第五次读书笔记

1. 自己提出的问题

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？

解答：朴素贝叶斯的目的是学习到生成数据的机制，如果有一个数据不符合模型，它就学习不到，因而需要混合模型以保证学习的完成。

2.co-training中，如果特征属性足够多，是否可以多分几个属性的子集，训练更多classifierf1，f2，f3……，来获得更准确的分类结果？如果是，那在这样的情况下，不同分类器的兼容要如何解决？

解答：很多人都提了这个问题。这理论上来说是能够做到的，然而实际上要考虑到，分的特征的自己越多，分类器越多，分类器之间的兼容就越难达成，因而如果能够达成多个分类器兼容自然最好，如果达不成就不强求增加分类器的数量。

1. 别人提出的问题

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。

解答：这是正常的。因为这是最早的半监督学习算法，没有假设，效果较差。我在讨论中也提出了一些想法尝试更多利用labeled data，如像评估分类器时取验证集一样，多次选取不同部分的labeled data，但都被指出了一些问题，因而还是认为要从利用数据量更多的unlabeled data着手。

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?

解答：有人认为这是单纯一个前提假设，我认为这是要视具体目标而决定。前提是两个特征自己之间无关（条件独立），剩下的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.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……

解答：这本质是为了达成仿佛这两个之间是无关的，是两个随机集。因而以买瓜为例，要挑选好瓜，一个特征是颜色，一个特征是品种。如果品种与颜色之间存在较强的决定关系，那这就不可行。极端一点，如果两个子集中每个特征互相有关的话，那这和只训练一个就没有区别了。

4.5.1.1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？

解答：在第四章中提到过类似的句子。第四章中提到过，可先对labeled data聚类，这些labeled data就是种子，label就是标识。

1. 读书进度

读书进度：读完了5.1，并对朴素贝叶斯与支持向量机做了额外的阅读

下周读书计划：统计机器学习第一章，视时间阅读第二章。

1. 读书摘要

5.14：支持向量机

原理：将训练集置身于样本空间，找到一个合适的划分超平面

如何找到一个合适的划分超平面：引入支持向量与间隔的概念，将两个异类支持向量到超平面的距离之和称为间隔（此处指硬间隔），而目的是找出一个具有最大间隔的划分超平面。由此可推导出间隔γ=2/||w||，根据数学推导，为使间隔最大，等价于最小化||w||^2。

再度引入概念：软间隔。由于未必存在一个超平面完全划分不同的类与过拟合可能性的存在，引入软间隔的概念，允许某些样本不满足约束。

Transductive Support Vector Machines：通过利用未被标注的数据，来找到划分超平面使间隔最大。

难点：由于计算的复杂性，将标签分配给未标记的例子，从而使分类器的结果边际最大化的问题已经不能有效、快速地解决了。

解决办法：先利用标签过的数据进行学习，用产生的classifier对未被标签过的数据分类，将最有可能是positive的标为positive，其他标为negative，多次迭代以优化软间隔成本函数。

问题与解决方法：Further improvements described in [259] include allowing the soft margin error of unlabeled examples to be penalized differently from the soft margin error of the labeled examples and penalizing the soft margin error on the positive unlabeled examples differently from the soft margin error on the negative unlabeled examples. The penalty on the unlabeled examples is also iteratively increased from a small value to the desired value.

5.15graph-based method

基本思想：将各个data point视作空间中的点，定义计算各点的边权重

方法：Mincut，Gaussian Fields，Spectral Graph Transducer

共同点：为使其得出的损失函数最小而优化。