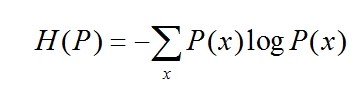
**熵与最大熵原理**

熵是随机变量不确定性的度量，不确定性越大，熵值就越大；若随机变量退化成定值，熵为0。均匀分布是“最不确定”的分布

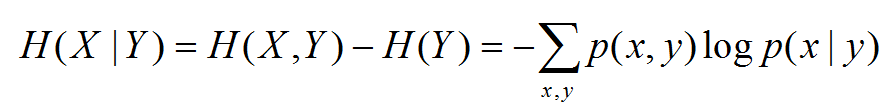
假设离散随机变量X的概率分布为P(x)，则其熵为：



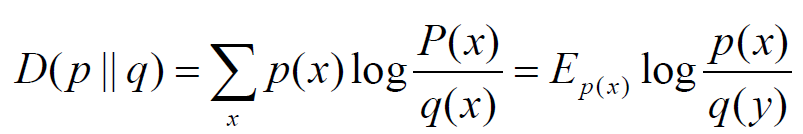
**联合熵和条件熵**

两个随机变量的X，Y的联合分布，可以形成联合熵，用H(X,Y)表示

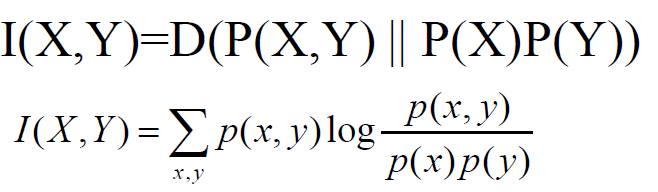
条件熵H(X|Y) = H(X,Y) - H(Y)



**相对熵与互信息** 设p(x),q(x)是X中取值的两个概率分布，则p对q的相对熵是：

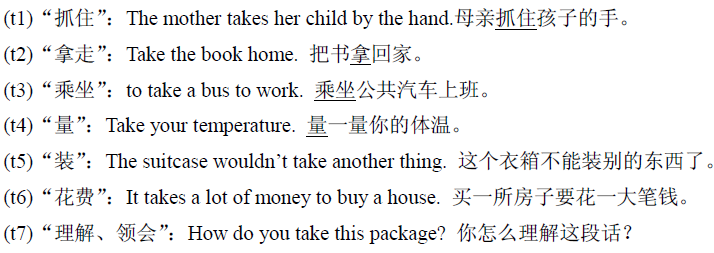


两个随机变量X，Y的互信息，定义为X，Y的联合分布和独立分布乘积的相对熵。



最大熵原理是统计学的一般原理，也是概率模型学习的一个准则。最大熵原理认为，学习概率模型时，在所有可能的概率模型中，熵最大的模型是最好的模型。

**最大熵模型的实例：**



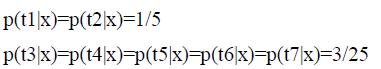
在英汉翻译中，take有多种解释例如上文中存在7中，在没有任何限制的条件下，最大熵原理认为翻译成任何一种解释都是等概率的。

https://images2015.cnblogs.com/blog/754644/201609/754644-20160910110134488-144248936.png

实际中总有或多的限制条件，例如t1,t2比较常见，假设满足

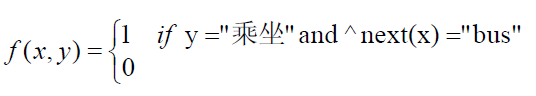
https://images2015.cnblogs.com/blog/754644/201609/754644-20160910110421754-1152671856.png

