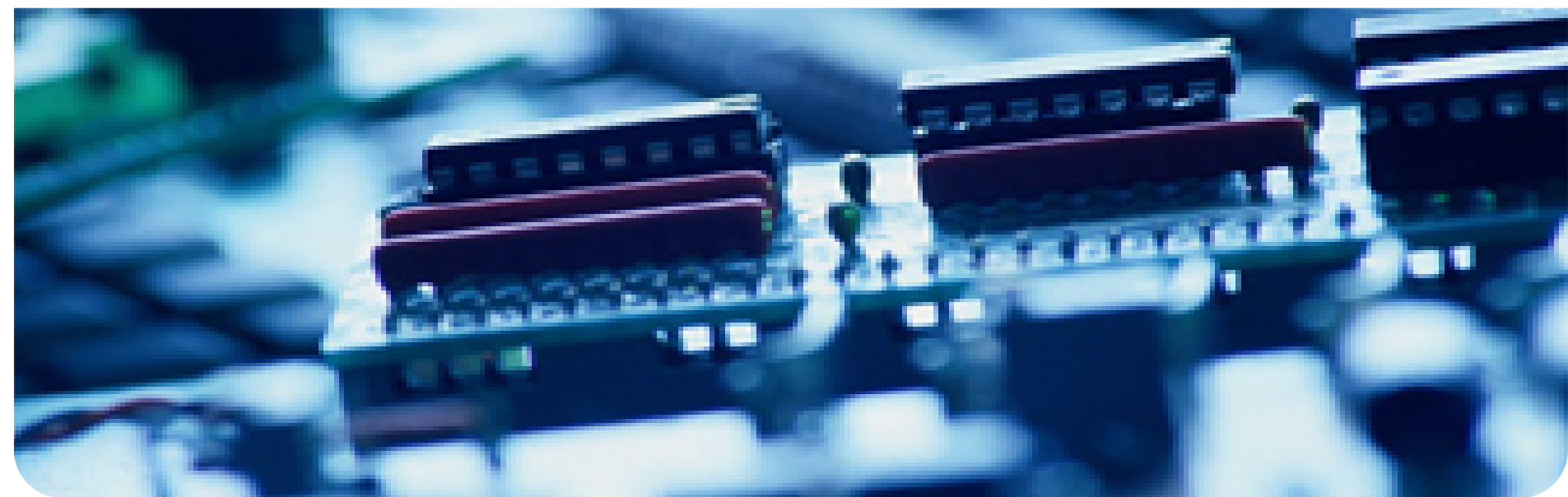


数据挖掘和大数据分析



Outline

① Review (Assignments & Kmeans & KNN)

② Decision Tree Algorithm



③ Decision Tree Project



Review (Assignments & Kmeans & KNN)

作业清单 (5/13)

【1】选择4名同学 A、B、C、D，两次小测成绩，利用 Kmeans 算法分为“优秀”和“及格”两类。@注意：不能直接调用 sklearn 第三方库的 KMeans 函数，根据课堂讲授的分类过程，编写代码。撰写实验报告。

学生姓名	小测 1	小测 2
A	1	1
B	2	1
C	4	3
D	5	4

Review (Assignments & Kmeans & KNN)

【2】根据下列成绩单，将 5 名同学成绩归为 A 类、B 类、C 类，利用 Kmeans 算法实现。@注意：不能直接调用 sklearn 第三方库的 KMeans 函数，根据课堂讲授的分类过程，编写代码。撰写实验报告。

