本文主要讨论标注问题及常用的模型，包括hmm，memm，crf等。标注算法是一个统计学习方法，其模型的形式为一个条件概率模型P(B|A)，其中A为所有可能的输入序列，B为所有可能的输出序列。hmm，memm，crf是用于解决标注问题的模型，所以它们的概率模型归根结底也是关于输入序列A和输出序列B的条件概率模型P(B|A)。

# 最大熵模型ME

最大熵模型是一个工具，并不是一种标注算法。但凡是通过训练样例集合来计算条件概率模型P(Y|X)都可以用最大熵模型来计算。这里只简单的介绍，如果想了解其具体运作的原理请参考上一篇文章。本节主要将最大熵模型的学习方法给步骤化。

## 1.1概率模型

最大熵模型的目标模型是一个条件概率模型P(Y|X)，其中X，Y分别是定义在输入空间和输出空间的随机变量。标识在输入X的条件下，输出Y发生的条件概率。



知道了最大熵模型的概率模型，下面就考虑如何学习这个模型，而学习模型之前首先要构造训练样本集合。

## 1.2训练样本集合

知道了待求的模型P(Y|X)，下面需要了解用什么格式的样本数据训练这个模型。定义训练样本集合S格式，每一个训练样本是一个输入和输出pair：



## 1.3经验概率

根据上面给出的训练样本集合S来训练最大熵模型，在此之前我们需要两个先验概率以备后续迭代算法使用：

1.输入X的边缘概率的先验概率P~(x)：



2.输入x和输出y的联合概率的先验概率P~(xy)：



这两个先验概率的作用上一篇文章中有介绍，用于构造最大熵模型的约束条件。

## 1.4特征函数

最大熵模型需要为输入和输出建立特征，每对在样例S集合中出现过的输入x和输出y都会对应一个特征函数，特征函数如下：



特征函数的意义在上一篇文章中有介绍，每个特征函数对应一个约束条件。更为关键的一点是有了特征函数后，就可以写出概率模型关于特征函数的表达式。

## 1.5模型公式

上一篇文章中详细介绍了这个公式是怎么推导来的，其实大家使用的时候并不需要知道公式怎么得来的，只需要记住它使用它就足够了。根据公式1.5定义的特征函数得到模型P(Y|X)计算公式为：





其中Z(x)为归一化因子，其保证对于任意一个x，所有的标记值y的条件概率P(y|x)之和为1。虽然公式的推导在这里不再重提，但可以尝试解释下这个公式的含义。这个看起来很可怕的公式实质上很简单，可以简化如下：





对于在样本集合中出现过的xy对，都会分配一个正数λi（上一篇文章中介绍的拉格朗日乘子），其归一化的权重就是e的λi次方，概率就是(e^λi)/Z(X)；而样本集合中没出现过的xy对，归一化权重就是一个保底的1，概率就是保底概率1/Z(X)。这样通过迭代算法不断调整λi的值，直到取得训练样本集合的极大似然值。

## 1.6训练算法

目标概率模型为条件概率P(Y|X)，样本集合为S，可根据样本集合S计算出先验概率P~(XY)和P~(X)；设计特征函数f(x,y)，为每个特征函数分配一个λi，可得出目标概率模型的计算公式如公式1.6和1.7，然后通过迭代算法计算出λi的值，从而可得出目标概率模型。

输入：经验概率P~(XY), P~(X)，参数λi，及每个特征函数关于P~(XY)的期望E~(fi(x,y))

输出：概率模型P(Y|X)

算法步骤：

* 1. 初始化操作：根据经验概率计算每个特征函数fi的经验期望E~(fi)，初始化拉格朗日乘子λi=0。
  2. 根据当前拉格朗日乘子λi计算概率模型P(Y|X)
  3. 计算每个特征函数fi关于模型P(Y|X)和P~(X)的期望E(fi)， 并通过以下公式调整λi



* 1. 若所有λi都收敛，则结束，否则跳至步骤2

## 1.7小结

经过本节的介绍我们可以将最大熵模型的学习分为固定的几个步骤：1. 确定输入X和输出Y的可能取值，这样就确定了模型P(Y|X)；2. 构造训练样本集合S，其中的每个样本是一个xy对；3.根据训练样本集合S计算输入的边缘概率的经验概率P~(x)和输入输出的联合概率的经验概率P~(xy)；4.为输入输出构造特征函数；5.根据特征函数写出概率模型的表达式；6.根据迭代算法求出最终的概率模型即为最大熵模型。

# 隐马尔科夫模型HMM

本节主要介绍隐马尔科夫模型的概率模型的学习，很多资料上有介绍HMM的两种学习方式：1. 直接通过训练样本计算经验概率（在第一篇文章“词性标注和隐马尔科夫中使用了这种方式”）；2.通过EM算法计算概率模型。本文不再介绍这两种，本文用最大熵方式计算经过条件独立简化后的模型。

