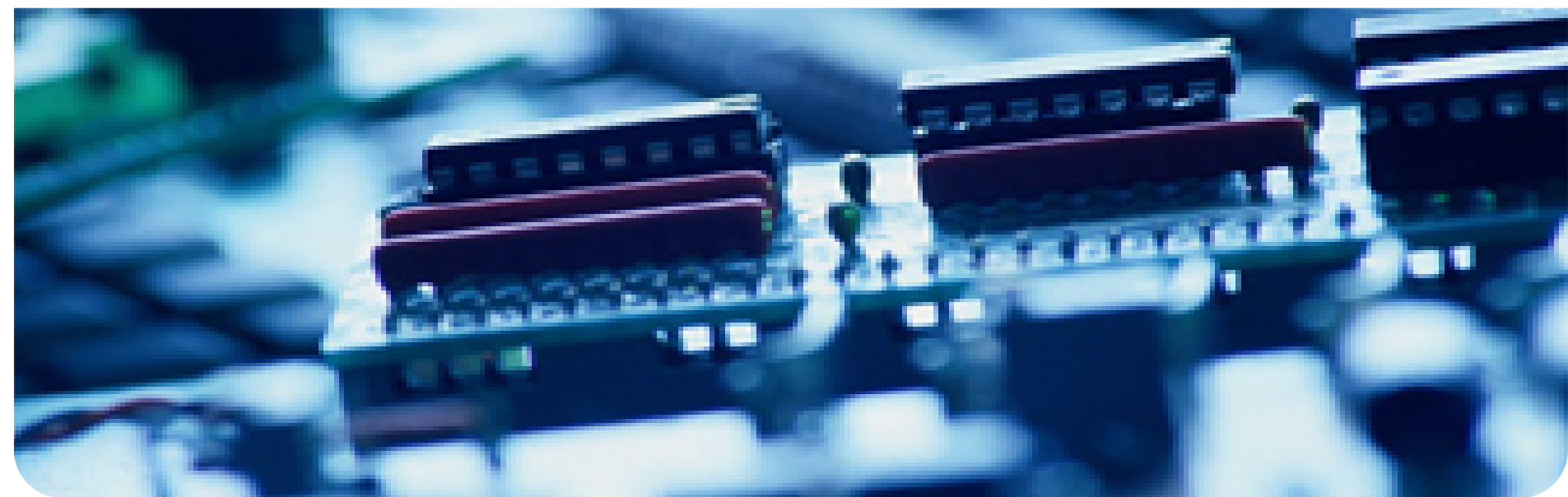


数据挖掘和大数据分析



Outline

① Review (Assignments & KNN & Distance)

② Kmeans Algorithm



③ KNN & Kmeans Project



Review (Assignments & KNN & Distance)

作业清单 (5/11)

【1】 Pandas Series 是什么? Pandas 中的 DataFrame 是什么? 如何将 numpy 数据转成 DataFrame 格式的数据? 如何将 Series 数据转成 DataFrame 格式的数据? 如何将 DataFrame 转换为 NumPy 数组? 如何对 DataFrame 进行排序? 什么是数据聚合? (注: 每一小问, 举例说明)

【2】 利用 iris.csv 数据集, 建立 KNN 模型, 预测 Sepal.Length\Sepal.Width\Petal.Length\Petal.Width 分别为 (6.3, 3.1, 4.8, 1.4) 时, 属于鸢尾花的哪个类别? 编写 KNN 源代码。

【3】 计算 $X = [1, 2, 3]$ 和 $Y = [0, 1, 2]$ 的曼哈顿距离 (Manhattan Distance), 切比雪夫距离, 闵可夫斯基距离, 标准化欧氏距离, 马氏距离。给出计算公式, 并根据公式计算。利用 Python 实现上述距离。

Do you finish it in 6 minutes?


