**统计学习方法读报告**

09017244 郑健雄

**本周进度:**统计学习方法十四章，负责10.3.2-10.3.3讲解

**下周进度:**统计学习方法十五章到十六章。

1. **问题与解答**

1.(我提出) P202页(10.22)如何推导？

讨论结果:该公式的推导并不能直接根据之前的两种形式得到，而是进行定义带入化简得方法进行求解，此处还发现t得取值其实是任意的，最终得概率都是一样的。

2.(我提出) 维持比算法动态规划原理的细节是什么？逆推的方式为什么可以得出最优的路径？

讨论结果:维持比算法的优势在于它极大的简化了计算的复杂性，每一步骤，对于每一个节点，只需要在N个数据中比大小，而又要在N个点里面比大小，通过记录的方式，最终只要得出最优路径，它就不用思考路径本身的整体格式，而是记录前一个路径的位置即可，因为前一个路径点一定是最优，不然就存在更优路径，而通过这样的方法就可以逆推出最优路径。

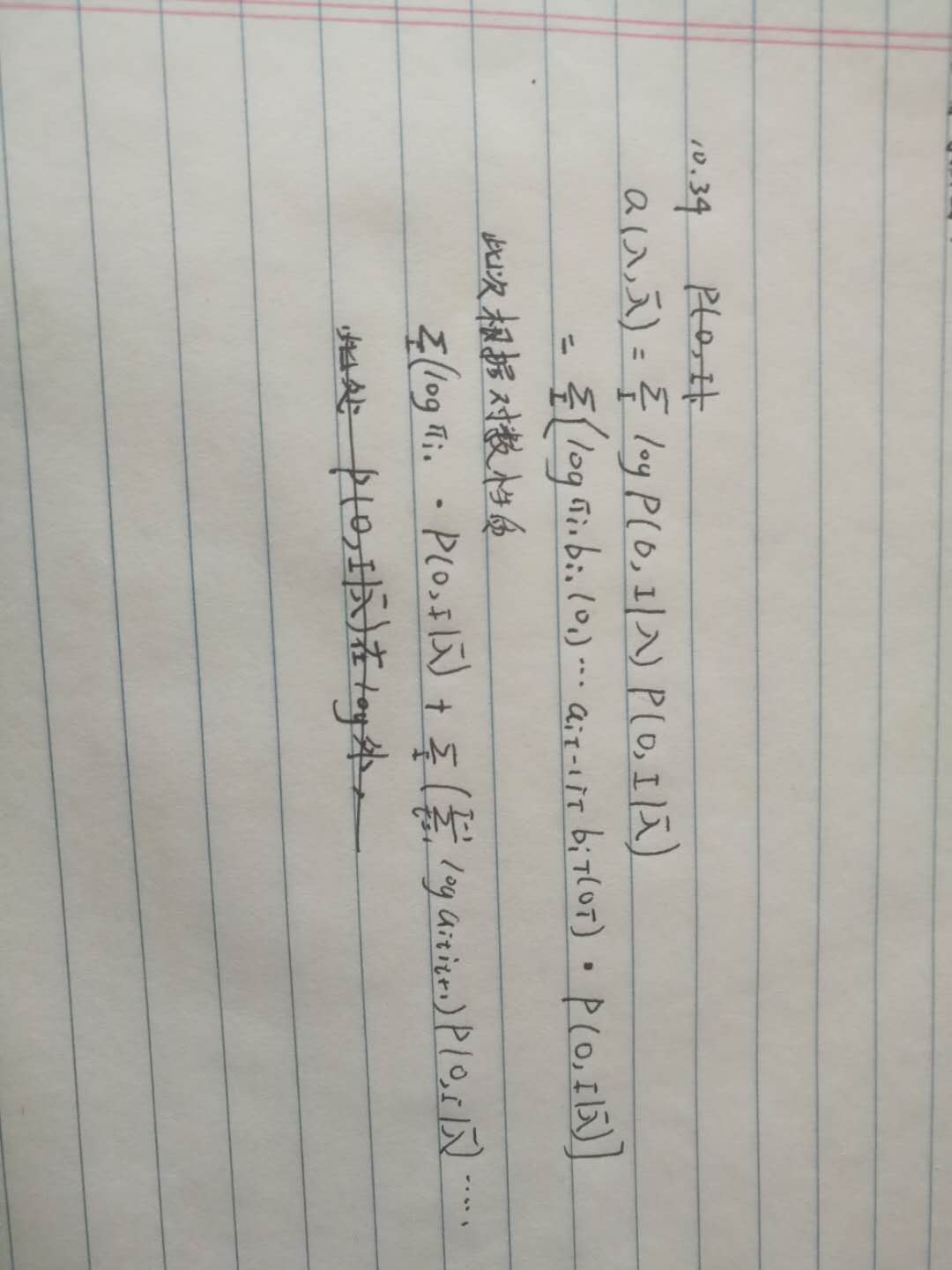
3.(我提出) 维持比算法中变量delta的计算和前向算法中变量a的计算有什么区别？