## 2.1模型

前文所述：标注算法是从观测序列到标记序列映射的条件概率模型，因此可以得出隐马尔科夫模型的概率模型为：



其中A为输入序列也就是观察序列a1a2…an，而B为输出序列也就是标注序列b1b2…bn。但隐马尔科夫模型是一个生成模型，生成模型就是先求出输入输出的联合概率P(AB)，然后通过条件概率公式由联合概率P(AB)求出条件概率P(B|A)，条件概率公式为：



而对于给定的观测序列A，其边缘概率P(A)是固定的，因此公式2.1可以转变为如下公式：



这样目标概率模型就转变为求输入输出的联合概率模型P(AB)。

## 2.2模型复杂度

上一小节将条件概率概率模型转变为求联合概率P(AB)，那么这个概率模型的复杂度是多少，我们来计算一下理论上存在多少种AB的组合：假设观测序列A的长度固定为m，观测集合大小为n，标注集合大小为n1，那么AB组合的可能性数目为m^n\*m^n1，这个数字太大了。更何况对于词性标注来说n的大小有上百万之多，而且观测序列的长度也是不固定的，理论上一个词序列可以无限长。也就是说AB组合的可能性有无数种，那么要计算这个概率模型是不可能的，因为在一个无限的样本空间中选中任何一个有限子集的概率为0，也就是说无论给定的训练样例集合多大都无法表达这个概率模型。

## 2.3独立假设

上一小节中介绍了模型P(B|A)是复杂到无法学习的，因此需要一些独立假设来简化这个模型，准确点应该说正是这些独立假设的存在，该模型才被称为是隐马尔科夫模型：

1. 齐次马尔科夫假设：下一时刻的标记值仅与当前时刻标记值相关，与其他信息无关



1. 观测独立假设：某位置观测值仅与相同位置的标记值相关，与其他信息无关



有了上述两个假设的从观测序列到标记序列映射的条件概率模型P(B|A)才叫做隐马尔科夫模型，而这个条件概率模型P(B|A)在前文中已经被转化为联合概率模型P(AB)。

## 2.4模型简化

利用上面的独立假设就可以对上述概率模型进行简化运算，其中每一步化简都是用到了独立假设和条件概率公式P(B|A)=P(AB)/P(A)及该公式的变型：











简化后模型为：



经过条件独立假设后原概率模型转变为两个条件概率的乘积，这大大简化了原有模型的复杂度，我们可以通过计算这两个条件概率模型，然后通过公式（2.7）计算整体联合概率。这两个概率模型为：

1. 标记值之间的转移条件概率模型：



2. 标记值到观测值的输出条件概率模型：



## 2.5样例集合

定义训练样本集合S如下，每个训练样本是一个三元组，包含一个观测状态ai，标记状态bi和 bj；每个三元组<ai,bj,bk>解读为：ai被标记为bj一次，bj由bk转移得来一次。



