**读书报告**

姓名：赖泽升

1. 自己提出的问题的理解（罗列全部）：

**1.提出的问题1：**5.1.1中通过leave-one-out在已标注数据集中交叉验证精度的方法选取μ是什么意思？

**讨论后的理解：**每次拿一个已标注的数据出来作为有效数据，然后把其他的已标注数据和无标注数据用这种方法训练，取不同的μ，然后再来验证准确度。

**2.提出的问题2：**5.1.2中为什么加入到L中的数据不是随机的话对学习没有什么帮助？

**讨论后的理解：**两个子集中每个特征互相有关的话，就和只训练一个就没有区别了。那就退化成self-training了。

别人提出的问题的理解（选择几个问题罗列，并给出理解）：

**问题3：** 在Co-Training中提到，用来训练分类器的两个特征的子集is sufficient for learning the target classification function。但是在训练之前我们是怎么得到这个结论的？还是说这只是一个assumption?

**自己的理解：** 这只要彼此之间无关（条件独立）即可，也可以按人的思维去思考哪些特征凑在一起更容易判定种类，然后人工选取。

**问题4：** 5.1.5节中，最小割方法的代价函数|vi-vj|代表的是什么？是两点之间的距离么？

**自己的理解：** 代表的是它们类别是否一样的标志。如果两个很相似的话，vi-vj就要为0。否则为1。

**问题5：** 5.1.1中种子的类别标识表示了聚类结果中各个簇的类别标识是什么意思呢？

**自己的理解：** 先对已标注的数据进行聚类，得到几个簇，然后再利用这些簇对无labeled的聚类。

**问题6：** 在5.1.2中，提到将U中标为类的一定数量、分类可信度高的数据加入到L中，请问这里的分类可信度是怎么衡量的，怎么理解？

**自己的理解：** 每一次都找出现的概率最高的那几个数据加入到L中。

读书计划：

1. 本周完成的内容章节：5.1和前面章节的回顾。
2. 下周计划：统计学习方法第一章。

读书摘要：

第五章讲的是半监督学习，第一节介绍了几种分类器和算法。分别有朴素贝叶斯分类器和EM算法，Co-training，自学习，直推式支持向量机，基于图的办法。半监督学习主要的思想是利用已标注数据加上未标注数据的辅助进行机器学习。这样既能增加学习效果，又能减少人的工作量。但同时，若想让无标注数据对学习起到促进作用，在进行半监督学习的算法时要满足一定条件，否则无标注数据会对学习产生负面影响。