### 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 & KMM)

#### 作业清单 (5/13) -

٦

【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。4

学生姓名₽	小测 1₽	小测 2₽	ø
A₽	1₽	1 🕫	٦
Β ↔	2₽	1 🕫	ته
C &	4.	3 ₽	ته
D	5.	4.0	ته

## 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。

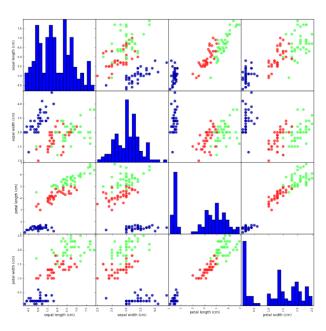
学生姓名↩	小测 1₽	小测 2₽	小测 3。	期末成绩。	项目答辩。	成绩↓
张三。	12 ₽	15 ₽	13 🕫	28 🕫	24 🕫	? ₽
李四⋄	7 ₽	11 ₽	10 🕫	19 🛭	21 0	? ₽
王五↩	12 🕫	14 🕫	11 ₽	27 ₽	23 🕫	? ₽
赵六。	6 ₽	7 ₽	4 ₽	13 🖟	20 ۵	? ₽
刘七。	13 ₽	14 ₽	13 🕫	27 ₽	25 ₽	? ₽

### 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



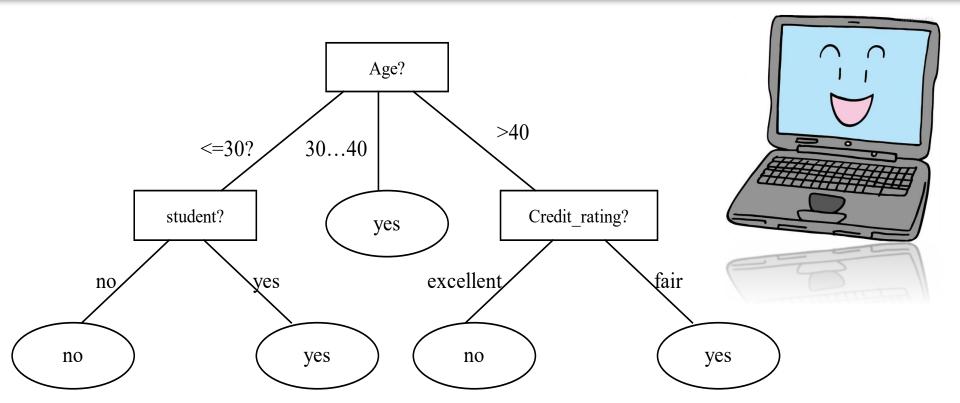
— Machine Learning 4





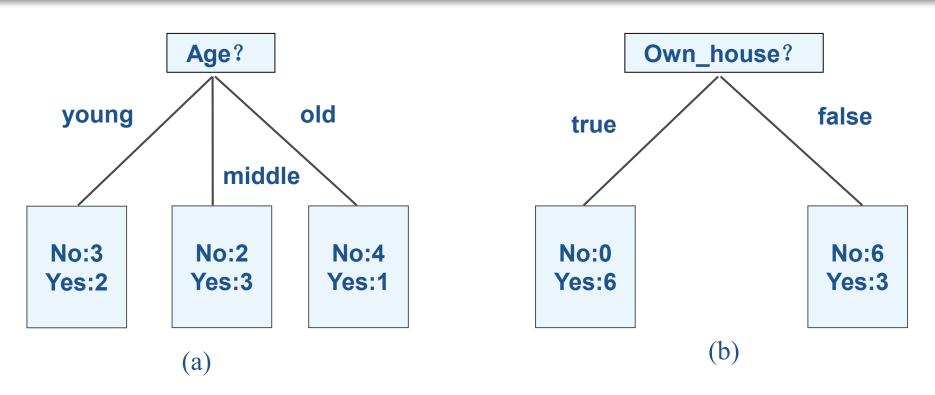
Grade	<b>Graduation Class</b>	Go Back to School in First Batch
Senior One	No	No
Senior Two	No	No
Senior Three	Yes	Yes
Junior One	No	No
Junior Two	No	No
Junior Three	Yes	Yes





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







### **Entropy**







$$H(X) = -\sum P(x) 1 \text{ o } P(\mathbf{g}x)$$

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

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

Age	Salary	STU	Credit	<b>Buy Computer</b>	
<30	Н	No	OK	No	AT INCHES
<30	Н	No	Good	No	
30-40	Н	No	OK	Yes	D =14
>40	M	No	OK	Yes	
>40	L	Yes	ОК	Yes	C1,D =5
>40	L	Yes	Good	No	C2,D =9
30-40	L	Yes	Good	Yes	
<30	M	No	OK	No	Info(D)
<30	L	Yes	ОК	Yes	$Info(D)$ = $-\frac{5}{14}\log_2\frac{5}{14} - \frac{9}{14}\log_2\frac{9}{14}$
>40	M	Yes	OK	Yes	$= -\frac{5}{14} \log_2 \frac{5}{14} - \frac{9}{14} \log_2 \frac{9}{14}$
<30	M	Yes	Good	Yes	
30-40	M	No	Good	Yes	=0.940
30-40	Н	Yes	ОК	Yes	
>40	М	No	Good	No	