A Yes

B No

提交

Student Name	T1	T2
A	1	1
B	2	1
C	4	3
D	5	4

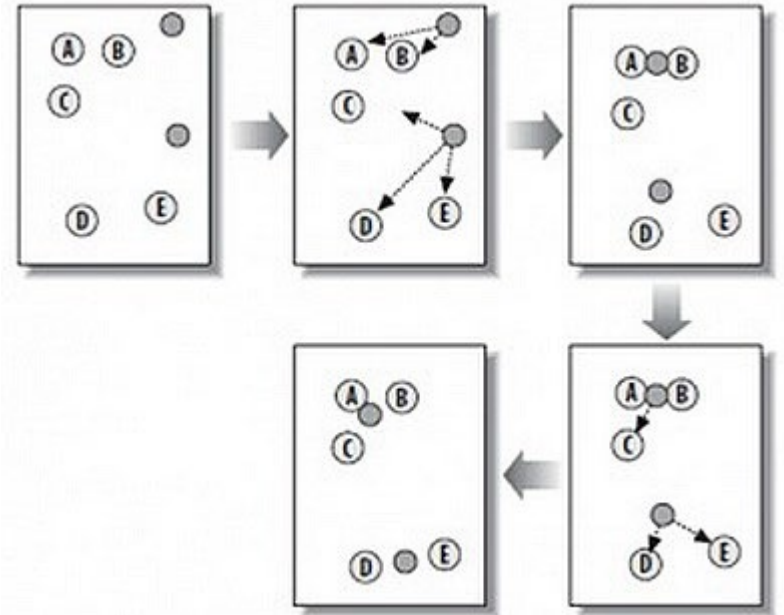
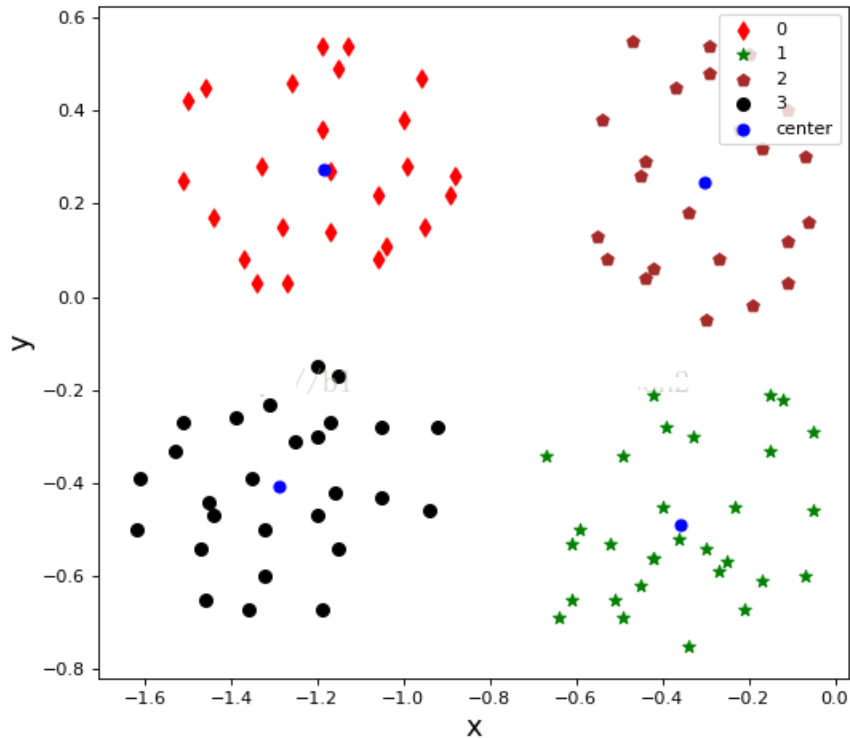
$C1 = (1,1)$ $C2 = (2,1)$



