方骏-2020年5月10日-读书报告

1. 自己提出的问题的理解：
2. 提出的问题1：隐式马尔科夫的具体应用问题是什么？

讨论后的理解：在模式识别上和生物信息序列上都可以用到。在自然语言处理上也可以用到，比如词性标注的问题。已知词序列和模型的情况下，找到最优的词性序列，就是隐式马尔可夫的标注问题。在生物信息上，比如得知了蛋白质的序列，我们可以通过马尔可夫模型在数据集上的训练得到的结果，来预测蛋白质的结构等。

1. 别人提出的问题的理解：
2. 问题2：近似算法和维特比算法的区别（在应用场景上、有什么优劣）。

自己的理解：近似算法是一个很想当然的算法，就是不断地找局部最优解然后拼凑起来，这个很显然得到地路径从整体上来说不一定是最优解。所以近似算法不一定能够得到全局最优解，而维特比算法利用动态规划，可以保证得到最优解，但是这种算法计算量有些大，可以用于一些效性小型场景中，对于精度要求不高的完全可以用近似算法。

1. 问题3：能否给出一个具体的示例来理解隐马尔可夫具体是如何解决概率计算问题、学习问题、预测问题的，分别是完成了什么样的任务。

自己的理解：就拿书上所说的投硬币的例子。概率计算问题就是给定了HMM问题，观察到给定结果的概率是多少，用的是前向或后向算法，就是直接用概率值计算；学习问题就是在模型参数都不知道的情况下，如何根据观察结果得到这些参数，这个用EM算法就可以得到；预测问题就是给定观察结果和模型，找到最佳的状态序列，即盒子序列的问题，这个维特比算法就可以解决，利用动态规划一点点算。

1. 问题4：维特比算法中变量δ的计算和前向算法中α的计算有什么区别？

自己的理解： 维特比算法中变量δ指的是在时刻t状态为i的所有单个路径中概率最大值，而前向算法中α指的是到时刻t部分观测序列为，，，…，且状态为的概率。一个是找到最大值，一个只是一个概率，这两个是完全不一样的。

1. 读书计划

1、本周完成的内容章节：《统计机器学习》第十章

2、下周计划：《统计机器学习》第十一章

四、读书摘要及理解

1、隐马尔可夫模型可以用于标注问题。这是一个关于时序的概率模型，描述由一个隐藏的马尔可夫链随机生成不可观测的状态随机序列，再由各个状态生成一个观测而产生观测随机序列的过程。隐藏的马尔可夫链随机生成的状态的序列，称为状态序列：每个状态产生一个观测，由此产生的观测的随机序列，称为观测序列。

2、Q是所有状态的集合，V是所有可能的观测的集合：



I是长度为T的状态序列，O是对应的观测序列：



A是状态转移概率矩阵：



其中：



B是观测概率矩阵：



其中：



是初始状态概率向量：



其中：



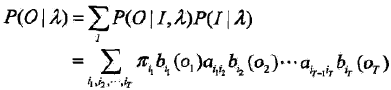
所以隐马尔可夫模型*λ*可以用三元符号表示，即

可以称为隐马尔可夫模型的三要素。

3、观测序列的生成遵循步骤：按照初始状态分布产生状态，然后从时刻1开始，按照状态的观测概率分布生成观测数据，按照该状态的状态转移概率分布再产生状态，最后已知增加时间循环，直到终止。

值得注意的是，隐马尔可夫有3个基本问题：概率计算问题、学习问题和预测问题。这个在问题中有过阐述。

4、概率计算方法可以采取直接计算法，方法非常直接，就是按照概率列举所有ke能长度为T的状态序列，然后求各个状态序列与观测序列的联合概率，然后求和：



这个方法计算量非常大，所以一般不采用这个方法。

前向算法比较简单，这里定义前向概率。给定隐马尔可夫模型，定义到时刻t部分观测序列为且状态为的前向概率，记作：

先计算初值：



递推：



终止：



后向算法先定义后向概率。给定隐马尔可夫模型，定义到时刻t状态为的条件下，从t+1到T的部分观测序列为的概率为后向概率，记作：

令初值为：



对

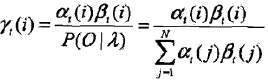


最后得到：

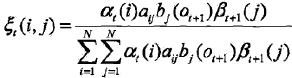


5、下面是一些相关概率的计算公式：

给定模型和观测，在时刻t处于状态的概率：

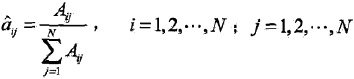


给定模型和观测，在时刻t处于状态且在时刻t+1处于状态的概率：

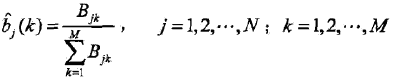


6、学习算法分成监督学习算法和非监督学习算法，就是EM算法。监督学习算法就是最大似然估计转移概率和观测概率以及初始状态概率，即为：

转移概率估计：

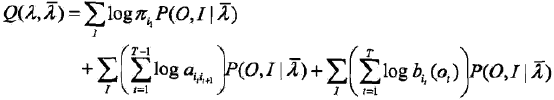


观测概率的估计：

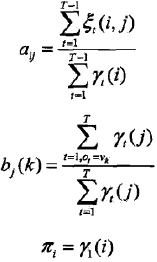


初始状态概率就是S个样本中初始状态为的频率。

Baum-Welch算法就是EM算法，状态数据就是隐变量。最大化完全数据的对数似然函数就是。EM算法的E步，求Q函数为：



EM算法的M步，极大化Q函数的各项即可，可以得到相关结果为：



根据上述公式进行递推，就可以得到最终的最优参数。

7、近似算法就是最基础的想法，在每个时刻t选择在该时刻最有可能出现的状态。从而得到一个状态序列，通过计算：

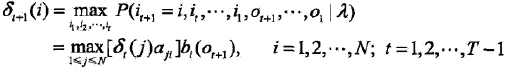


即可得到所有的状态序列。这个算法虽然简单，但是得到的序列从整体上来说未必就是最优状态序列。在某些精度要求不高的应用场景中，完全可以用这种方法。

维特比算法是利用动态规划，先求路径的子集的最优路径，然后一点点扩大路径，求得最优的路径。定义在时刻t状态为i的所有单个路径中概率最大值为：



可得：



以及在时刻t状态为i的所有单个路径中概率最大的路径的第t-1个结点为：



维特比算法先初始化，再一步步利用上述定义式进行递推，最终求得最大的和相应的最优路径点，再利用公式进行回溯，即可求得最优路径。