大数据分析

——期末复习

杨明达

2023年12月15日

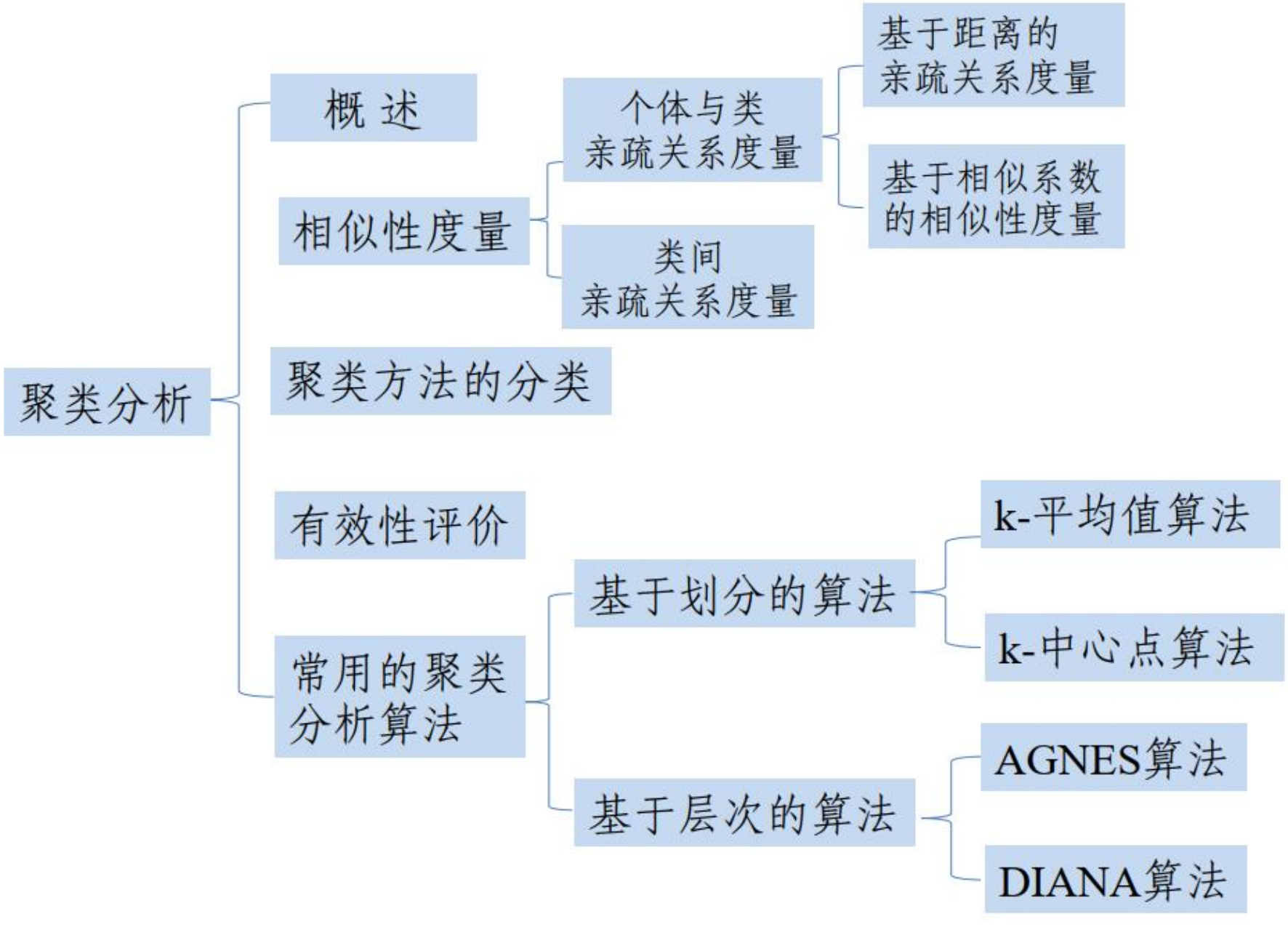
免责声明

- 一、所有内容根据个人理解总结，由于本人水平有限，所以难免会有错误，欢迎大家批评指正。
- 二、讲述内容以聚类、分类、关联三大分析中的计算题为主，不涉及这几章中的简答题。同时，挑选的题目和知识点是比较有代表性的，没讲的不代表不考，讲了的也不代表一定考。
- 三、由于既有必修同学，又有选修同学，且同学们的目前复习进度和熟练程度相差较大，所以本次讲座会照顾绝大多数同学，讲的比较慢和细致，大家如果觉得不太符合自己的情况，直接离开即可。