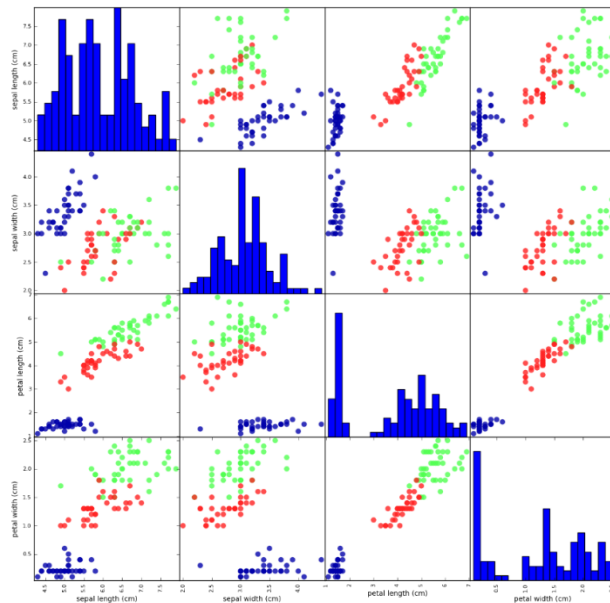
学生姓名	小测 1	小测 2	小测 3	期末成绩	项目答辩	成绩
张三	12	15	13	28	24	?
李四	7	11	10	19	21	?
王五	12	14	11	27	23	?
赵六	6	7	4	13	20	?
刘七	13	14	13	27	25	?


Review (Assignments & Kmeans & KNN)

【3】 利用 Sklearn 的标准 KNN 和 KMeans 方法,数据集为“wine.csv”
(见[微信群](#)),通过 KNN 算法,对葡萄酒的测试集进行标注,然后对比预测标签值和已知标签值,得到 KNN 算法的预测准确率。通过 Kmeans 算法,对无标签的“wine.csv”进行分类,自己设定 K 值和初始中心点值。

Review (Assignments & Kmeans & KNN)

- 【4】 利用 KMeans 算法对 “iris.csv” 数据集的无标签数据分为 3 类，用三维图形可视化分类结果。
- 【5】 利用 KMeans 算法对 “iris.csv” 数据集的无标签数据分为 3 类，任取 2 个特征值，显示分类结果，用二维图形可视化分类结果，类似下图。