讨论结果:前向算法中a的计算属于单纯的计算概率并且进行求和，但是维持比算法则是每一步对概率值进行极大化的选择，它们的过程类似，但是方法不太一致。

4.(别人提出) 10.34的公式如何推导得出？

讨论结果:

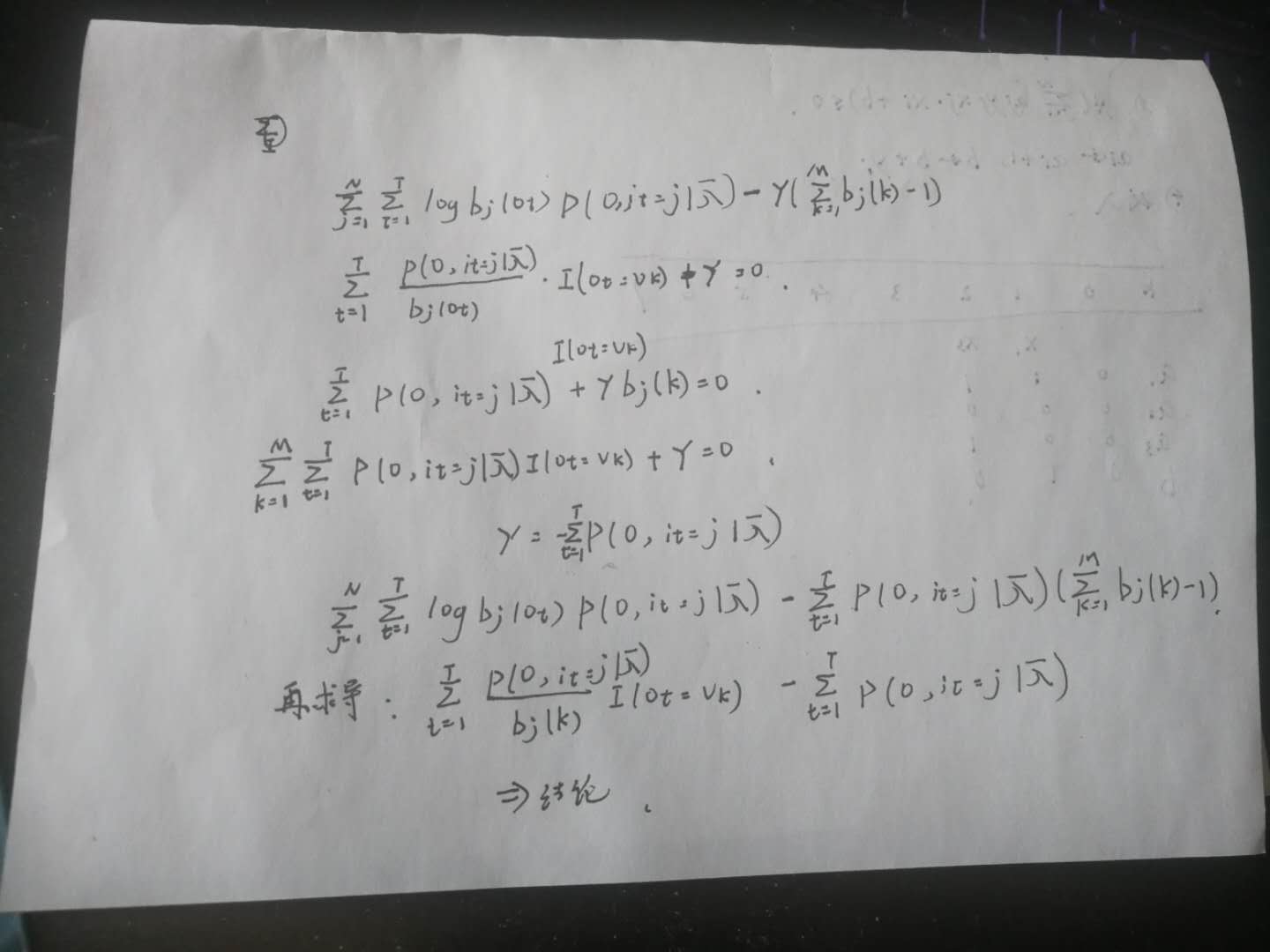
推导过程如下:



