# 一、熵（Entropy）

**直观理解：**事件的混乱程度—>事件的不确定性—>事件所含的信息量。

**基本假设：**发生概率越小的事件所包含的信息量就越大（bits），即

其中，表示事件所含的信息量。

**信息熵（Shannon Entropy）：**指整个系统中的不确定性的程度、信息量。

其中，表示信息熵，表示期望，表示事件的离散分布，表示事件所含的信息量。对上式进行进一步化简，可得：

**举例：**

|  |  |  |
| --- | --- | --- |
| **系统** | **事件的分布** | **信息熵** |
| Case1 |  |  |
| Case2 |  |  |
| Case3 |  |  |

可见，当一个系统中的事件越复杂，其所包含的信息熵越大，即信息量越大。

**必要知识补充1：**

**香农（Shannon）第一定理**指出信息熵是无失真信源编码的极限值，若编码的平均码长小于信息熵值，则必然发生差错（即有损）。例如在上述举例中，当我们采用二进制编码时，所需的最短无损编码为，所需的最短无损编码为，所需的最短无损编码为。

而应用在机器学习的损失函数计算中，我们通常采用自然常数作为对数的底，来进行信息熵的计算，即：

# 二、交叉熵（Cross Entropy）

**直观理解：**用于度量两个概率分布之间的差异性信息。

**交叉熵公式：**

若假设为目标分布（通常未知），为模型（model）输出的预测分布，则Cross Entropy为：

其中，表示用分布替换分布所需的编码量。而根据**香农第一定理**，我们知道是传达某一信息最少需要的编码量，因此：

当且仅当时，。

**结论：**

上述公式解释了——为什么在机器学习中，当我们采用交叉熵损失函数时，我们的训练目标是降低Cross Entropy的大小，因为只有这样才能使我们模型的预测分布不断接近于目标分布（真实分布）。

# 三、KL散度（Kullback-Leiber Divergence）

**一般理解：**KL Divergence衡量了两个分布之间的差异程度。

**公式：**

而实际上，，即：

因此，KL散度表示了以当前的编码方式最多还可以减少多少的编码量。此外，还需要特别注意，KL散度不是距离，它不具有对称性，即：

**思考：**

为什么在机器学习中，我们不使用KL散度来作为损失函数？

实际上，最小化（Minimize）交叉熵损失函数和最小化KL散度是等价的，因为我们已知，而在分类问题中，对于我们的学习模型（）来说是不变量，因此它无需梯度下降。所以，仅计算的效果和计算的效果是一样的。因此，采用交叉熵损失函数的效率更高。

**总结：**

* Shannon Entropy：

·代表传达一个系统所需要的最小信息量。

* Cross Entropy：

·代表用来编码所需要的信息量。

* KL Divergence：

·代表以目前的编码方式用的信息量还有多少下降空间。

# 四、最大似然估计（MLE）

**问题描述：**

给定一组Dataset 以及训练模型model的超参数，现需要找到一组model的权重参数，使得model能以最大概率生成的分布。

**解决思路：**

我们可以先随机设定一组权重参数，然后根据模型目前的和来计算产生分布的可能性，即，而越大则代表模型的输出越接近目标分布。然后，通过梯度下降，来不断更新优化的值，直到最优。