有了训练样本集合我们可以根据第一节介绍的学习最大熵模型的方法来分别计算条件概率模型P(bi|bj)和条件概率模型P(ai|bj)的最大熵模型。

## 2.6经验概率

求概率模型P(ai|bj)的输入b和输出a的联合概率的经验概率P~(ab)：



求概率模型P(ai|bj)的输入b的边缘概率的经验概率P~(b)：



求概率模型P(bj|bi)输入bi和输出bj的联合概率的经验概率P~(bibj)：



求概率模型P(bj|bi)的输入bi的边缘概率的经验概率P~(bi)：



## 2.7特征函数

输出概率模型P(a|b)的输入输出对应的特征函数：



转移概率模型P(bj|bi)的输入输出对应的特征函数：



## 2.8模型公式

根据1.5节提供的模型公式和前面给出的特征函数，我们可以给出输出概率模型和转移概率模型的计算公式如下：

转移概率模型的表达式和归一化公式：





输出概率模型的表达式和归一化公式：





## 2.9模型训练和使用

模型的训练就是1.6节介绍的最大熵模型的训练方式；模型的使用就是用viterbi算法根据观测序列算出概率最大的标注序列。

# 最大熵马尔科夫MEMM

前文讲过标注算法的模型是在输入A已知的条件下输出B的条件概率模型P(B|A)，最大熵马尔科夫是标注算法，所以最大熵马尔科夫也是条件概率模型。

## 3.1模型



其中A为所有可能的输入序列a1a2…an；B为所有可能的输出序列b1b2…bn。

## 3.2模型复杂度

由2.2小节的分析得出P(B|A)在没有独立假设的情况下是复杂到无法计算的，因此MEMM模型也是在一定条件的独立假设情况下的观测序列到标记序列的条件概率模型。

## 3.3独立假设

1．标记序列独立假设：每个标记值不仅和前一个标记相关，还和相同位置的观测值相关，和除此之外的其他信息均无关：



2. 观测序列独立假设：每个观测值只和相同位置的标记值相关，和其他信息均无关：



有了上述两个假设的从观测序列到标记序列映射的条件概率模型P(B|A)才叫做最大熵马尔科夫模型，而这个条件概率模型已经被我们转化为联合概率模型P(AB)。

## 3.4模型简化

利用上面的独立假设就可以对上述概率模型进行简化运算，其中每一步化简都是用到了独立假设和条件概率公式P(B|A)=P(AB)/P(A)及该公式的变型：











简化后模型为：



经过条件独立假设后原概率模型转变为一个条件概率的乘积，这大大简化了原有模型的复杂度，我们可以通过计算这个简化后的条件概率，然后通过公式（3.5）计算整体联合概率。这个条件概率模型为：



其中模型的输入为ai，bk，模型的输出为bj。

## 3.5样例集合

根据公式3.6所得的要计算的条件概率公式，我们设计训练样本格式如下：



## 3.6经验概率

计算输入ai，bk和输出bj和联合概率的经验概率如下：



计算输入ai，bk的边缘概率的经验概率如下：



## 3.7特征函数

定义关于输入ai，bk和输出bj的特征函数如下：



## 3.8模型公式

根据1.5节提供的模型公式和前面给出的特征函数，我们可以给出概率模型的表达式和归一化公式如公式3.11和3.12所示：





## 3.9模型训练和使用

模型的训练就是1.6节介绍的最大熵模型的训练方式；模型的使用就是用viterbi算法根据观测序列算出概率最大的标注序列。

# 条件随机场CRF

前文讲过标注算法的模型是在输入A已知的条件下输出B的条件概率模型P(B|A)，条件随机场是标注算法，所以最大熵马尔科夫也是条件概率模型。

## 4.1模型



其中A为所有可能的输入序列a1a2…an；B为所有可能的输出序列b1b2…bn。

## 4.2模型复杂度

由2.2小节的分析得出P(B|A)在没有独立假设的情况下是复杂到无法计算的，因此crf模型也是在一定条件的独立假设情况下的观测序列到标记序列的条件概率模型。

## 4.3独立假设

1. 标记序列独立假设：每个标记值和整个观测序列相关及相同位置的观测值以及该位置相关：



2. 观测序列独立假设：某序列a的某个位置观测值ai和该序列及对应位置标记值bi以及位置i相关：



## 4.4模型简化

利用上面的独立假设就可以对上述概率模型进行简化运算，其中每一步化简都是用到了独立假设和条件概率公式P(B|A)=P(AB)/P(A)及该公式的变型：









模型简化为：



经过条件独立假设后原概率模型转变为一个条件概率的乘积，这大大简化了原有模型的复杂度，我们可以通过计算这个简化后的条件概率，然后通过公式（4.5）计算整体联合概率。这个条件概率模型为：



其中模型的输入为bj，a1…an，pos；模型的输出为bi。

## 4.5样例集合

根据公式4.6所得的要计算的条件概率公式，我们设计训练样本格式如下：



## 4.6经验概率

输入输出的联合概率的先验概率为：



输入的边缘概率的先验概率为：



## 4.7特征函数

样本<bi,bj,a1…an,pos>表示观测序列a1…an在pos位置的观测值被标注为bi，同时前一个位置的观测值被标注为bj。定义关于输入bj，a1…an，pos和输出bi的转移特征函数如下：



定义关于标记值到观测值的输出函数如下：



## 4.8概率公式

根据1.3节提供的模型公式和前面给出的特征函数，我们可以给出概率模型的表达式和归一化公式如公式4.12和4.13所示：





根据公式4.5带入上述公式得到标注算法的概率模型P(B|A)为：





而这个公式正式李航老师的统计学习中给出的公式。

## 4.9模型训练和使用

模型的训练就是1.6节介绍的最大熵模型的训练方式；模型的使用就是用viterbi算法根据观测序列算出概率最大的标注序列。

# 总结

个人理解无论是HMM，MEMM，CRF归根结底都是标注算法，而标注算法的概率模型为P(B|A)，其中A为所有可能的输入序列，B为所有可能的输出序列；而区分这三种算法的就是它们的独立假设程度上的不同。三个算法里HMM的独立假设最多，CRF的独立假设最少。独立假设一方面简化了模型，另一方面丢失了一部分关联信息。独立假设越多模型越简单，计算量越小；而同时丢失的信息就越多，计算出来的模型就越不能反映原始模型的特性。