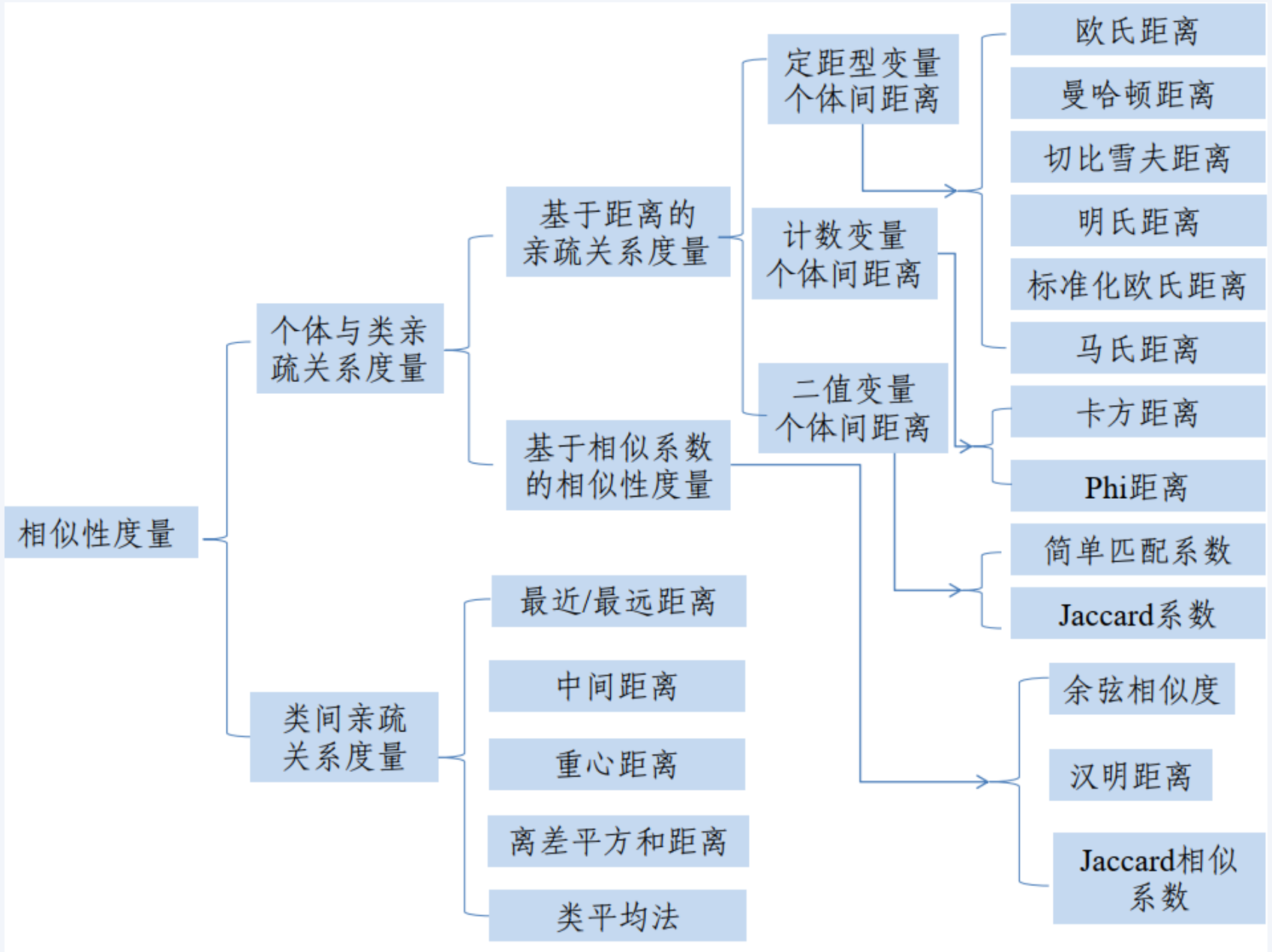
目录

一、聚类分析

聚类分析


聚类分析
概 述
相似性度量
个体与类
亲疏关系度量
基于距离的
亲疏关系度量
基于相似系数
的相似性度量
类间
亲疏关系度量
聚类方法的分类
有效性评价
常用的聚类
分析算法
基于划分的算法
k-平均值算法
k-中心点算法
基于层次的算法
AGNES算法
DIANA算法