Age	Salary	STU	Credit	Buy Computer	THE RESERVE TO THE PARTY OF THE
<30	Н	No	ОК	No	The same of the sa
<30	Н	No	Good	No	
30-40	Н	No	ОК	Yes	Info年龄(D)
>40	М	No	OK	Yes	5 , 3 , 3 2 , 2
>40	L	Yes	ОК	Yes	$= \frac{5}{14} \left( -\frac{3}{5} \log \frac{3}{5} - \frac{2}{5} \log \frac{2}{5} \right)$
>40	L	Yes	Good	No	4 4 4 0 0
30-40	L	Yes	Good	Yes	$+$ — $(\log\log -)$
<30	М	No	OK	No	14 4 4 4 4
<30	L	Yes	OK	Yes	$+\frac{5}{14}\left(-\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5}\right)$
>40	М	Yes	OK	Yes	14 5 5 5 5 5
<30	М	Yes	Good	Yes	=0.694
30-40	М	No	Good	Yes	Cain(在於)
30-40	Н	Yes	ОК	Yes	Gain(年龄) = Info(D) - Info年龄(D)
>40	М	No	Good	No	= 0.940 - 0.694 = 0.246

Age	Salary	STU	Credit	<b>Buy Computer</b>	
<30	Н	No	OK	No	The same of the sa
<30	Н	No	Good	No	
30-40	Н	No	OK	Yes	Info收入(D)
>40	M	No	OK	Yes	4 2 2 2 2 2 2
>40	L	Yes	OK	Yes	$= \frac{4}{14} \left( -\frac{2}{4} \log \frac{2}{4} - \frac{2}{4} \log \frac{2}{4} \right)$
>40	L	Yes	Good	No	
30-40	L	Yes	Good	Yes	$+ - ( \log \log -)$
<30	М	No	OK	No	14 6 6 6 6
<30	L	Yes	OK	Yes	$+\frac{4}{14}\left(-\frac{1}{4}\log\frac{1}{4}-\frac{3}{4}\log\frac{3}{4}\right)$
>40	M	Yes	OK	Yes	14 4 4 4 4 4
<30	М	Yes	Good	Yes	= 0.911
30-40	M	No	Good	Yes	
30-40	Н	Yes	OK	Yes	Gain(收入)
>40	M	No	Good	No	= Info(D) - Info收入(D) = 0.940 - 0.911 = 0.029
					31313 313 11

Age	Salary	STU	Credit	Buy Computer	
<30	Н	No	ОК	No	AT PARTY OF THE PA
<30	Н	No	Good	No	
30-40	Н	No	ОК	Yes	
>40	M	No	OK	Yes	Info学生(D)
>40	L	Yes	ОК	Yes	7 1 1 6 6
>40	L	Yes	Good	No	$= \frac{7}{14} \left( -\frac{1}{7} \log \frac{1}{7} - \frac{6}{7} \log \frac{6}{7} \right)$
30-40	L	Yes	Good	Yes	14 7 7 7 7
<30	M	No	OK	No	$+\frac{7}{14}\left(-\frac{4}{7}\log\frac{4}{7}-\frac{3}{7}\log\frac{3}{7}\right)$
<30	L	Yes	ОК	Yes	$14 \left( 7 \right) 7 7 7 7 7 $
>40	M	Yes	ОК	Yes	=0.788
<30	M	Yes	Good	Yes	
30-40	M	No	Good	Yes	Gain(学生)
30-40	Н	Yes	ОК	Yes	= Info(D) - Info学生(D) = 0.940 - 0.788 = 0.152
>40	M	No	Good	No	

Age	Salary	STU	Credit	<b>Buy Computer</b>	
<30	Н	No	OK	No	The state of the s
<30	Н	No	Good	No	
30-40	Н	No	OK	Yes	Info信用(D)
>40	М	No	OK	Yes	6 3 3 3 3
>40	L	Yes	OK	Yes	$= \frac{6}{14} \left( -\frac{3}{6} \log \frac{3}{6} - \frac{3}{6} \log \frac{3}{6} \right)$
>40	L	Yes	Good	No	
30-40	L	Yes	Good	Yes	$+\frac{8}{14}\left(-\frac{2}{8}\log\frac{2}{8} - \frac{6}{8}\log\frac{6}{8}\right)$
<30	М	No	OK	No	14 8 8 8 8 8
<30	L	Yes	OK	Yes	=0.892
>40	М	Yes	OK	Yes	
<30	М	Yes	Good	Yes	Gain(信用)
30-40	М	No	Good	Yes	= Info(D) - Info信用(D) = 0.940 - 0.892 = 0.048
30-40	Н	Yes	OK	Yes	
>40	M	No	Good	No	

#### Decision Tree Project



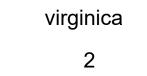


setosa



versicolor







KNN & K-means & Decision Tree





贵在坚持!