李航统计学习：

第五章讨论部分：

读书报告内容：

1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：:cart算法对比id4.5算法有什么优势，为什么要选择cart算法？

讨论后的理解：

CART的全称是分类和回归树，既可以做分类算法，也可以做回归。

决策树的优缺点：

优点：

1.可以生成可以理解的规则。

2.计算量相对来说不是很大。

3.可以处理连续和种类字段。

4.决策树可以清晰的显示哪些字段比较重要

缺点：

1. 对连续性的字段比较难预测。

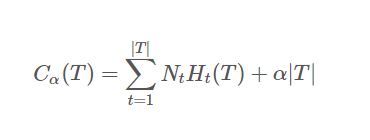
2.对有时间顺序的数据，需要很多预处理的工作。

3.当类别太多时，错误可能就会增加的比较快。

4.一般的算法分类的时候，只是根据一个字段来分类。

1. 提出的问题2：

损失函数Ca（T）的后面只是简单的用T的叶节点数来衡量树的复杂程度，有没有什么更好的办法？



右边第一项表示误差大小，第二项表示模型的复杂度，也就是用叶节点表示，防止过拟化。

损失函数中的第一项表示模型对训练数据的预测误差，也就是模型的拟合程度，第二项表示模型的复杂程度，通过参数 α 控制二者的影响力。一旦 α 确定，那么我们只要选择损失函数最小的模型即可。

二、（必填）别人提出的问题的理解（选择几个问题罗列，并给出理解）：

问题3：P80进行剪枝的条件，如果损失函数值变小就剪枝掉，会不会因为样本过少导致分类结果更不准确，有没有其他的剪枝方法或者影响因素？

自己的理解：

在构建决策树的时候是通过测试集合进行决策树的构建的，但是测试集合的数据很容易产生过拟合的现象，为了防止过拟合，导致对实验数据预测结果不佳，所以使用验证集对决策树进行剪枝防止决策树过拟合，在剪枝的过程中，通过验证机比如说交叉验证的方式，能够有效的防止过拟合，只要数据正常分布，是不会出现剪枝过度而分类结果不好的情况。

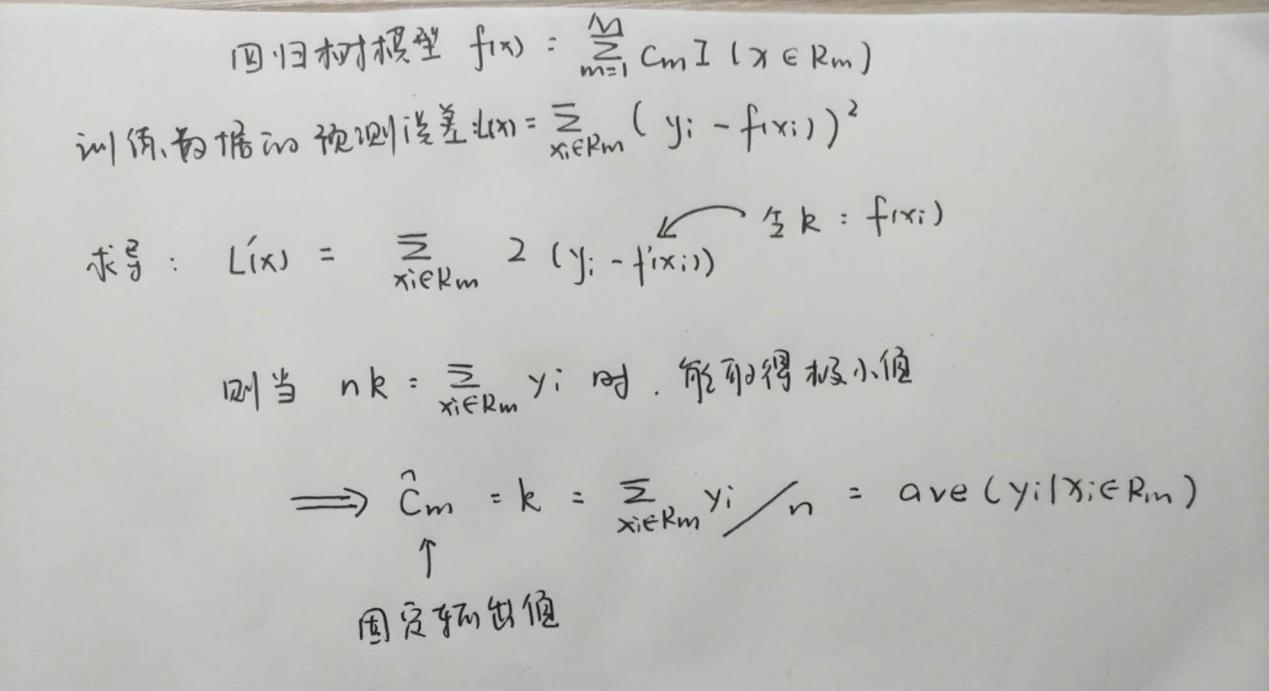
问题4：以信息增益作为划分训练数据集的特征，为什么会存在偏向于选取值较多的特征？

