SHANGHAI JIAO TONG UNIVERSITY

研究生课程作业

GRADUATE COURSE PROJECT REPORT

学生姓名: 季华伟

学生学号: 023034910026

任课教师: 陈海宝 副教授

学院(系)： 电子信息与电气工程学院 （微纳电子学系）

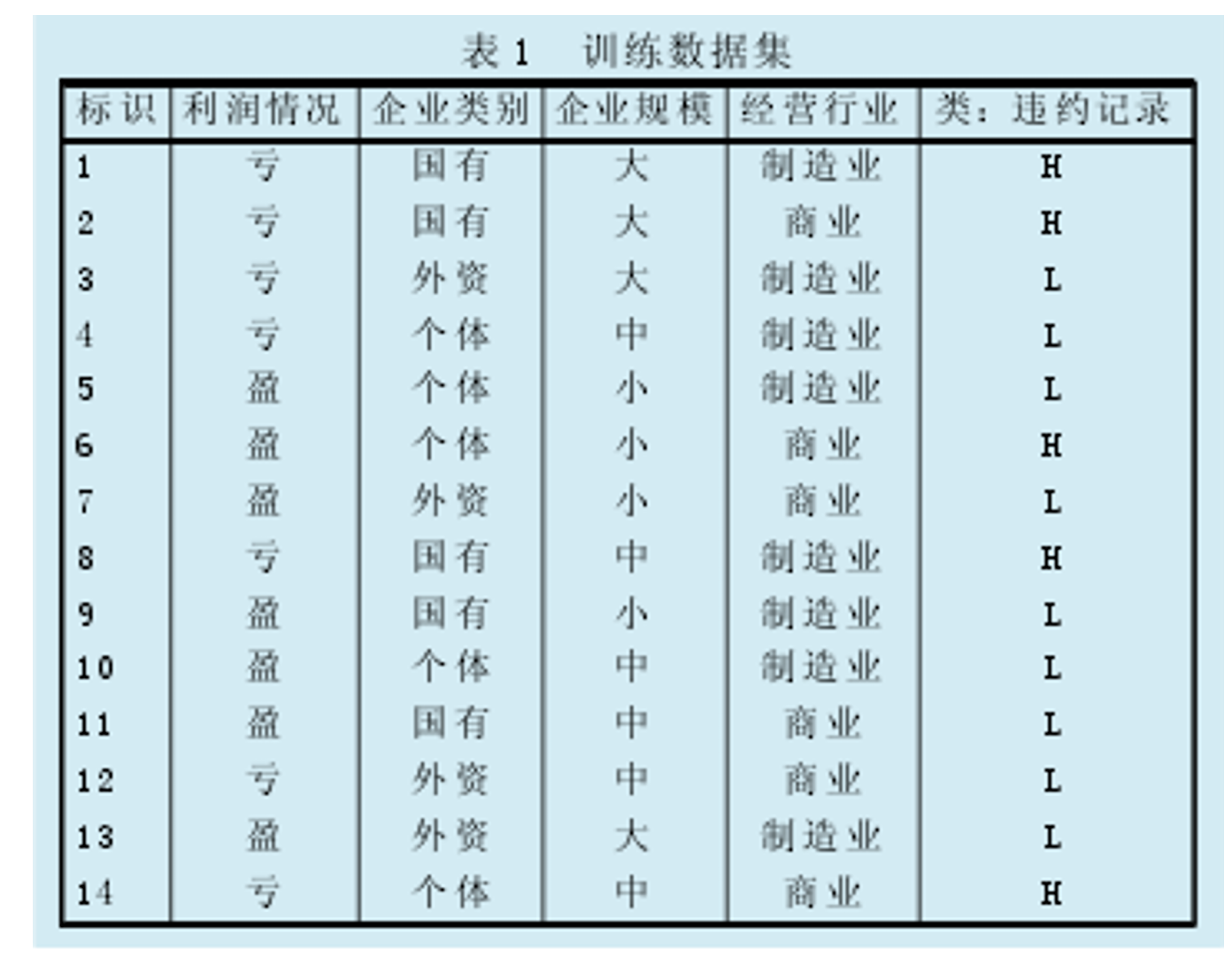
开课学期： 2024年 （春季）

# 问题描述

题目1（K均值法或K中心点法）：已知有20个样本（可根据需要自行扩充数据），每个样本有2个特征，数据分布如下表所示，使用K均值法或K中心点法实现样本分离。



题目2（决策树）：请针对Lecture10第29页PPT中的银行违约率（也可根据需要自行扩充数据）模型构造一棵决策树。



# 方法描述

方法描述大部分来自sklearn，以及一些相关blog。

1. KMeans算法

KMeans算法的核心是想用一种类方差近似的方法，将数据分成n份。其训练目标是最小化组内均方距离。即优化下式：

其中，n是样本总数，是聚类C中每个簇的均值。这些均值又被称为“质心”。上式可以用来衡量簇内的相关程度。

KMeans算法一般采用Lloyd算法，即首先选择一些初始的“质心”，比如从N个样本中随机选择k个样本点。之后计算质心其他各个点之间的距离，距离近的点分配到对应的质心上。然后通过求分配到不同质点上的点的均值，更新质点。重复上述两个步骤，并计算新旧质点之间的距离，知道距离小于某个阈值，即质点稳定，停止训练。

1. 决策树算法

假设训练数据，其中，标签。第m个节点处的数据表示为。对于其中每个数据点按照规则分割成和两个部分。其中j是某个特征，表示按该特征被分割的阈值。

分割之后得到的两个分支，计算不纯度：

其中是计算每个分支的不纯度函数。其一般有如下两种形式：

Gini :