同样根据最大熵原理，可以得出：



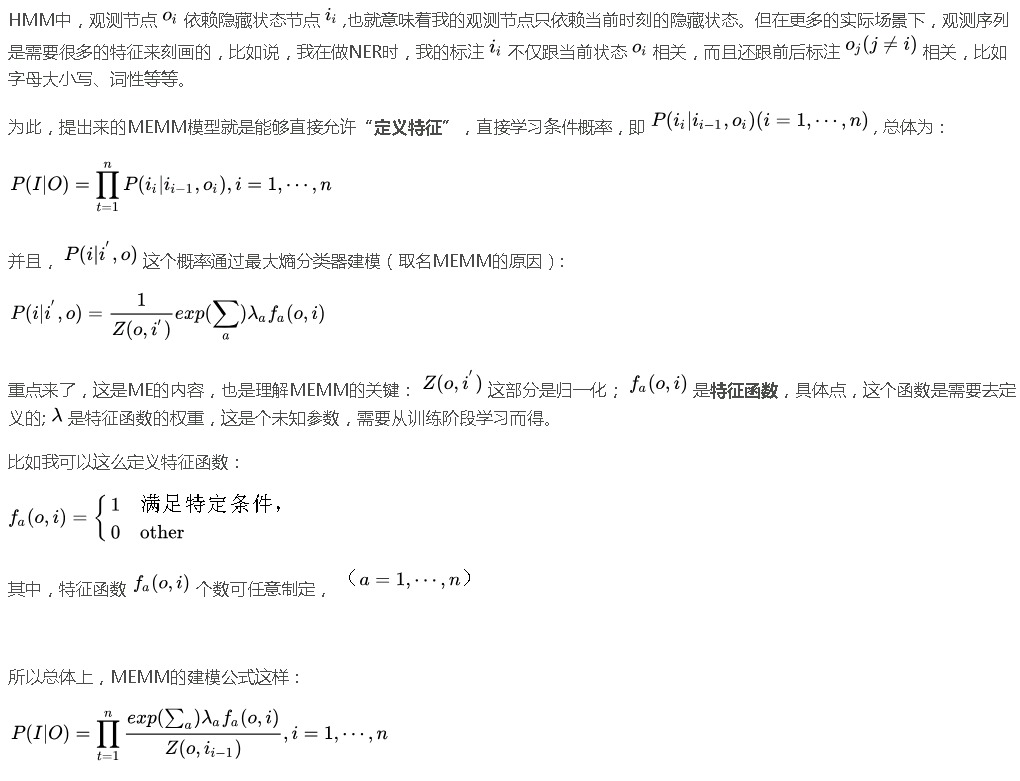
实际的统计模型中，还需要引入特征提高准确率。例如take翻译为乘坐的概率小，但是当后面跟着交通工具的名词“bus"，概率就变得非常大。

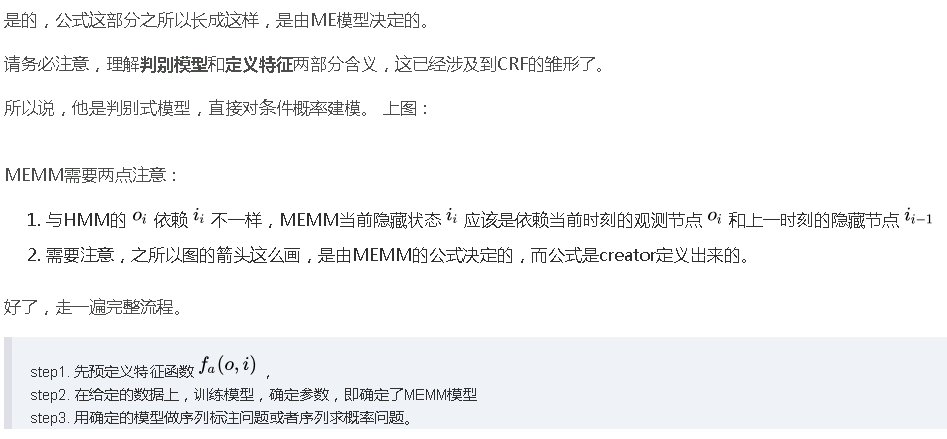
用**特征函数**f(x,y)描述输入x,输出y之间的某一个事实，只有0和1两种值，称为二值函数。例如：



最大熵模型根据最大熵原理在类似上面的特征限制下选择最优的概率分布。最大熵模型的学习过程就是求解最大熵模型的过程，可以形式化为约束最优化问题。

MEMM：





**模型运行过程**

MEMM模型的工作流程也包括了学习训练问题、序列标注问题、序列求概率问题。

**学习训练过程**

一套MEMM由一套参数唯一确定，同样地，我需要通过训练数据学习这些参数。MEMM模型很自然需要学习里面的特征权重λ。

不过跟HMM不用的是，因为HMM是生成式模型，参数即为各种概率分布元参数，数据量足够可以用最大似然估计。而判别式模型是用函数直接判别，学习边界，MEMM即通过特征函数来界定。但同样，MEMM也有极大似然估计方法、梯度下降、牛顿迭代发、拟牛顿下降、BFGS、L-BFGS等等。

**序列标注过程**

还是跟HMM一样的，用学习好的MEMM模型，在新的sample观测序列 上找出一条概率最大最可能的隐状态序列  。

只是现在的图中的每个隐状态节点的概率求法有一些差异而已,正确将每个节点的概率表示清楚，路径求解过程还是一样，采用viterbi算法。

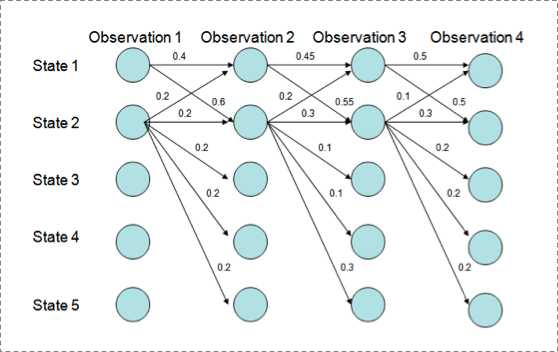
**序列求概率过程**

跟HMM举的例子一样的，也是分别去为每一批数据训练构建特定的MEMM，然后根据序列在每个MEMM模型的不同得分概率，选择最高分数的模型为wanted类别。

**标注偏置**

MEMM讨论的最多的是他的labeling bias 问题。

**现象**



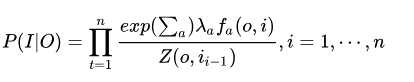
用Viterbi算法解码MEMM，状态1倾向于转换到状态2，同时状态2倾向于保留在状态2。 解码过程细节（需要会viterbi算法这个前提）：