自己的理解：

从公式出发，信息增益是整个数据集的经验熵与特征A对整个数据集的经验条件熵的差值，信息增益越大即经验条件熵越小，那什么情况下的属性会有极小的的经验条件熵呢？举个极端的例子，如果将身份证号作为一个属性，那么，其实每个人的身份证号都是不相同的，也就是说，有多少个人，就有多少种取值，如果用身份证号这个属性去划分原数据集，那么，原数据集中有多少个样本，就会被划分为多少个子集，这样的话，会导致信息增益公式的第二项整体为0，虽然这种划分毫无意义，但是从信息增益准则来讲，这就是最好的划分属性。其实从概念来讲，就一句话，信息增益表示由于特征A而使得数据集的分类不确定性减少的程度，信息增益大的特征具有更强的分类能力。

问题5：P81为什么说Rm上的最优值是Rm上所有输入实例x对应的y的均值？

自己的理解：



可以通过数学方法求导，最后的极值就是Rm上的最优值。

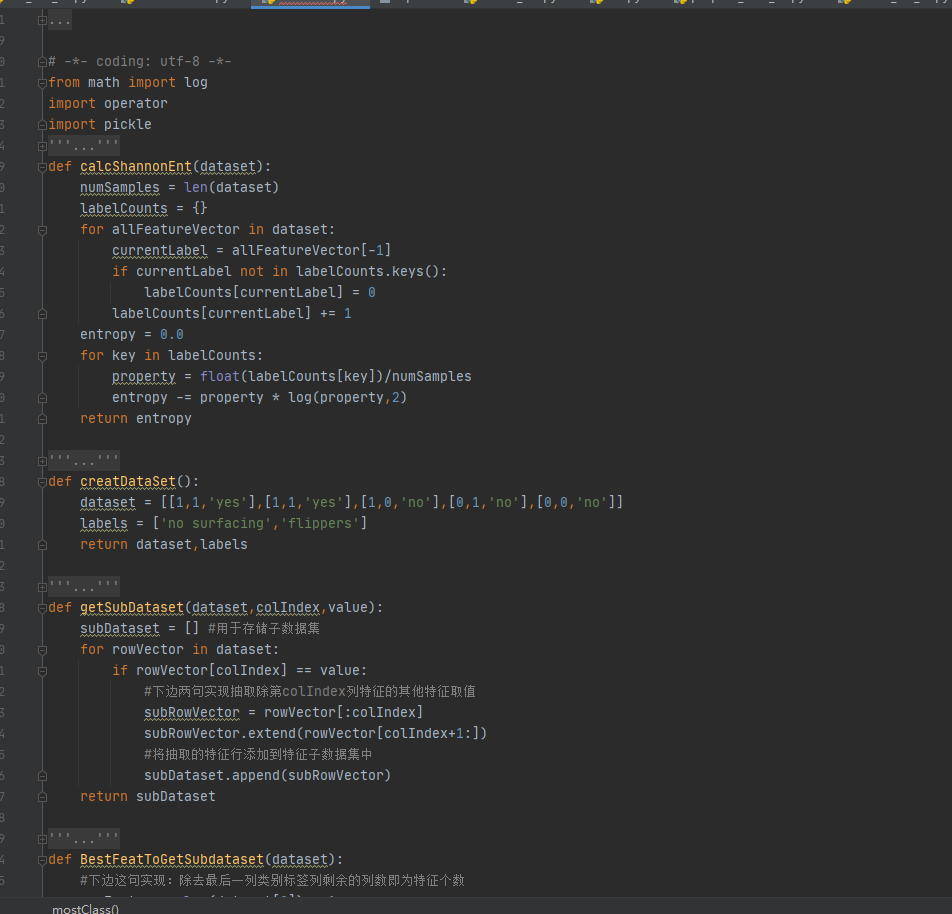
1. （必填）读书计划
2. 本周完成的内容章节：

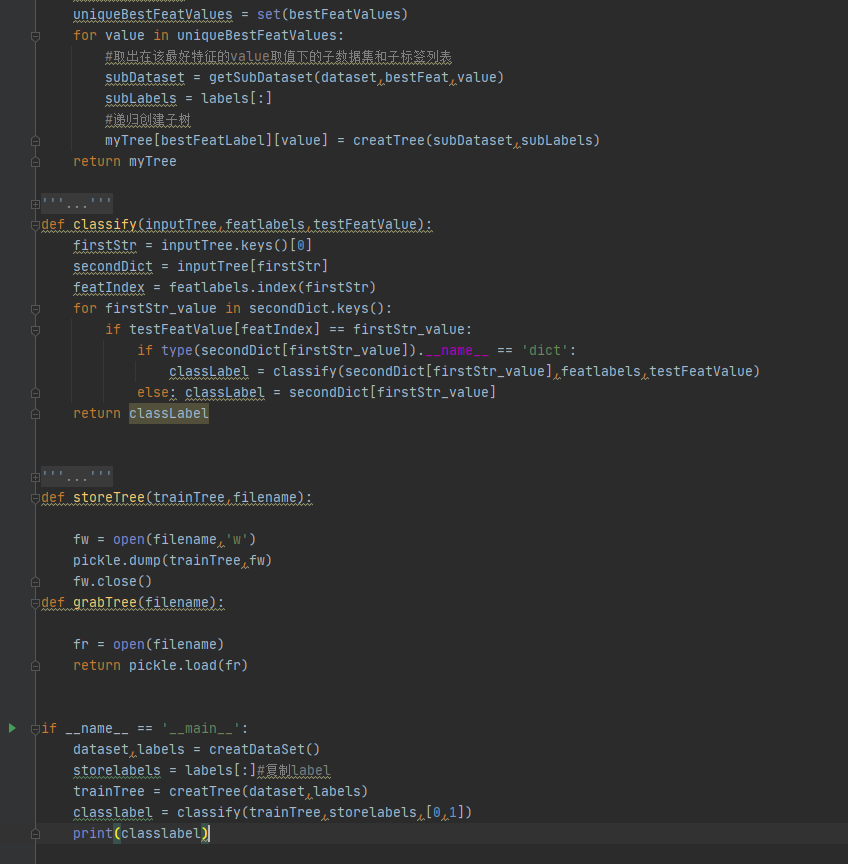
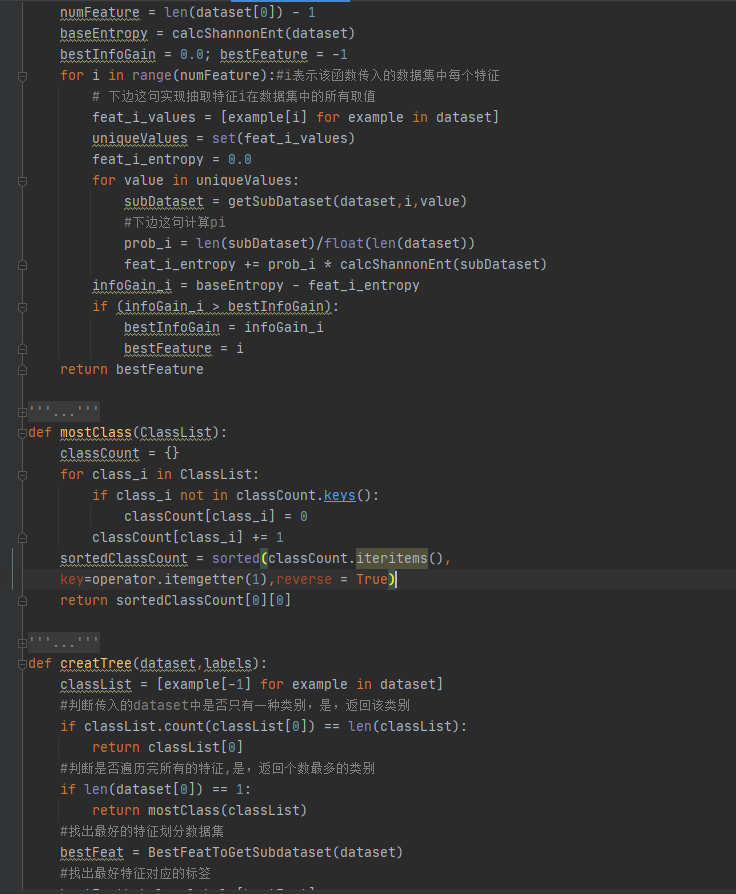
（1）第五章完成并且梳理知识点，寻找问题，自己思考并且于小组会之前完成了自己的思考。

