0、概述

给定一个随机变量的一些样本，我们如何根据这些样本去估计随机变量的分布？我们已经知道，若给定分布的含未知参数的形式，那么可以使用统计推断中的参数估计方法；若不给出分布的形式，而是给定一些关于分布的先验的“认识”作为约束，则应当使用最大熵估计原则。

最大熵原则的思想是，在所有可能的概率分布（即所有满足先验认识的分布）中，熵最大的分布是最好的分布，直观地说，我们在估计分布时，除了使分布满足已有的认识外，对不确定的部分都应当保持“等可能性”。

将最大熵原则应用于机器学习模型，就是最大熵模型。最大熵模型属于概率模型，并属于判别模型。

注意，机器学习需要求解的最优分布不是简单的一个随机变量的分布P(X)，而是条件分布P(Y|X)，但最大熵原则仍然是适用的，但此时“最大熵原则”变为“最大条件熵原则”。

一、模型的目标函数及优化问题

与其他机器学习模型一样，我们给定训练集

此外，需对“先验认识”进行形式化。我们定义“特征函数”，用来表示先验认识：



那么如何用利用特征函数去表示先验认识对条件分布的约束呢？这个我们后面再说。

为了使用最大熵模型，我们还需要在训练集上统计联合分布P(X,Y)和边缘分布P(X)，为了区别于“真实分布”，我们将它们称作经验分布（就是在训练集上直接统计出来的分布），并分别记作和：





其中，表示训练集中样本(x,y)出现的次数，表示训练集中样本输入x出现的次数。

则条件熵为：



先验认识形成的对分布的约束为：



等式右边项表示先验认识在训练集上的可信度；等式左边项表示先验认识基于所求分布P(y|x)的可信度。可见，先验认识对所求分布的约束就是要求基于分布计算出的先验认识的可信度等于训练集上先验认识的可信度。这相当于使用了训练集的信息，但没有使用其全部信息，只是使用了关于先验认识的那部分信息（而朴素贝叶斯模型则是使用了训练集上全部的信息）。

我们再解释一下，为什么上面这个约束等式中这两个求和项表示先验认识的“可信度”：以右边项为例，根据特征函数的意义，其对应于一个“认识”，若样本(x,y)不符合该认识，则特征函数在该样本上的值=0，若(x,y)符合该认识，则特征函数在该样本上的值=1，所以整个求和项就表示符合该认识的样本占总样本数的比例，也就表示该认识的可信度——比例越高，说明该认识的可信度越高。可以想象，若该认识是“绝对正确”的，那么训练集中所有样本都应该符合该认识，则求和项的值就等于1。

注意，在最大熵模型中，我们并不是用“先验认识”本身去约束所求分布，而是用先验认识在训练集上的“可信度”去约束。

最终的优化问题为：



写成极小化的形式：



注：总共有N个先验认识，对应N个特征函数。

最大熵模型与朴素贝叶斯模型类似，它们都不像logistic回归模型那样，给出了分布（条件分布或者联合分布）的形式——它们都没有事先给出分布的形式，它们的思路是根据定理或准则，基于训练集上的信息给出最优分布。

二、模型形式

尽管最大熵模型没有给出分布的形式，但是其形式隐含在目标函数和约束条件之中了。因此可以从其优化问题中推导出其形式。如何推导呢？与SVM类似，最大熵模型的优化问题也是约束优化问题，它可以通过对偶转化而推导出模型的形式。

首先，构造拉格朗日函数：



注意，拉格朗日函数的第一参数中的P是条件概率分布，这一点与SVM不同，SVM中拉格朗日函数的第一参数是模型的参数向量w。

则原始问题的对偶问题为极大极小化问题：



对偶问题内部的极小化问题的结果是关于w的函数，记为，并称为对偶函数。对应的最优解（也是关于w的函数）记为，即：



如何得到（关于w）的形式？由于是极小化得到的，所以对求关于P的偏导数，并令其等于0，得到：



其中，。

这就是所求最优分布的形式，它是关于未知参数w的函数。它依赖于先验认识。未知参数的求取需要极大化，容易推导出的形式：



这就是最大熵模型优化问题的最终的目标函数。求出使极大的参数，带到的表达式中，即可得到最优条件分布。

三、模型的优化算法

上面推导出来的最大熵模型优化问题的最终形式，是一个凸优化问题，很多优化方法都适用，比如梯度下降、牛顿法或拟牛顿法等。