熵描述事物的不确定性，或者说事物的信息量。可能性越多，不确定性越大，被获知时得到的信息量越大。反之，当事物是确定的，其信息量为０。

【特征函数的理解】

代码参考

http://blog.csdn.net/u014688145/article/details/55003910?fps=1&locationNum=2

总结很好

最大熵模型需要统计对应特征和标签的联合概率。比如按照样例，每个特征对应一个标签，都会得到你所说的[x\_1, y], [x\_2, y]等等，样本总数是self.\_N，所以是单独的每个特征加样本标签 除以 样本总数，表明在这单个特征结合样本标签在数据集中的联合概率。你所说的除以self.\_n是能够构成的特征数，但GIS模型并不是算这些特征在特征数的概率，其次字典是天然去重的，所以真正的特征数也是不知道的，self.\_n 只反映没有重复的（特征，标签）数。

【算法理解】

理论解释：https://vimsky.com/article/714.html

某人1的代码：https://vimsky.com/article/776.html

某人2改进的代码：http://blog.csdn.net/huludan/article/details/72742614

更易于理解的理论解释：

http://blog.csdn.net/u014688145/article/details/55003910?fps=1&locationNum=2

总结很好

<http://www.hankcs.com/ml/the-logistic-regression-and-the-maximum-entropy-model.html>

**最大熵建模**

输入：包含单词"in"的英文句子

输出：给出选择某个法语单词f 做为"in"的翻译结果的概率p(f)。

过程：从样本中抽取规则（特征） 基于这些规则建立模型

在自动英译汉的例子中，输出y∈Y = {dans, en, à, au cours de, pendant}，会受到句子中"in"上下文信息x∈X的影响。

**任务**：构建一个模型，给定上下文x的情况→输出y的概率p(y|x)

**1）训练数据**

样本数据集为(x1, y1), (x2, y2), ..., (xN, yN)

每个样本包括：在英文句子中"in"周围的单词x，"in"的翻译y。

用样本的经验分布p~来表示所有样本的分布特性：



其中N为训练样本的大小, num(x,y)是样本中某一对(x,y)同时出现的次数。

**2）特征和约束**

定义特征如下：



它表示某个特定的x和某个特定的y之间是否有一定的关系。例如，如果April这个词出现在in之后，那么in会被翻译成en，那么这个特征可以表示成：



* 特征f(x,y)关于训练样本的经验分布p~(x,y)的期望如下：

https://vimsky.com/wp-content/uploads/2015/05/11.gif

* 特征f(x,y)关于模型分布p(y|x)的理论期望值是：

https://vimsky.com/wp-content/uploads/2015/05/21.gif

约束期望概率值等于经验概率值

https://vimsky.com/wp-content/uploads/2015/05/31.gif

即

https://vimsky.com/wp-content/uploads/2015/05/4.gif

**3）最大熵模型**

条件分布p(y|x)的均匀度就是条件熵：

https://vimsky.com/wp-content/uploads/2015/05/5.gif

最大熵原理的定义：在满足约束条件的概率分布集合中，选择模型p\*∈C ，使得熵H(p)最大。即：

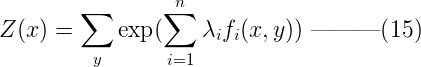
https://vimsky.com/wp-content/uploads/2015/05/61.gif

https://vimsky.com/wp-content/uploads/2015/05/71.gif

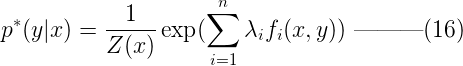
其中，C表示所有满足约束的模型集合，n为特征或者说特征函数fi的数量

------------------------------------------------求解-----------------------------------------------

定义归一化因子Z(x)如下：



则模型p(y|x)的最优解为：



接下来，需要求解参数，更新参数直到收敛。

【算法-书上】

用特征函数描述输入x和输出y之间的某个事实。其定义如下：



是一个二值函数。

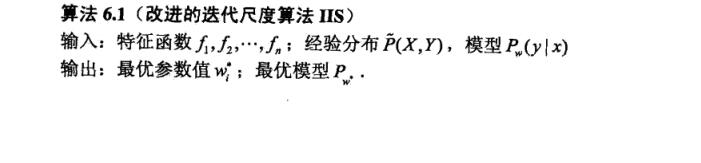
首先引入一个量表示所有特征在（x,y）出现的次数：

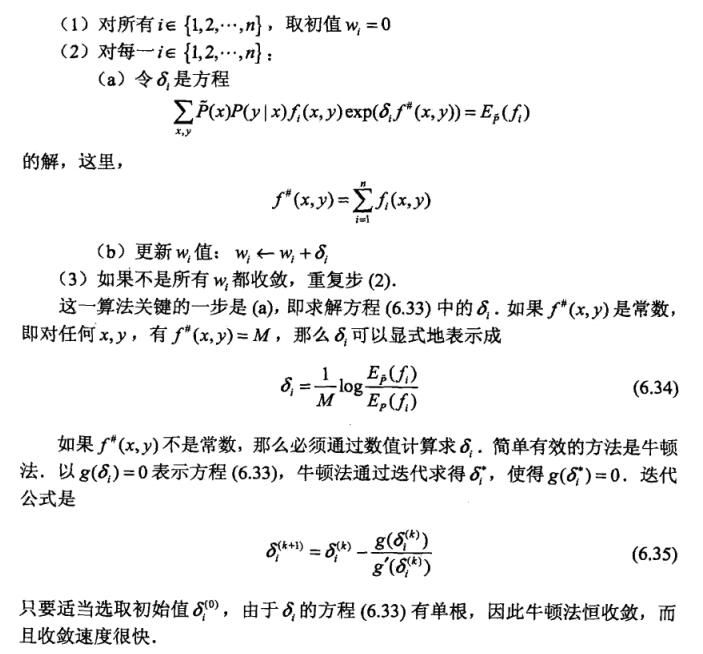


算法如下：

GIS算法：假设为常数，则迭代更新参数表达式可显示写出；

IIS算法：考虑了利用牛顿迭代法求解迭代更新参数。

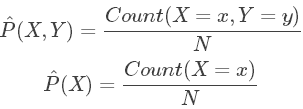




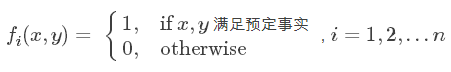
【代码】

**用到的公式：**

P(X,Y)和P(X)的经验分布



从样本中选取或者设计n个特征函数：



特征函数关于经验分布P^(X,Y)的期望：



特征函数关于理论模型P(X,Y)的期望：