是第m个节点上的概率。

在m节点找到一个使得最小的规则，将m节点分割。

# 实验结果



图1 聚类结果

我们给出了题目一的聚类结果，如图1所示。可以发现聚类效果非常好，KMeans算法将两个簇完全分开。我们同时画出了聚类的质心，可以发现质心未必是数据点本身。

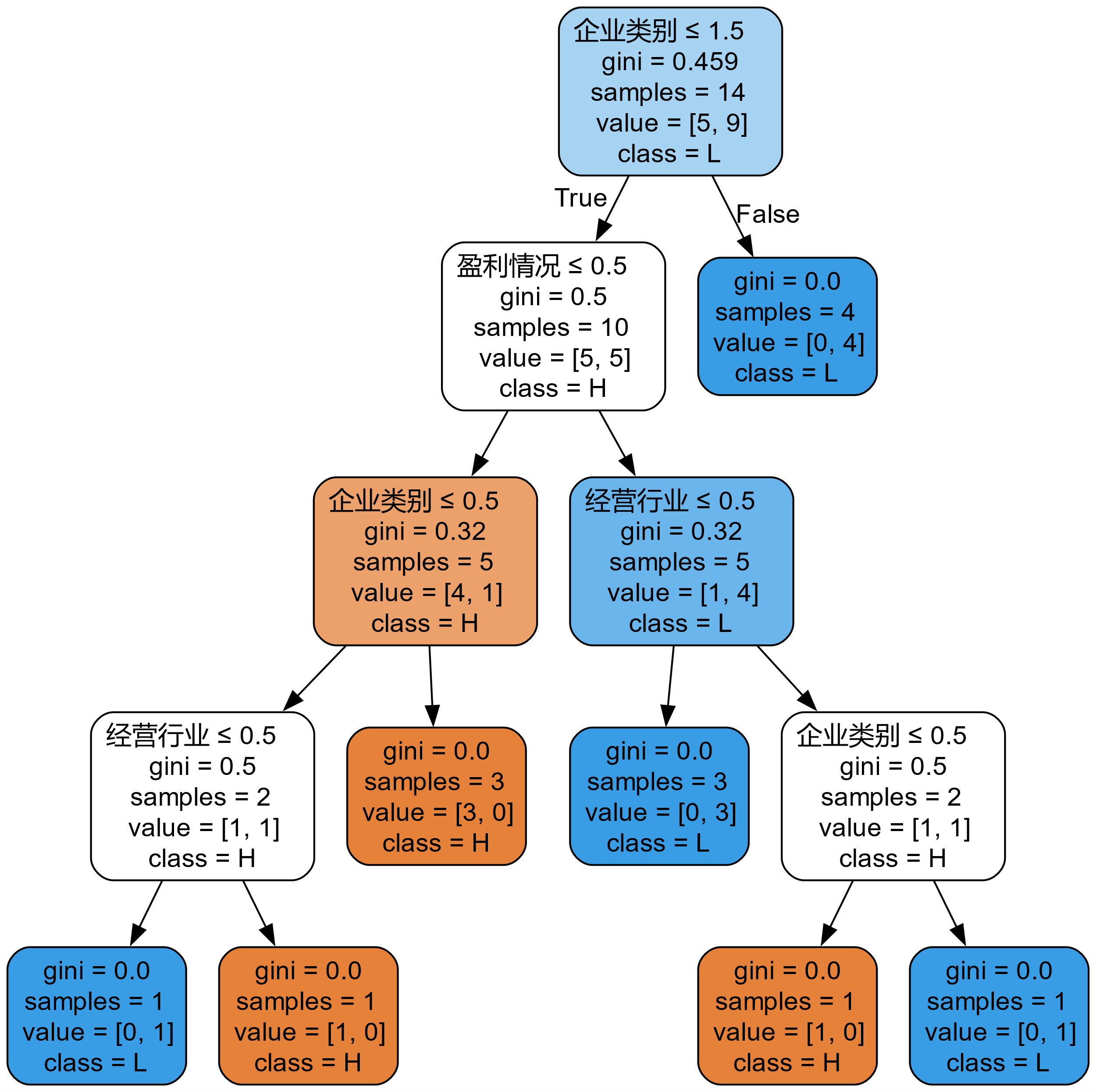


图2决策树生成结果

我们给出了题目二决策树生成的结果，其中渲染后的颜色表示类别，深浅表示不纯度的大小，颜色越深表示不纯度越小。另外，经过决策树训练之后，模型针对原始数据的拟合达到了100%。

# 结论

使用KMeans这种无监督式的训练方法对数据进行聚类，可以使得密度近似的数据得到很好的聚类，题目一就是一个很好的例子。决策树用于制定规则进行分类，题目二中的问题可以使用决策树得到很好的解决。

# 参考文献：

[1] <https://scikit-learn.org/stable/modules/tree.html#mathematical-formulation>

[2] [https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#](https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html%23)

[3] [https://scikit-learn.org/stable/modules/clustering.html#k-means](https://scikit-learn.org/stable/modules/clustering.html%23k-means)

[4] [https://scikit-learn.org/stable/auto\_examples/tree/plot\_unveil\_tree\_structure.html#understanding-the-decision-tree-structure](https://scikit-learn.org/stable/auto_examples/tree/plot_unveil_tree_structure.html%23understanding-the-decision-tree-structure)

[5] [https://www.biaodianfu.com/decision-tree.html#%E5%89%AA%E6%9E%9D%E7%9A%84%E5%88%86%E7%B1%BB](https://www.biaodianfu.com/decision-tree.html%23%E5%89%AA%E6%9E%9D%E7%9A%84%E5%88%86%E7%B1%BB)