**公式化描述：**

其中，表示用方式求出的一组最优权重参数；表示当取到某一值时，此时括号内函数的值取到最大；表示用和生成目标分布的概率。

现在，若假设中有很多不同的数据集，且从每个数据集中进行一次抽样，则上述公式可改写为：

若每次抽样都是独立的，则：

若每次都是从同个分布进行抽样的，则：

**（以上就是基于独立同分布的假设所进行的推导。）**

进一步，对上述公式取对数，可得：

若所进行的训练是监督学习，而data的形式为，则：

若的产生是依赖于的，则：

又因为和模型model是独立的，所以：

根据对数的性质——相乘变相加，可得：

因为与优化无关，所以可以去掉多余项，得到：

将上述公式进一步改写：

接着，可进一步等价为：

因此，求一组权重参数的问题实际上就转变为了求最小交叉熵的问题。

**思考：**

MLE有什么不足？缺陷？

在MLE中，有一个基本假设，即所有出现的概率都是均等的，而这会产生一些无法避免的问题，举例如下：

假设有以下两组都可以得到相同的，那么你觉得哪一组更好？

而根据经验，通常我们认为参数组所构成的模型更稳定，较不易。但是MLE本身无法区别此类情况。所以这就是MLE本身所存在的缺陷。

# 五、最大后验估计（MAP）

**必要知识补充2：**

**贝叶斯定理：**

所以可得：

其中，同MLE中的；代表先验概率（Prior Probability），由人为给定，比如正态分布（normal distribution）；代表资料概率，通常与m无关；代表后验概率（Posterior Probability），即给定后出现的概率。

**问题描述：**同MLE中的问题描述。

**解决思路：**

采用贝叶斯定理来处理上述问题。再简单来说，相比MLE，多了一个先验概率的条件。

**公式化描述：**

若采样基于独立同分布假设，则：

进一步，对上述公式取对数，可得：

又因为和模型model是独立的，所以：

然后，将替换为，并把移到后面，则：

进一步，因为和模型model是独立的，所以：

因此，MAP相当于在优化MLE，而添加的优化项为。

**思考1：**

若为均匀分布（uniform distribution），则为一常数。此时，。因此，在贝叶斯定理的观点下，MLE仅为当没有先验假设时的MAP特例。

**思考2：**

由于MLE本身存在缺陷，而为了解决缺陷，我们希望的大小总是接近于0，因此，我们希望的分布平均值接近于0且方差有限。所以，当我们假设为正态分布（normal distribution）时，即：

于是，可得：

最终，我们就导出了**L2正则化**！因此，在机器学习中，我们在损失函数中使用的**L2正则化**其实就隐含着希望参数呈现正态分布的假设！

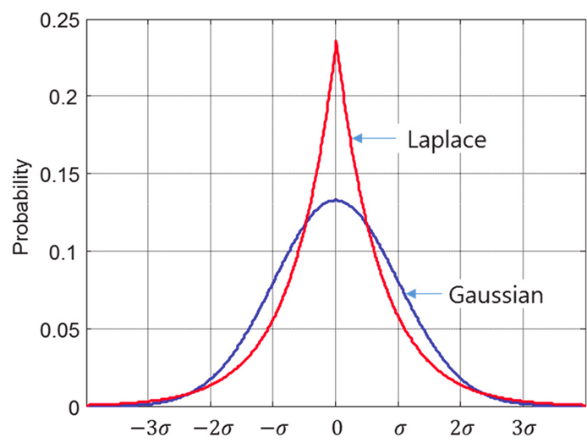
**思考3：**

若我们假设为均值为0的拉普拉斯分布（Laplace Distribution）时，即：

于是，可得：

最终，我们就导出了**L1正则化**！因此，在机器学习中，我们在损失函数中使用的**L1正则化**其实就隐含着希望参数呈现拉普拉斯分布的假设！

**补充说明：**



上图展示的是正态分布和拉普拉斯分布。

通常，在L2正则化下，因为是对权重参数的平方做惩罚，所以权重参数大小往往会比较均匀，而在L1正则化下，参数分布会比较稀疏。而这背后的实际原因来自于MAP中对于权重参数的Normal & Laplace Distribution假設。

**相关参考资料：**

1. [**https://www.bilibili.com/video/BV1hr4y1X7q5/?spm\_id\_from=333.999.0.0&vd\_source=36e948dc2acdb36d7055f879d377b529**](https://www.bilibili.com/video/BV1hr4y1X7q5/?spm_id_from=333.999.0.0&vd_source=36e948dc2acdb36d7055f879d377b529)
2. [**https://www.bilibili.com/video/BV1zj41127uG/?spm\_id\_from=333.999.0.0&vd\_source=36e948dc2acdb36d7055f879d377b529**](https://www.bilibili.com/video/BV1zj41127uG/?spm_id_from=333.999.0.0&vd_source=36e948dc2acdb36d7055f879d377b529)