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



—— Machine Learning 4

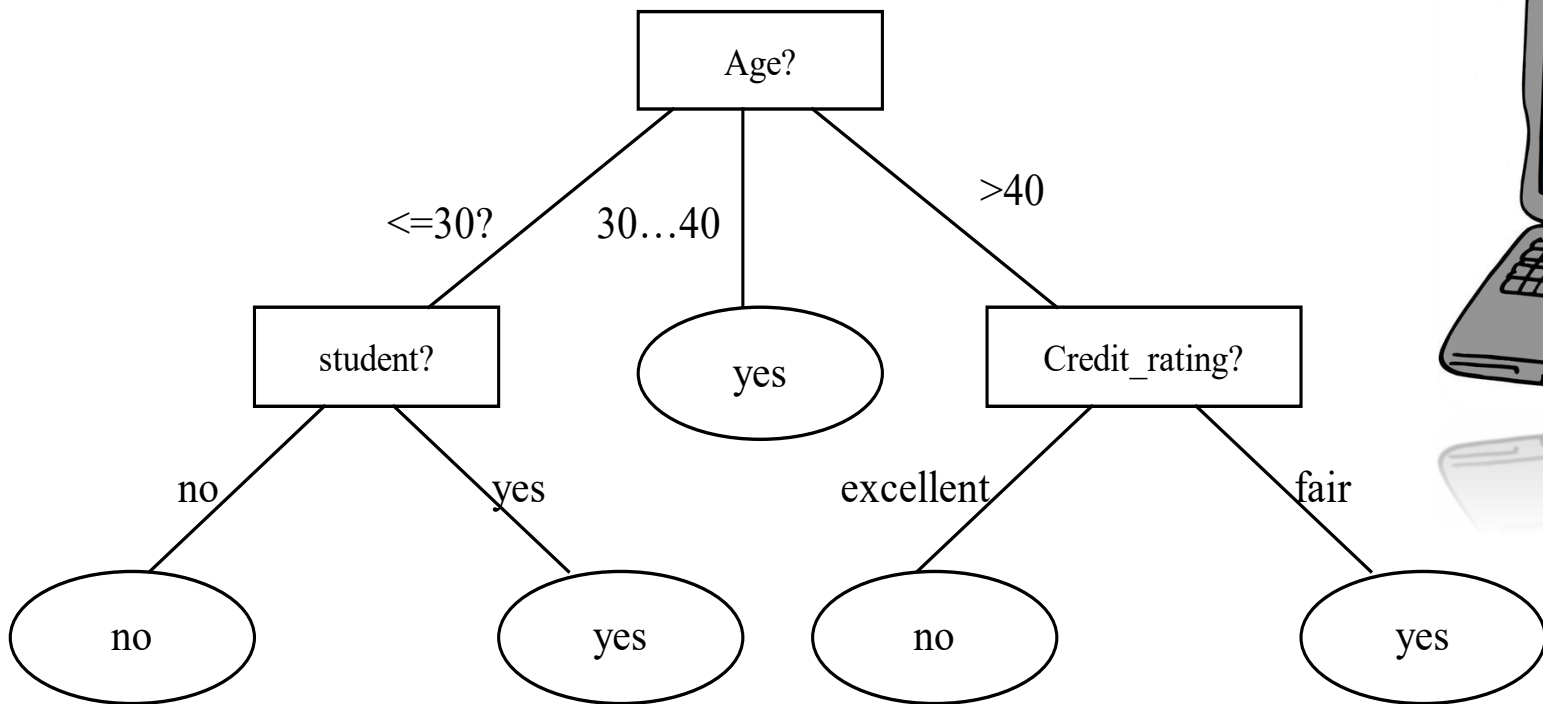


Decision Tree Algorithm



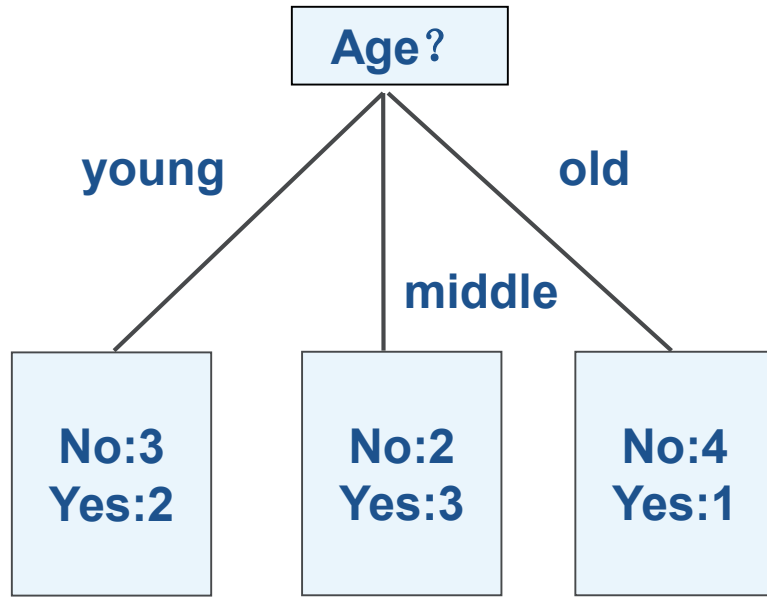
Grade	Graduation Class	Go Back to School in First Batch
Senior One	No	No
Senior Two	No	No
Senior Three	Yes	Yes
Junior One	No	No
Junior Two	No	No
Junior Three	Yes	Yes

