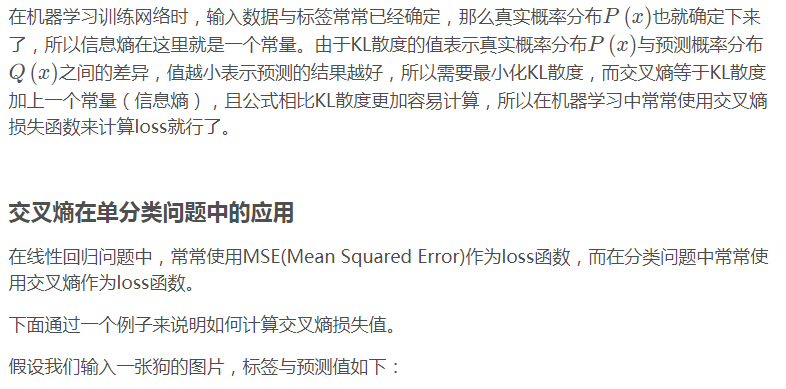
相对熵（KL散度）

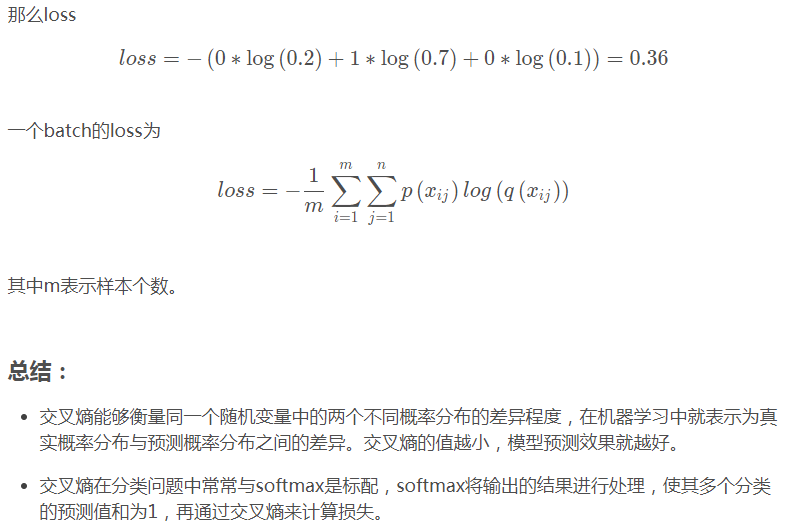


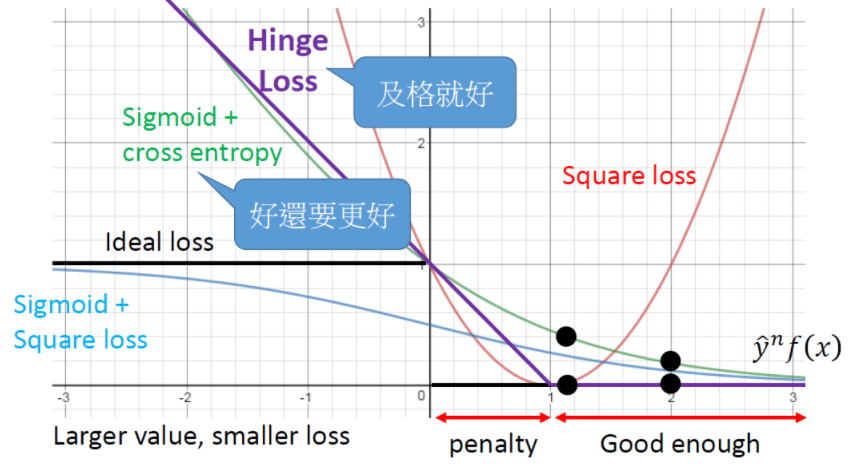












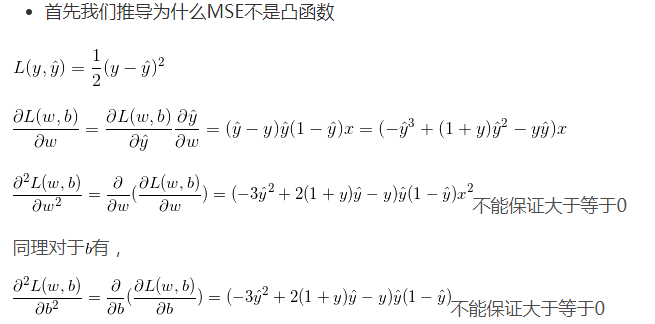
一般而言，我们认为预测值和实际值的乘积越大越好，即：乘积越大损失越小。相比较而言：

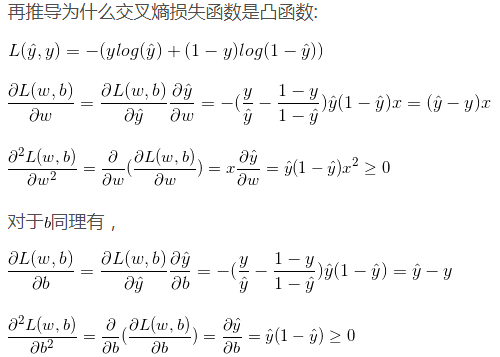
1、单独使用MSE：不合适，当yhat\*f(x)大于1的时候，loss应该下降才对；