DATA ANALYTICS: **DATA MINING AND BIG DATA**



K-means Algorithm



K-means Algorithm



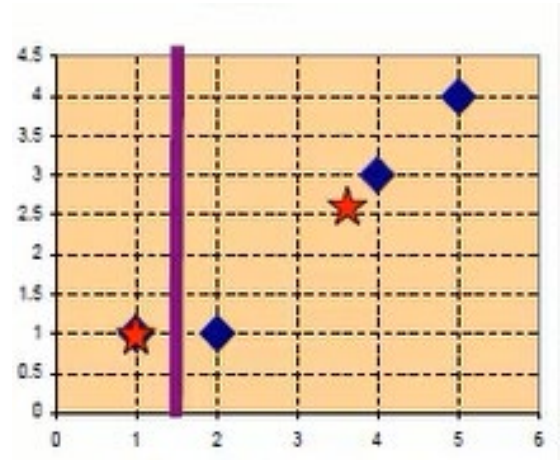
Student Name	T1	T2	T3	Final Exam	Project	Rank
张三	12	12	13	28	24	?
李四	7	11	10	19	21	?
王五	12	14	11	27	23	?
赵六	6	7	4	13	20	?
刘七	13	14	13	27	25	?

Student Name	T1	T2
A	1	1
B	2	1
C	4	3
D	5	4

K-means Algorithm

Student Name	T1	T2
A	1	1
B	2	1
C	4	3
D	5	4

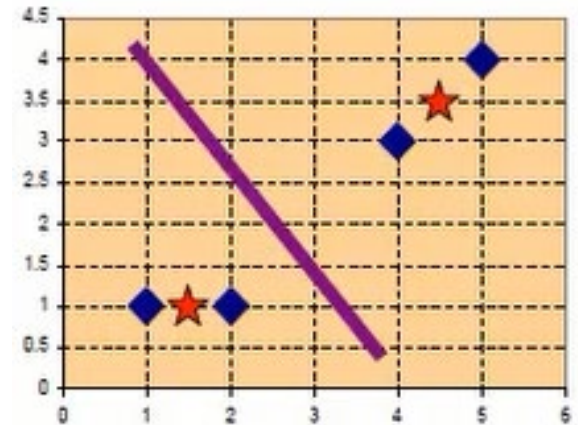
$$D_0 = \begin{bmatrix} 0, 1, 3.61, 5 \\ 1, 0, 2.83, 4.24 \end{bmatrix} \quad G_0 = \begin{bmatrix} 1, 0, 0, 0 \\ 0, 1, 1, 1 \end{bmatrix}$$



K-means Algorithm

Student Name	T1	T2
A	1	1
B	2	1
C	4	3
D	5	4

$$D_2 = \begin{bmatrix} 0.5, 0.5, 3.20, 4.61 \\ 4.30, 3.54, 0.71, 0.71 \end{bmatrix} \quad G_1 = \begin{bmatrix} 1, 1, 0, 0 \\ 0, 0, 1, 1 \end{bmatrix}$$