P(1-> 1-> 1-> 1)= 0.4 x 0.45 x 0.5 = 0.09 ，  
P(2->2->2->2)= 0.2 X 0.3 X 0.3 = 0.018，  
P(1->2->1->2)= 0.6 X 0.2 X 0.5 = 0.06，  
P(1->1->2->2)= 0.4 X 0.55 X 0.3 = 0.066

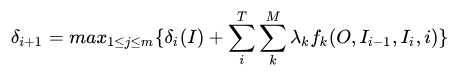
但是得到的最优的状态转换路径是1->1->1->1，为什么呢？因为状态2可以转换的状态比状态1要多，从而使转移概率降低,即MEMM倾向于选择拥有更少转移的状态。

**解释原因**

直接看MEMM公式：



 求和的作用在概率中是归一化，但是这里归一化放在了指数内部，管这叫local归一化。 来了，viterbi求解过程，是用dp的状态转移公式（MEMM的没展开，请参考CRF下面的公式），因为是局部归一化，所以MEMM的viterbi的转移公式的第二部分出现了问题，导致dp无法正确的递归到全局的最优。



**其它说明：**

最大熵模型是在已知经验分布的基础上求解有关特征函数f(x,y)的最优的P(y|x)概率分布，但它的随机变量y有相互独立的假设，所以不能很好的描述y\_i, x\_i与y\_{i-1}的关系，而HMM又有观测独立性假设不能自由的选择特征，所以我们希望找到一个能同时服从马尔可夫性假设和服从最大熵假设的模型解决序列标注的问题。

**模型形式**

对比隐马尔可夫模型(HMM)


P(X)=\sum_y \prod_{i=1}^T p(y_i|y_{i-1})p(x_i|y_i)
  
状态序列Y，观测序列X，两个状态转移概率: 从y\_{i-1}到y\_i的条件概率分布 p(y_i|y_{i-1})，状态y\_i的输出观测概率 p(x_i|y_i)，初始概率p_0(y)。隐马尔可夫模型依赖于已知数据的概率分布，已经历史经验来决定现实决策，但实际能提供训练的数据是少量且稀疏的，我们不能枚举所有的数据分布状况，所以需要**在数据稀疏的条件下估计未知x,y的条件概率。**

最大熵马尔可夫模型(MEMM)


p_{y_{i-1}}(y_i|x_i)=\frac{1}{Z_{\tilde{\lambda}}(x_i,y_{i-1})}exp(\sum_{a}^m \lambda_a f_a(x_i,y_i)),i=1,2,...,T


用p(y_i| y_{i-1},x_i)分布来替代HMM中的两个条件概率分布，它表示从先前状态，在观测值下得到当前状态的概率，即根据前一状态和当前观测预测当前状态。每个这样的分布函数p_{y_{i-1}}(y_i|x_i)都是一个服从最大熵的指数模型。

[](http://wiki.swarma.net/index.php/File:Maxent-03.png)

左边是hmm模型，右边是memm模型

HMM 双状态转移条件概率的输出独立性假设，在训练过程通过统计x,y,y_{i-1}的联合概率p(y_{i-1},y),p(y,x)，在解码中通过联合概率反向递推p(y_i|y_{i-1},x)的条件概率

MEMM 限定条件求最优条件概率分布，在训练过程中收敛一组有关特征函数f_i(x,y)的参数向量\lambda_i，以及有关x_i与y_{i-1}的正则化因子，在解码过程中直接求得条件概率p(y_i| y_{i-1},x_i)

**优势与局限**

隐马尔可夫模型因为其严格的观测独立性假设，对于具有多个互相作用的观测状态组合以及中长范围的元素依赖的数据环境识别特征的能力有限，而最大熵模型特征选择灵活，且不需要额外的**独立性假设**或内在约束。

隐马尔可夫是生成模型，学习的是联合概率，必须列举所有观察序列的可能值，而最大熵马尔可夫模型可以在**不完整**信息下有推导出未知数据的能力来。

最大熵马尔可夫模型本质是有向概率图模型，即每个y\_i只依赖于x\_i，并且每个component(y\_i, y\_{i-1}, x\_i)(特征组合)之间是独立且做**局部归一**，可能存在**label bias(标记偏置)**的问题(概率很低)。