Decision Tree Algorithm

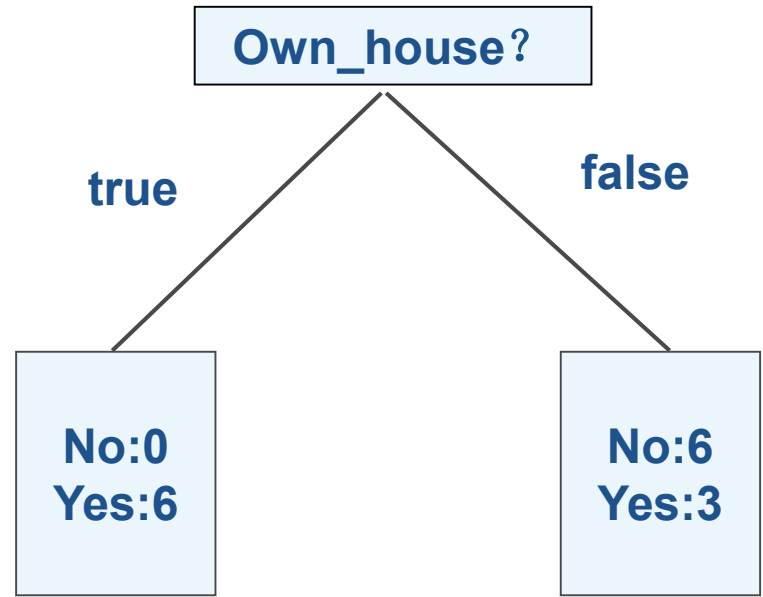


ID	Age	Algorithm	Own_house	Credit_rating	Class
1	Young	False	False	Fair	No
2	Young	False	False	Good	No
3	Young	True	False	Good	Yes
4	Young	True	True	Fair	Yes
5	Young	False	False	Fair	No
6	Middle	False	False	Fair	No
7	Middle	False	False	Good	No
8	Middle	True	True	Good	Yes
9	Middle	False	True	Excellent	Yes
10	Middle	False	True	Excellent	Yes
11	Old	False	True	Excellent	Yes
12	Old	False	True	Good	Yes
13	Old	True	False	Good	Yes
14	Old	True	False	Excellent	Yes
15	Old	False	False	fair	no

Decision Tree Algorithm



(a)



(b)

Decision Tree Algorithm



Entropy

熵



$$H(X) = - \sum_{x \in X} P(x) \log_2 P(x)$$

$$Info(D) = - \sum_{i=1}^m p_i \log_2 p_i$$

$$Info_A(D) = \sum_{j=1}^v \left[\left(\frac{|D_j|}{|D|} \right) * Info(D_j) \right]$$

Age	Salary	STU	Credit	Buy Computer
<30	H	No	OK	No
<30	H	No	Good	No
30-40	H	No	OK	Yes
>40	M	No	OK	Yes
>40	L	Yes	OK	Yes
>40	L	Yes	Good	No
30-40	L	Yes	Good	Yes
<30	M	No	OK	No
<30	L	Yes	OK	Yes
>40	M	Yes	OK	Yes
<30	M	Yes	Good	Yes
30-40	M	No	Good	Yes
30-40	H	Yes	OK	Yes
>40	M	No	Good	No



$$|D|=14$$

$$|C1,D|=5$$

$$|C2,D|=9$$

$$Info(D)$$

$$= -\frac{5}{14} \log_2 \frac{5}{14} - \frac{9}{14} \log_2 \frac{9}{14}$$

$$= 0.940$$

Age	Salary	STU	Credit	Buy Computer
<30	H	No	OK	No
<30	H	No	Good	No
30-40	H	No	OK	Yes
>40	M	No	OK	Yes
>40	L	Yes	OK	Yes
>40	L	Yes	Good	No
30-40	L	Yes	Good	Yes
<30	M	No	OK	No
<30	L	Yes	OK	Yes
>40	M	Yes	OK	Yes
<30	M	Yes	Good	Yes
30-40	M	No	Good	Yes
30-40	H	Yes	OK	Yes
>40	M	No	Good	No



Info年齡(D)

$$\begin{aligned}
 &= \frac{5}{14} \left(-\frac{3}{5} \log \frac{3}{5} - \frac{2}{5} \log \frac{2}{5} \right) \\
 &+ \frac{4}{14} \left(-\frac{4}{4} \log \frac{4}{4} - \frac{0}{4} \log \frac{0}{4} \right) \\
 &+ \frac{5}{14} \left(-\frac{2}{5} \log \frac{2}{5} - \frac{3}{5} \log \frac{3}{5} \right) \\
 &= 0.694
 \end{aligned}$$

