## 2020.3.1 第五次读书报告

09018330 孙毅远

### 一、自己提出的问题的理解

本章概念较为抽象，实例较少，仅大概理解了一些理念，未能产生有效问题，会在之后的学习再回来复习

### 二、别人提出的问题的理解

#### 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

理解：可能会有误差，自学习这种方法，核心在于如何自动为数据产生标签，其能的高低，主要通过模型学出来的feature的质量来评价，可以通过不同数据的选择输入来调整达到需求

#### 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

理解：这应该是一个前提条件，需要对应具体目标来分类

#### 3. 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？

理解： 取值为0或者1，代表的是类别， 是判别其类别是否相同，如果不同再计算相似度

#### 4. 在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表示是什么意思

理解：先对标注数据进行聚类，然后得到几个簇，再根据这些数据对无标注数据进行聚类

#### 5. Co-training 能否将属性分为三个集合训练三个分类器进行学习呢？

理解：理论上应该可行，但实际操作可能无法实现

### 三、读书计划

#### 本周 统计学习方法学到2.1

#### 下周 学到第三章

### 四、读书收获

#### 1.1 统计学习

* 统计学习的对象是数据
* 统计学习的方法
  + 监督学习（supervised learning）
  + 非监督学习（unsupervised learning）
  + 半监督学习（semi-supervised learning）
  + 强化学习（reinforcement learning）

#### 1.2 监督学习

* 任务：学习一个模型，使模型能对任意输入，对其相应的输出做出一个和的预测。
* 输入空间，输出空间，特征空间（）
* 输入与输出看作定义在输入和输出空间上的随机变量的取值，一般用X,Y表示
* 输入实例 的特征向量记作
* 监督学习假设输入与输出的随机变量X和Y遵循联合概率分布
* 监督学习的模型可以是概率/非概率模型，由条件概率分布 或决策函数 表示，输出写作

#### 1.3 统计学习三要素

* 模型
* 模型的假设空间 () 包含所有可能的条件概率分布或决策函数。
* 假设空间用 表示，可以定义为决策函数的集合
* 也可以定义为条件概率的集合
* 策略
  + 损失函数和风险函数
  + 损失函数度量模型一次预测的好坏，风险函数度量平均意义下模型预测的好坏。
  + 损失函数是预测 和 的非负实值函数，记作
  + 常用损失函数
    - 0-1 损失函数
* 平方损失函数
* $L(Y,f(X))=(Y-f(X))^2$
  + 绝对损失函数
  + 对数损失函数
* 损失函数值越小，模型越好。由于模型的输入输出是随机变量，遵循联合分布，所以损失函数的期望是
* 即风险函数，对于给定的数据集，平均损失称为经验损失，记作

* 算法

#### 1.4 模型评估与模型选择

* 训练误差与测试误差
* 过拟合：过于追求提高对训练数据的预测能力，所选模型复杂度往往比真模型更高

#### 1.5 正则化与交叉验证

* 正则化是在经验风险上加一个正则化项，形式如下
* 交叉验证：简单交叉验证，S折交叉验证，留一交叉验证

#### 1.6 泛化能力

* 泛化误差，泛化误差上界

#### 1.7 生成模型与判别模型

#### 1.8 分类问题

* 精确率，召回率
* TP——将正类预测为正类数 FN——将正类预测为负类数
* FP——将负类预测为正类数 TN——将负类预测为负类数
* 精确率定义为
* 召回率定义为

#### 1.9 标注问题

#### 1.10 回归问题

#### 2 感知机

#### 2.1 感知机模型

* 定义
* 输入空间 ，输出空间 ，输入 表示实例的特征向量，输出 表示实例类型
* 称为感知机，其中
* 感知机是一种线性分类模型，属于判别模型。