**2019/03/01读书报告**

学号 71117228 姓名 李泓烨

读书进度：按要求读完web data mining

1. **读书报告内容**

**1. 自己提出问题的理解**

① 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？

**讨论后的理解**：其实vi和vj的取值范围只是{0, 1}，如果i和j要是很相似的话，vi – vj就要为0，也就是它们类别相同，才能使得这个代价函数取到最小值。类别不同，则考虑相似度，使得代价函数最小。

② Co-training的第二个假设，能否举一个通俗易懂的例子，我没有太看懂书上举的例子

**讨论后的理解**：组里的同学举的例子，是一个反例，如果用于训练的这两个特征，一个是身高，一个是年龄，那么这两个特征在一定程度上是有关系的，Co-Training的目的就无法达到

**2. 别人提出问题的理解**

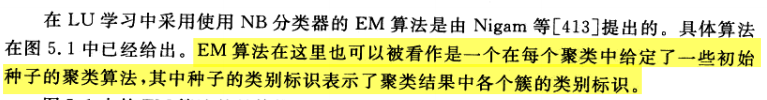
1. **提出的问题1**：Self-Training是不是会导致泛化能力不够强，因为它是用自己训练的结果来教自己，感觉会产生很大的误差？

**讨论后的理解**：其实这个算法最重要的，就是挑选最有把握的预测数据，因为没有假设（前提条件），是最简单的预测方法，所以其实效果相对不是很好，误差也相对较大

2. **提出的问题2**：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？

**小组成员的解释**：需要看训练目标，不同训练目标需要的特征数不同，只要彼此之间无关（条件独立）即可，当然我们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，这样子人工选取

3. **提出的问题3**：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？

**自己的理解**：过程（1）先对label的数据聚类，得到几个cluster。（2）再利用这些cluster对无label的数据聚类，这些label的数据就是种子

4. **提出的问题4**：Co-Training可否将属性分为三个集合训练三个分类器进行学习呢（或者更多）？

**自己的理解**：实际情况可能无法出现

5. **提出的问题5**：5.11中mixture model究竟指什么？意义何在？为什么要有这前提假设？

**自己的理解**：为了假设数据是同分布的，这样可以生成模型；如果数据没有一个分布的话，感觉就生成不了模型。朴素贝叶斯的目的是学习到生成数据的机制，也就是数据如何生成的，但是如果数据不符合一个model，它就学习不到，mixture model正式为了避免这种情况

6. **提出的问题6**：如何理解EM算法的两个假设？Mix-model以及one-one correspondence between mixture components and classes？

**自己的理解**：理解——每个混合成分对应唯一的类别

7. **提出的问题7**：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**：理解——可能是置信度，如果置信度高了，就说明概率高了

8. **提出的问题8**：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？

**自己的理解**：为了得到u的具体值，对不同的取值进行验证，看哪个取值的效果最好

**3. 读书计划**

① 本周完成的内容章节：看完了第5章

② 下周计划：统计机器学习，按照要求看一章

**4. 读书摘要、理解&伪代码的具体实现**

1. 读书摘要

这里我以思维导图的形式记录了我的读书笔记（见下一页）