Gain(年齡)

= Info(D) - Info年齡(D)

= 0.940 - 0.694 = 0.246

Age	Salary	STU	Credit	Buy Computer
<30	H	No	OK	No
<30	H	No	Good	No
30-40	H	No	OK	Yes
>40	M	No	OK	Yes
>40	L	Yes	OK	Yes
>40	L	Yes	Good	No
30-40	L	Yes	Good	Yes
<30	M	No	OK	No
<30	L	Yes	OK	Yes
>40	M	Yes	OK	Yes
<30	M	Yes	Good	Yes
30-40	M	No	Good	Yes
30-40	H	Yes	OK	Yes
>40	M	No	Good	No



Info收入(D)

$$\begin{aligned}
 &= \frac{4}{14} \left(-\frac{2}{4} \log \frac{2}{4} - \frac{2}{4} \log \frac{2}{4} \right) \\
 &+ \frac{6}{14} \left(-\frac{2}{6} \log \frac{2}{6} - \frac{4}{6} \log \frac{4}{6} \right) \\
 &+ \frac{4}{14} \left(-\frac{1}{4} \log \frac{1}{4} - \frac{3}{4} \log \frac{3}{4} \right) \\
 &= 0.911
 \end{aligned}$$

Gain(收入)

= Info(D) - Info收入(D)

= 0.940 - 0.911 = 0.029

Age	Salary	STU	Credit	Buy Computer
<30	H	No	OK	No
<30	H	No	Good	No
30-40	H	No	OK	Yes
>40	M	No	OK	Yes
>40	L	Yes	OK	Yes
>40	L	Yes	Good	No
30-40	L	Yes	Good	Yes
<30	M	No	OK	No
<30	L	Yes	OK	Yes
>40	M	Yes	OK	Yes
<30	M	Yes	Good	Yes
30-40	M	No	Good	Yes
30-40	H	Yes	OK	Yes
>40	M	No	Good	No



Info学生(D)

$$\begin{aligned}
 &= \frac{7}{14} \left(-\frac{1}{7} \log \frac{1}{7} - \frac{6}{7} \log \frac{6}{7} \right) \\
 &+ \frac{7}{14} \left(-\frac{4}{7} \log \frac{4}{7} - \frac{3}{7} \log \frac{3}{7} \right) \\
 &= 0.788
 \end{aligned}$$

Gain(学生)

$$\begin{aligned}
 &= \text{Info(D)} - \text{Info学生(D)} \\
 &= 0.940 - 0.788 = 0.152
 \end{aligned}$$

Age	Salary	STU	Credit	Buy Computer
<30	H	No	OK	No
<30	H	No	Good	No
30-40	H	No	OK	Yes
>40	M	No	OK	Yes
>40	L	Yes	OK	Yes
>40	L	Yes	Good	No
30-40	L	Yes	Good	Yes
<30	M	No	OK	No
<30	L	Yes	OK	Yes
>40	M	Yes	OK	Yes
<30	M	Yes	Good	Yes
30-40	M	No	Good	Yes
30-40	H	Yes	OK	Yes
>40	M	No	Good	No



Info信用(D)

$$\begin{aligned}
 &= \frac{6}{14} \left(-\frac{3}{6} \log \frac{3}{6} - \frac{3}{6} \log \frac{3}{6} \right) \\
 &+ \frac{8}{14} \left(-\frac{2}{8} \log \frac{2}{8} - \frac{6}{8} \log \frac{6}{8} \right) \\
 &= 0.892
 \end{aligned}$$

Gain(信用)

= Info(D) - Info信用(D)

= 0.940 - 0.892 = 0.048

Decision Tree Project



setosa

0



versicolor

1



virginica

2



KNN & K-means & Decision Tree



贵在坚持！