KNN & K-means Project



setosa

0



versicolor

1

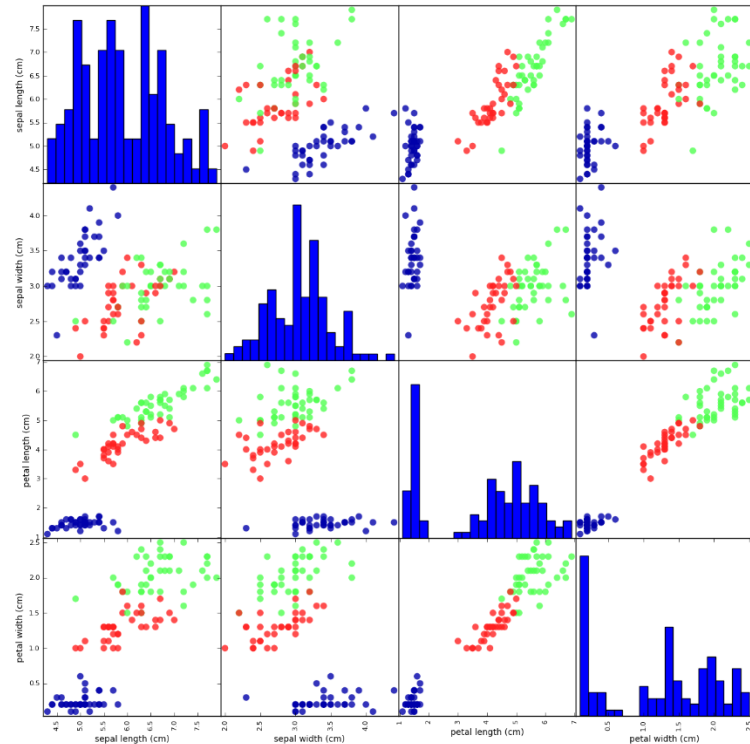
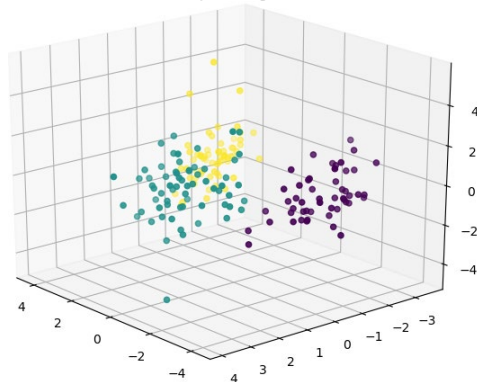
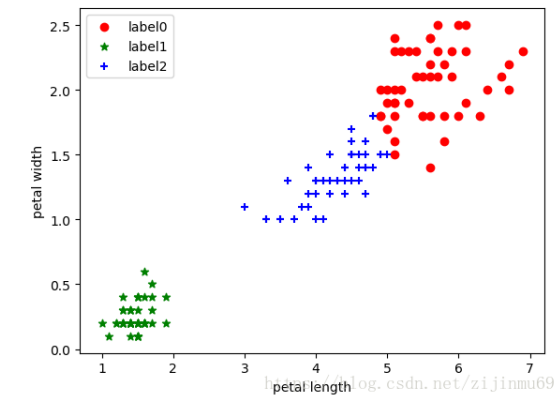
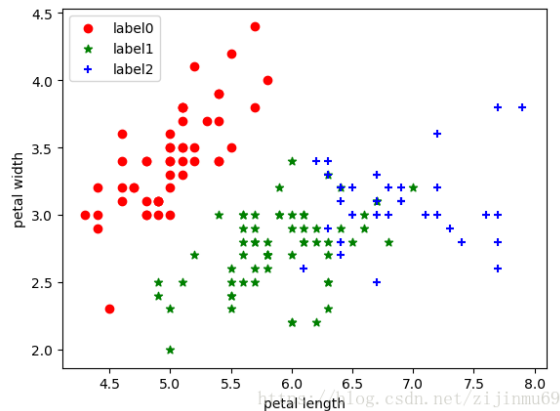
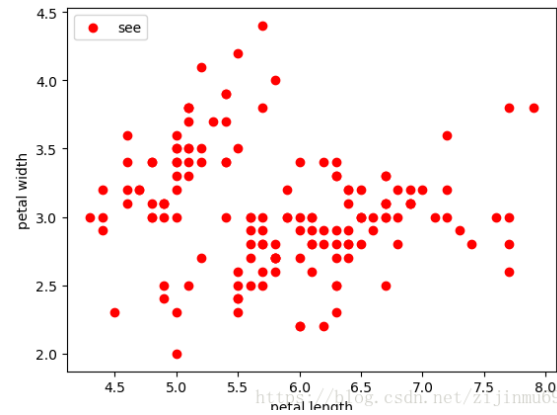


virginica

2



KNN & K-means Project



Do you understand Decision Tree and Random Forest?

A

Yes

B

No

提交



贵在坚持！