实验一

分别使用一种监督学习方法和一种非监督学习方法,各自解决一个实际问题,在该问题对应的测试数据上进行实验,分析实验结果,提交实验报告。

1.监督式学习

待解决的问题：通过温度特征，预测西雅图是否下雨。

数据集：西雅图 1948-2017 年的每日天气情况[1]。

- DATE = the date of the observation
- PRCP = the amount of precipitation, in inches
- TMAX = the maximum temperature for that day, in degrees Fahrenheit
- TMIN = the minimum temperature for that day, in degrees Fahrenheit
- RAIN = TRUE if rain was observed on that day, FALSE if it was not

数据筛选：使用 RAIN 作为标签，使用除去 PRCP 列的其他列作为特征。

学习算法：决策树（sklearn.tree.DecisionTreeClassifier）

测试方法：交叉验证（Cross-validation）

实验结果：

	precision	recall	f1-score	support
0	0.79	0.80	0.80	3703
1	0.72	0.71	0.72	2684
avg / total	0.76	0.76	0.76	6387

准确率和召回率均达到 76%，在天气预测领域，该结果比较理想。

2.非监督式学习

待解决的问题：根据运输车队驾驶员的驾驶特征，对其进行分类。

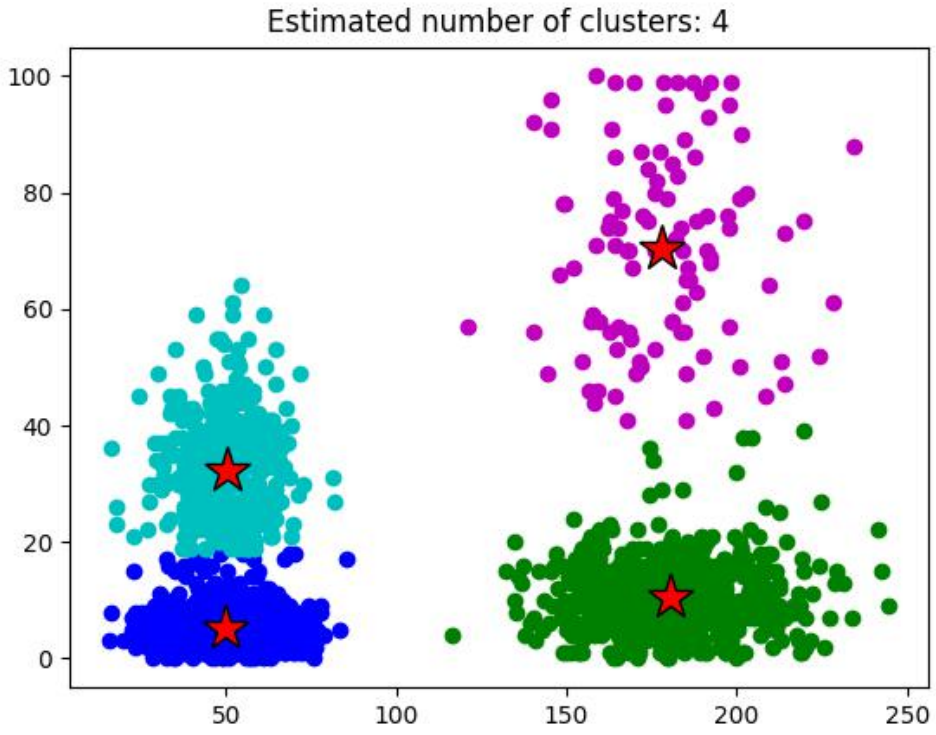
数据集：某运输车队驾驶员的驾驶数据[2]。

- Driver_ID = 驾驶员编号
- Distance_Feature = 每日驾驶里程数
- Speeding_Feature = 驾驶速度超限速 5mph 的时间占比

学习算法：K-Means (sklearn.cluster.KMeans)

测试方法：无

实验结果：



通过聚类算法，将驾驶员分为 4 类。

Reference:

[1]
https://www.kaggle.com/rtatman/did-it-rain-in-seattle-19482017/downloads/did-it-rain-in-seattle-19482017.zip
[2]
https://raw.githubusercontent.com/datascienceinc/learn-data-science/master/Introduction-to-K-means-Clustering/Data/data_1024.csv