## 2020.3.15 第七次读书报告

09018330 孙毅远

### 一、自己提出的问题的理解

#### 1.如何证明习题2.3

理解：形象化来说两个凸包不相交便可以找到分割的方法，但严谨证明经过思考还是不太会

### 二、别人提出的问题的理解

#### 1. 为什么“损失函数的一个自然选择是误分类点的总数”？

理解：这是一个由损失函数定义而来的初始想法

#### 2.怎么理解当训练数据集线性可分时，感知机学习算法存在无穷多个解，其解由于不同的初值或不同的迭代顺序而可能有所不同。

理解：指的是最终的超平面的解会由于初值的不同或者迭代顺序而变化，从结果上看，会存在一定范围内的超平面都满足要求

#### 3. 如何理解实例点更新次数越多，它距离分离超平面的距离越近？

理解：每次操作都是调整w和b,使得超平面离实例点更接近，而满足分割

### 三、读书计划

#### 本周 统计学习方法学完第三章

#### 下周 学完第四章

### 四、读书笔记

#### 3.1 k近邻算法

* 给定一个数据集，对新的输入实例，在训练数据集中找到与该实例最邻近的k个实例，这k个实例的多数属于某个类，就把该输入实例分为这个类

#### 3.2 k近邻模型

* k近邻模型相当于根据训练集、距离度量与k值及分类决策规则，将特征空间划分为子空间，确定子空间每个点所属的类
* 特征空间中，对于每个训练实例点xi，距离该点比其它点更近的所有点组成一个域，叫单元（）
* 距离度量
* 特征空间中的两个实例点的距离是两个点的相似程度的反映：
* $当p=\infty时，它是各个坐标距离的最大值，L\_{\infty}(x\_{i},x\_{j})=max\_{l} \ |x\_{i}^{(l)}-x\_{j}^{(l)}|$
* 不同距离度量所确定的最近邻点是不同的
  + 较小的k值，相当于用较小的邻域中的训练实例进行预测，近似误差（）会减小，但估计误差()会增大，k值的减小意味着整体模型变复杂，容易发送过拟合
  + 较大的k值，估计误差会减小，但近似误差会增大，k值的增大意味着整体模型变简单
* 在应用中，一般选取一个较小的k值，然后用交叉验证法选取最优值
* 分类决策规则
* 往往采用多数表决规则（）即由输入实例的k个邻近的训练实例中的多数类决定输入实例的类
* 等价于经验风险最小化

#### 3.3 k近邻法的实现：kd树

* 构造kd树
  + kd树是一种对k维空间的实例点进行存储以便对其进行快速检索的树形数据结构，是二叉树，表示对k维空间的一个划分（）
  + kd树每个结点对应于一个k维超矩形区域
  + 依次选择坐标轴对空间切分，选择训练实例点在选定坐标轴上的中位数为切分点，得到的kd树是平衡的，但效率未必是最优的
  + 构造方法
  + $2)重复：对深度为j的顶点，选择x^{(l)}为切分的坐标轴，l=j(mod\ k)+1，以该结点的区域中所有实例的x^{(l)}坐标的中位数\\为切分点，将该结点对应的超矩形区域切分为两个子区域$
* 搜索kd树
* k近邻从当前实例数据点位于的叶节点出发，层层向上退，搜索限制在局部，效率大为提高，目标的最近邻一定在以目标为中心且通过当前最近点的超球体内部
* 每一层搜索更近邻的点可能出现在当前父节点的另一个子节点内，搜索复杂度是log N