5.(别人提出) P206（3）部分的推导，主要是不太明白加入I()怎么推导的吗？

讨论结果:

推导过程如下:



6.(别人提出) 10.3节学习算法分成监督学习和非监督学习（Baum-Welch算法）。可这两种算法的区别仅在于数据上一个是有人工标注的数据和无人工标注的数据。但这区分监督与非监督的标准应该是是否需要生成类。

讨论结果:实际上类本身就是人为规定的一种标签，聚类方法或者人为标注以及现在的马尔可夫模型都是人们寻找类的一种形式，不同的模型表现出人们对于最终产生类的不同预判，来选择合适的模型进行模拟。而该模型假设了隐藏变量的存在，使用了EM算法，所以可以理解为是无监督学习方法。

7.(别人提出) P198直接计算方法以及前向、后向算法的时间复杂度，如何得到的？

讨论结果:直接计算是一种排列组合的思想，总共存在N的T次方种路径，而每条路径计算需要T的复杂度，所以直接计算的复杂度可以得出。而前向算法则是每一步的计算复杂度为N平方，而总共的步数为T,所以计算复杂度也可以求出，而后向是类似于前向计算的，所以复杂度也可以这样得出。

**二、 下周计划安排**

下周计划阅读完第十五章和第十六章。

**三、 读书收获**

读书摘要:

1. 隐式马尔可夫模型是一个时序模型，它本身要考虑数据的顺序，也是一个生成模型。因为观测数据本身是具有随机性的，不是很好预测，所以该模型假设实际上存在一个隐藏状态序列在控制着观测序列的生成，而隐藏状态的形式比较简单。我们的目标就是求出该隐藏序列，并通过这个序列生成观测序列。其包含初始状态向量，状态转移概率矩阵以及观测概率矩阵来控制整个模型。
2. 马尔可夫模型有三种基本问题要进行研究，

概率计算问题:给定模型参数以及观测序列O,求O的概率。

学习问题:已知O，估计模型参数，使得概率最大，使用了极大似然的思路。

预测问题:已知O和模型参数，预测状态序列I,该部分的方法可以用于数据标注上，状态序列就是一种标注。

1. 概率计算问题:可以根据定义来直接计算，但是过于复杂，一般通过前向-后向算法实现数据的重用，避免重复计算。
2. 学习问题:一般如果数据完备，就使用近似算法来进行估计。但一般情况下隐藏状态是缺失的，所以使用基于EM算法的Baum-Welch算法来进行实践，其对隐藏变量进行了一个估计，可以很好的训练模型。
3. 预测问题:处理该问题使用了维持比算法，其使用了动态规划和极大化的思路，从而很好的提高了计算的效率。其比近似算法更合理也更有效。