读书报告内容：（复制下面部分到一个新文档，名称命名为自己的名字，填写完交给组长，没做第四部分可以不复制第四部分）

1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：co-Training算法中，假如在某一轮f1f2将分类结果加入L的时候，某一个元素在f1中被标为正标记，又在f2被选中标为负标记，那算法会如何应对这种冲突呢？

讨论后的理解：我想这是co-Training算法的弊端之一吧，计算机不知道这两条数据是一样的，加入L后只会依据算法流程和标记进行训练，这也是误差产生的原因之一

1. 问题2：co-Training算法同时使用两个分类器的意义是什么？

讨论后的理解：这是协同训练的算法，一个分类器训练出的结果作为另一个的训练集。在每次循环中，它首先从已标数据集L的每个特征集中学习得到一个分类器，然后利用这个分类器去分类（或标注）U中的无标注数据。将U中标为类Ci的一定数据（Ni），分类可信度高的数据加入到L中。这个过程一直进行到U成为空集（或达到一定的循环次数）。

二、（必填）别人提出的问题的理解（选择几个问题罗列，并给出理解）：

1. 问题3：如何选择半监督算法，有哪些可借鉴的原则？

自己的理解：首先，我们使用由LU学习算法得到的分类器来对无标注数据或测试集进行分类，然后人工评估一些随机采样的标注好的数据来估计分类的精度。如果分类有足够的精度，则得到的分类器可以被使用否则，还需要进一步改进算法。

1. 问题4：如何检验是否满足P126的两个假设？

自己的理解：所有的方法都做了一些很强的假设，当在满足这些假设的应用数据集上工作时，无标注数据是有帮助的。而当这些假设不被满足时，无标注数据可能会对学习产生负面的影响。但是，事先用一种自动的方法判定问题的结构至今还未解决。

1. 问题5：Em算法能够保证必定收敛嘛？

自己的理解：Em算法可以保证收敛到一个稳定点，但是却不能保证收敛到全局的最大点，因此它是局部最优的算法。

三、读书计划

1、本周完成的内容章节：如5.1.1-5.1.6

2、下周计划：《统计学习方法》第一章

四、（选做）读书摘要及理解或伪代码的具体实现（读书摘要、伪代码的具体实现代码等可以写到这个部分）

5.1部分监督学习

半监督学习方法可以对同时含有已标记的和未标记的数据集进行聚类，然后通过聚类结果中，每一类中所含有的任何一个已标记数据实例来确定该聚类全体的标签。  
更进一步的，考虑聚类算法可能存在误差。我们对聚类结果（某个分类）中未标记的的标记需要更加慎重。这个时候可以考虑表决法，根据聚类结果中某已标记分类的样本多少来决定是否“感染”，也就是否未未标记样本进行标记。而处在聚类边缘的样本，可以考虑通过其他算法完成标记。这样可以保证情报收集过程中小样本标记工作更高的准确性。

5.11朴素贝叶斯EM算法

EM算法是一种在存在不完整数据情况下广泛使用的进行最大似然估计的迭代算法。  
EM算法分两个步骤组成：  
期望过程（Exception Step,E-step）：使用现有的参数估计对数据的不完整部分进行填充。  
最大化过程（Maximization Step,M-step）：各个参数被重新计算。  
进入下一个循环，直到模型的参数不再发生变化时，EM算法收敛到一个局部最优点。

5.12Co-training

思想如下：首先我们需要根据已标记数据的两组不同特征来划分出不同的两个数据集，然后根据这两个不同的数据集分别训练出两个分类器。每个分类器用于无标识的数据集的分类，并且给出分类可信程度的概率值。

事实上，不同分类器可以给出不同的概率值。概率值高低分类加入数据集，然后进而影响每个判别器的重新生成，逐步提升判别器泛化能力。

5.13自学习

思想如下：首先训练带有标记的数据（这一步也可以理解为监督训练），得到一个分类器。然后我们就可以使用这个分类器对未标识的数据进行分类。根据分类结果，我们将可信程度较高的未标记数据及其预测标记加入训练集，扩充训练集规模，重新学习以得到新的分类器。

5.14直推式支持向量机

半监督支持向量机（Semi-supervised support vector machines,S3VM或S3VMs）是由直推学习支持向量机（Transductive Support Vector Machines, TSVM）变化而来。S3VM算法同时使用带有标记和不带标记的数据来寻找一个拥有最大类间距的分类面

5.15基于图的方法

基于图论的方法思想：首先从训练样本中构建图。图的顶点是已标记或者未标记的训练样本。两个顶点xi，xj之间的无向边表示两个样本的相似性，又称两个样本的相似性度量。根据图中的度量关系和相似程度，构造k-聚类图。然后再根据已标记的数据信息去标记为未标记数据。