softmax 回归相关理论

熵

信息论是将信息量化的理论，其最出名的概念叫 **熵**，香农给它的定义是 **无损编码事件信息的最小平均编码长度**。

其实用什么字符集编码都差不多，这里就默认用 $\{0,1\}$ 来编码了。

对一个概率分布来说，服从它的每个事件都有其发生的概率，假如为这些事件附两两不同的编码且这些编码互不为前缀（可以为后缀（想想为什么只要禁止互为前缀或互为后缀之一就可以避免歧义）），也就是无损编码，那么所谓平均编码长度，其实说白了就是编码长度的期望。

那么我们通过调整对各事件的编码，使编码长度的期望变得尽可能小，直到最小，这时候编码长度的期望就是这些事件服从的那个概率分布的熵了。

熵的计算公式是 $H[P] = -\sum_i p_i \log(p_i)$，这个公式是唯一的，证明有些繁琐，可参考深入理解信息熵 - 知乎 (zhihu.com)。

交叉熵

交叉熵的意义是衡量预测分布与真实分布之间的偏差，通常来说，交叉熵大于等于真实熵。

交叉熵的公式是 $H[P,Q] = -\sum_i p_i \log(q_i)$，也就是用 P 计算概率，用 Q 计算编码长度。

用交叉熵来计算分类模型的损失是十分自然的。

通过极大似然估计得到损失函数

如何定义在 softmax 回归模型上每个样本出现的概率呢？

书上的方法是直接把预测值 $\hat{y}$ 看成一个条件概率向量，向量中的每个数都是在输入为 $\mathbf{x}$ 下类别为对应物体的概率，那么对于每个样本，它们都是形如“输入为 $\mathbf{x}$ 下类别为某个物体”，所以 $\hat{y}$ 的对应位置的概率就是我们要的东西了，用它来进行极大似然估计。

我这里主要写我在另一个方向上的尝试和失败。

这里我们假设样本的类型向量之所以是独热编码而不是我们模型输出的概率向量，是因为噪声导致的，假设预测值加上噪声等于真实值，也就是 $\hat{y} + \varepsilon = y$，那么由于我们 $y.sum()$ 和 $\hat{y}.sum()$ 都是等于 1 的，那么这个噪声 ε 就必须满足 $\varepsilon.sum() = 0$。

只要能假设出这种噪声的分布，就可以在此假设的基础上进行极大似然估计。

但可惜……

所以我失败了，还是用书上的方法。