2、下周计划：第五章的阅读.

四、（选做）读书摘要及理解或伪代码的具体实现（读书摘要、伪代码的具体实现代码等可以写到这个部分）

ID3算法实现：





读书摘要：

### 决策树三种算法优缺点比较：

### **ID3**

D3算法十分简单，核心是根据“最大信息熵增益”原则选择划分当前数据集的最好特征，信息熵是信息论里面的概念，是信息的度量方式，不确定度越大或者说越混乱，熵就越大。在建立决策树的过程中，根据特征属性划分数据，使得原本“混乱”的数据的熵(混乱度)减少，按照不同特征划分数据熵减少的程度会不一样。在ID3中选择熵减少程度最大的特征来划分数据（贪心），也就是“最大信息熵增益”原则。

同时这是最早提出的一种决策树方法，使用上述信息增益的方式建立。

缺点：只能处理离散型属性，并且对倾向于选择取值较多的属性；（为什么？）

****原因****：

信息增益反映的给定一个条件以后不确定性减少的程度,必然是分得越细的数据集确定性更高,也就是条件熵越小,信息增益越大

### **C4.5**

C4.5算法流程与ID3相类似，只不过将信息增益改为信息增益比，以解决偏向取值较多的属性的问题，另外它可以处理连续型属性。

### **CART**

CART是一棵二叉树，采用二元切分法，每次把数据切成两份，分别进入左子树、右子树。而且每个非叶子节点都有两个孩子，所以CART的叶子节点比非叶子多1。相比ID3和C4.5，CART应用要多一些，既可以用于分类也可以用于回归。CART分类时，使用基尼指数（Gini）来选择最好的数据分割特征，gini描述的是纯度，与信息熵的含义相似。CART中每一次迭代都会降低GINI系数

相对于ID3使用的信息增益，CART中用于选择变量的不纯性度量是Gini指数，总体内包含的类别越杂乱，GINI指数就越大（跟熵的概念很相似）。

****GINI指数****：   
1、是一种不等性度量；   
2、通常用来度量收入不平衡，可以用来度量任何不均匀分布；   
3、是介于0~1之间的数，0-完全相等，1-完全不相等；   
4、总体内包含的类别越杂乱，GINI指数就越大（跟熵的概念很相似）

****CART分析步骤****：

1、从根节点t=1开始，从所有可能候选S集合中搜索使不纯性降低最大的划分S\*，然后，使用划分S\*将节点1（t=1）划分成两个节点t=2和t=3；   
2、在t=2和t=3上分别重复划分搜索过程。

