1. **自己提出的问题的理解：**

问题1：对于自学习算法，本身具有一定的分类能力，但是通过这种已经存在的分类能力去增加自己的训练集规模，如果分类错误的话，不就会产生“一错再错”的问题吗？

讨论后的理解：因为自学习算法的限制比较小，所以不可避免的会有比其他算法更加多的问题，这种“一错再错”的问题也会出现。如果把置信度阈值设置的过低的话，则会出现把未标注的数据分类到和未标注数据的正确种类不同的分类中，之后再用这些错误分类的数据去训练分类器的话，就会使得分类器越来越“错”。

问题2：监督训练中的svm能够保证最后将两个种类的数据分开，但是SVM使用已经标注的正反例来产生决策边界，可能会产生未标注的点出现在决策边界上的情况，这种情况要如何处理？

讨论后的理解：因为在半监督学习中的SVM的训练是使用的标注好的数据，对于未标注的数据则先不关心，先用标注的数据训练好一个决策边界，然后将未标注的数据分类到两边，这个时候难免会出现未标注的数据出现在决策边界上而不是在某个分类的领域里，处理这种情况的方式也非常多，可以通过不考虑这些点或者随机分配一个类别，因为这种点占比非常非常小，所以对最后结果影响不大。

1. **别人提出的问题的理解：**

问题1：P128提到特征选择算法用来去除冗余的特征，其中的逻辑是什么？

自己的理解：在分类过程中，数据的某些特征对分类并没有做出什么贡献，比如我们将一个班级中的人分为男人和女人，那么对于每个人来说，性别是他们唯一重要的属性，我们需要唯一考虑的属性就是性别，考虑任何其他的属性比如年龄、身高、体重都是没必要的，对最后的分类结果都不会做出什么贡献。

2、问题2：co-Training算法同时使用两个分类器的意义是什么？

自己的理解：和普通分类器相同，半监督学习到的分类器同样会存在不稳定的情况，即两次训练出来的结果不同，如果同时考虑所有的特征来将未标注的数据点进行分类，那么一些特征会被某些非常相似的特征所覆盖，导致这些特征未能对整体分类做出贡献，如果将这些独立的特征分开，使用每个特征集合训练出来的分类器都能有非常高的准确率，此时可以通过投票或者将两个分类器相乘等方式来将两个分类器进行结合，就可以对那些不太稳定的数据点（即两次分类有可能有不同的类别）有比较稳定的结果，从而可以提高分类器的稳定性和准确性。

3、问题3：自学习算法因为自己训练自己，会不会出现严重的过拟合结果？

自己的理解：和我的问题1比较相似，是自学习算法的两个极端，即如果使用现有已标注数据训练出来的分类器去分类其他的未标注数据，如果置信度阈值设置非常高，那么最后分类出来的结果中每个类别的相似度都会非常高，然而因为初试数据的单一性，并不能考虑到全部的特征，那么后来被分类的数据训练出来的分类器就会出现这个错误累加的情况，最终出现过拟合。

1. 读书计划

1、本周完成的内容章节：第5章

2、下周计划：《统计机器学习》第一章

1. 读书摘要及理解或伪代码的具体实现

**(1)读书摘要：**

（1）LU算法需要确保每个类别都有已标注数据。

（2）LU算法非常像是在进行无监督的聚类，因为聚类算法是通过每个数据中的属性进行分类，其中的类别是预测出来的，而半监督更像是已经知道了分类，再进行聚类，这样无监督的数据会根据已经分好类的数据进行分类，会比简单的聚类算法更加准确和稳定。

（3）co-training 和之前的混合分布不同的一点在于cotrianing中各个属性都能将数据进行正确分类，而混合分布则将数据分入不同的类别。

（4）基于图的方法时间复杂度为O(V3),空间复杂度为O(V2)