读书报告

71117203-袁佳怡

1. 自己的问题
2. 问题1：5.1.5中，计算代价函数使得其最小，这个怎么理解？是什么意思？

讨论后的理解：这是对算法的cost function,前面又说道w越相似越大，而等于0或1，如果两个很相似的话，就要为0，也就它们类别相同，这样才能min这个代价函数，相似度是根据函数已经得到了，然后要判断他们的类型，目标是使得这个代价函数最小。

1. 问题2：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？

讨论后的理解：在分类的每一次迭代出的结果中，选取概率高置信度高的，从U中加入到U中，这样做使得每次的加入都是相对更可靠的，也是为了能够提升算法的执行效率。

1. 别人的问题
2. 问题3：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

自己的理解：这是一个无基础的初始学习，主要是为了训练出分类器，具体质量得判断，因为可能本身labeled的数据很少，那么它的误差可能是非常大的，但这种算法主要取决于怎么选取无标签数据，主要是有选择数据输入决定。

1. 问题4：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：根据训练目标的不同，所需要的特征数也不同，在没有训练之前，选取彼此之间无关（条件独立的）特征即可，也可以按人的思维去思考那些特征凑在一起更容易判断种类，这样子去人工选取。

1. 问题5：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

自己的理解：举例是这样的，比如f1是根据年龄来训练，f2是根据身高来训练，因为年龄和身高之间可能存在正相关的关系，那么co-training的目的就达不到了，因为在co-training中要求两个特征之间尽量没有关系，是为了达成仿佛这两个之间是无关的，是两个随机集。退化成极端情况理解，如果两个子集中每个特征互相有关的话，那这和只训练一个没有区别了，退化成self-training了，因为大部分的数据都是没有标签的。

1. 问题6：请问一下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：首先对labeled的数据聚类，得到几个簇，然后再利用这些簇对无labeled的进行聚类,这些labeled的数据就是种子。

1. 问题7：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：如果能够满足兼容的假设，那么理论上是可以的，但在现实中，这样的假设很难成立。

1. 问题8：1.5.11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：这是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，可能就生成不了模型。朴素贝叶斯的目的是学习到生成数据的机制，但如果数据不符合一个model它就学习不到，有mixture model就是为了避免这种情况。

1. 问题9: 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：对不同的取值进行验证，看那个取值的效果最好，得到μ的具体值。每次拿一个labeled数据出来作为validation data然后把其他的labeled数据和unlabeled数据用这种方法训练，去不同的μ，然后再用那个一开始拿出来的验证准确度。

1. 学习计划

1、本周学习：第五章和统计学习第一章

2、下周学习计划：统计学习第二章

1. 数据挖掘第五章学习笔记