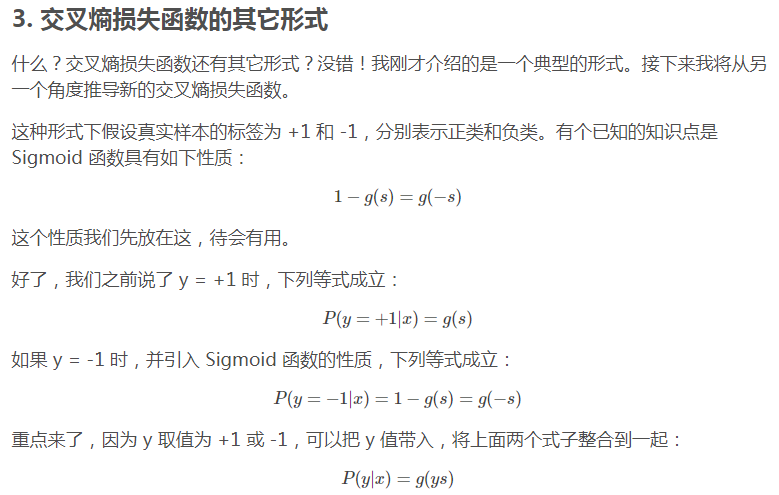
2、sigmoid+交叉熵优于sigmoid+mse：当乘积很小的时候，需要快速的下降，但是sigmoid+mse梯度很小。

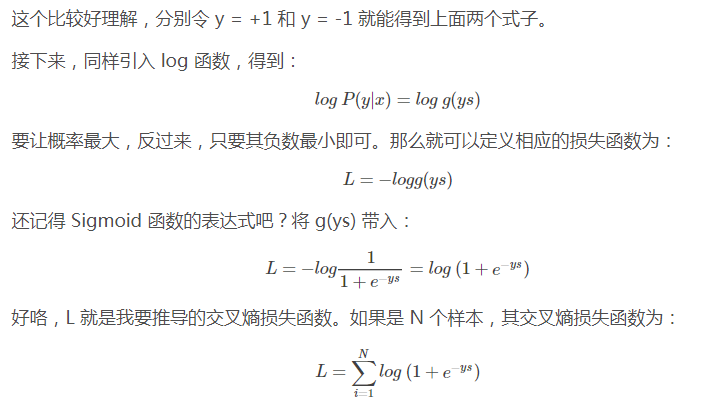
3、sigmoid+交叉熵与hinge loss相比：区别主要在于，当预测已经足够好的时候，hinge不再学习，交叉熵会继续学习，所以hinge的鲁棒性应该会更好。

## 03 分类问题代价函数选取

****





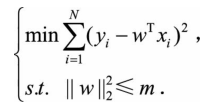


## L1正则与参数稀疏(葫芦书178)

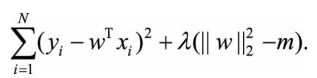
为什么L1正则化更容易产生稀疏解？

**从解空间的角度分析：**

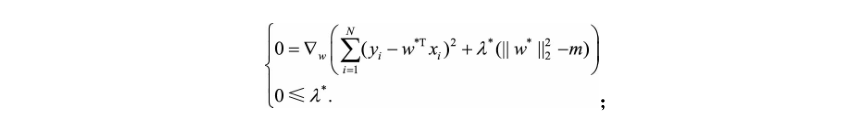
正则化可以理解为约束最优化问题，例如可以把目标函数改写为：



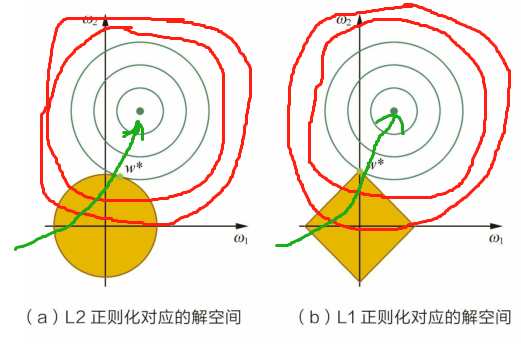
拉尔朗日函数为：



KKT条件包含以下两项：



可以得出：w的二范数小于等于m，也就是说解空间为圆形，同理，L1正则化的解空间为菱形。



当参数沿着目标函数的等高线梯度下降的时候，结合解空间可以发现：最优解要么在解空间内部，要么在解空间的边界，当时后者的时候，最优解更容易出现在菱形的尖点，这时候往往会导致参数稀疏。

**函数叠加角度：**

L1正则化,损失函数添加C|w|：在w<0的位置，导数为-C，w>0的位置，导数为C，假如不加正则的时候，导数的绝对值小于c，那么正则化以后，w<0区域导数恒为负，w>0的区域导数恒为正，所以最优解的位置出现在w为0。更容易导致参数稀疏。

L2正则化,损失函数添加C|w|2，导数为2wc，只在w为0的位置导数为0，只要原函数在原点的导数不为零，那么最优解就不会出现在原点处。

所以：L2只有减小w绝对值的作用，对参数稀疏没有贡献。

**贝叶斯先验角度**