****基尼不纯度指标****：

在CART算法中, 基尼不纯度表示一个随机选中的样本在子集中被分错的可能性。基尼不纯度为这个样本被选中的概率乘以它被分错的概率。当一个节点中所有样本都是一个类时，基尼不纯度为零

****离散和连续目标变量的区别****：

同时，如果目标变量是标称的，并且是具有两个以上的类别，则CART可能考虑将目标类别合并成两个超类别（双化）；   
如果目标变量是连续的，则CART算法找出一组基于树的回归方程来预测目标变量。

知识图谱部分：

工作报告内容：

本周主要做了两方面的工作

1.一个是安卓开发的自学为吴老师项目做安卓做准备，继续推进自己安卓开发的学习

2.使用ConnectionE的模型做type inference

针对上周ConnectionE模型出现的问题：

1源码太老了，实际下载跑出有很多问题，正在调试。作者自己说不是最新的版本，跑出大量报错，我在改改，试试能不能用。

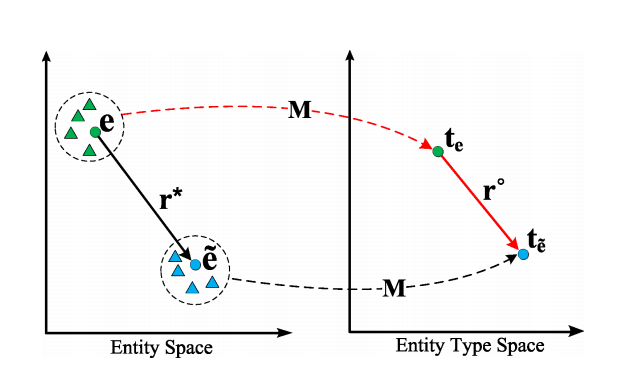
2该算法的trt模型需要尾实体的type进行推断，现有的数据只有头实体的type，需要重新做出尾实体的entity\_type的键值对。

针对问题本周工作：

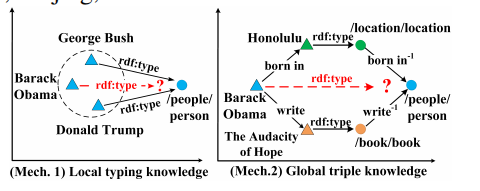
1. 补充学习了深度学习的相关知识，调试了他本来的代码，问题较多，带着问题去搜索相关的博客学习，同时也有编译器版本的问题，python3.8很多不兼容，进行了版本的降级，并且原模型的type是层次结构的type，讨论后进行了相应的修改，可以在他本来的数据集上跑通了。
2. 但是模型需要的数据集有三种。
3. type\_entity键值对分别作为模型的训练集，验证集和测试集
4. 关系三元组entity\_relation\_type
5. 关系三元组type\_relation\_type
6. 数据集的处理

A.我手上有的是之前手工和基于规则推导的两万条不到的较为正确的entity\_type键值对，数据量对于深度学习训练模型来说比较小，所以为了进行ConnectionE模型的训练，我采取交叉验证的方式，按照9：1的比例进行数据集的划分。抽取了一万五千多条（e，t）数据集作为训练集。剩余的作为测试集和验证集。

B.对于关系三元组，我手上的只有properties三元组，并不能直接使用。对此，我去跟李志强讨论后，选择了进行头尾entity的实体匹配，并且对于relation出现的频次进行了排序和清洗，可以认为太高频次出现的relation没有太大的作用，进行了筛选，最终得到了关系rdf。

C.将关系三元组entity\_relation\_entity将entity替换为type构成type\_relation\_type的三元组

如图所示：将entity通过矩阵M变换为type。

例如：people替换掉本来的Barack Ibama所对应的rdf应该是 仍然成立的。那么有如图所示的

即为由（Barack Obama,born in,Honolulu）变成了（people,born in,location）

下周工作内容：

1继续推进安卓开发的学习.

2跑起来ConnectionE模型，并且进行参数的调整，得到最后的结果

3进行数据结果的评估，计算准确率召回率，f1值等指标与其他算法进行对比分析