聚类分析


相似性度量
个体与类亲疏关系度量
基于距离的亲疏关系度量
基于相似系数的相似性度量
类间亲疏关系度量
定距型变量个体间距离
计数变量个体间距离
二值变量个体间距离
欧氏距离
曼哈顿距离
切比雪夫距离
明氏距离
标准化欧氏距离
马氏距离
卡方距离
Phi距离
简单匹配系数
Jaccard系数
余弦相似度
汉明距离
Jaccard相似系数
最近/最远距离
中间距离
重心距离
离差平方和距离
类平均法

基于距离的个体与类亲疏关系度量：

定距型变量个体间距离：

欧氏距离

曼哈顿距离

切比雪夫距离

明氏距离

标准化欧式距离

马氏距离

计数变量个体间距离：

卡方距离

Phi距离

二值变量个体间距离：

简单匹配系数

Jaccard系数

聚类分析

计算向量（0,0），（1,0），（0,2）两两间的欧式距离、曼哈顿距离、切比雪夫距离、标准化欧式距离，假设两个分量的标准差分别为0.5和1

聚类分析

已知二维正态总体G的分布为：$G\sim N(\mu, \Sigma)$，其中

$$\boldsymbol{\mu}=\begin{pmatrix}0\\0\end{pmatrix}, \boldsymbol{\Sigma}=\begin{pmatrix}1 & -0.9\\-0.9 & 1\end{pmatrix}$$

分别求点$A=(1, 1)^{\mathrm{T}}$和点$B=(1, -1)^{\mathrm{T}}$到均值μ的欧氏距离和马氏距离。

基于相似系数的个体与类亲疏关系度量：

余弦相似度

汉明距离

Jaccard相似系数

Pearson相关系数

聚类分析

求向量（3,2,0,5,0,0,0,2,0,0）和向量（1,0,0,0,0,0,0,1,0,2)的余弦相似度

聚类分析

求集合X={1,2,3,4}和集合Y={3,4，5,6}的Jaccard相似系数

例子：有两个物品A, B，调查7位用户是否购买了这两样物品，得以下向量：A=(0,0,1,1,1,0,1)，B=(1,0,1,0,1,0,0)

忽略0-0匹配，求A和B的Jaccard距离

聚类分析

例1：计算压力x和压缩量y之间的相关系数r。

表 2-3　绝缘材料的压缩量和压力表

压力 x(10 lb/in^2)	压缩量 y(0.1 in)
1	1
2	1
3	2
4	2
5	4

类间亲疏关系度量:

最近/最远距离

中间距离

聚类分析

分别采用最近距离、最远距离、中间距离对

下面6个点进行层次聚类，采用欧式距离度量

数据集		
	x	y
x_1	1	1
x_2	2	1
x_3	1	3
x_4	4	1
x_5	4	4
x_6	5	4

	$C_1=\{x_1\}$	$C_2=\{x_2\}$	$C_3=\{x_3\}$	$C_4=\{x_4\}$	$C_5=\{x_5\}$	$C_6=\{x_6\}$
$C_1=\{x_1\}$	0					
$C_2=\{x_2\}$	1	0				
$C_3=\{x_3\}$	4	5	0			
$C_4=\{x_4\}$	9	4	13	0		
$C_5=\{x_5\}$	18	13	10	9	0	
$C_6=\{x_6\}$	25	18	17	10	1	0

聚类分析

（1）最近距离

	$C_1=\{x_1\}$	$C_2=\{x_2\}$	$C_3=\{x_3\}$	$C_4=\{x_4\}$	$C_5=\{x_5\}$	$C_6=\{x_6\}$
$C_1=\{x_1\}$	0					
$C_2=\{x_2\}$	1	0				
$C_3=\{x_3\}$	4	5	0			
$C_4=\{x_4\}$	9	4	13	0		
$C_5=\{x_5\}$	18	13	10	9	0	
$C_6=\{x_6\}$	25	18	17	10	1	0

聚类分析

（2）最远距离

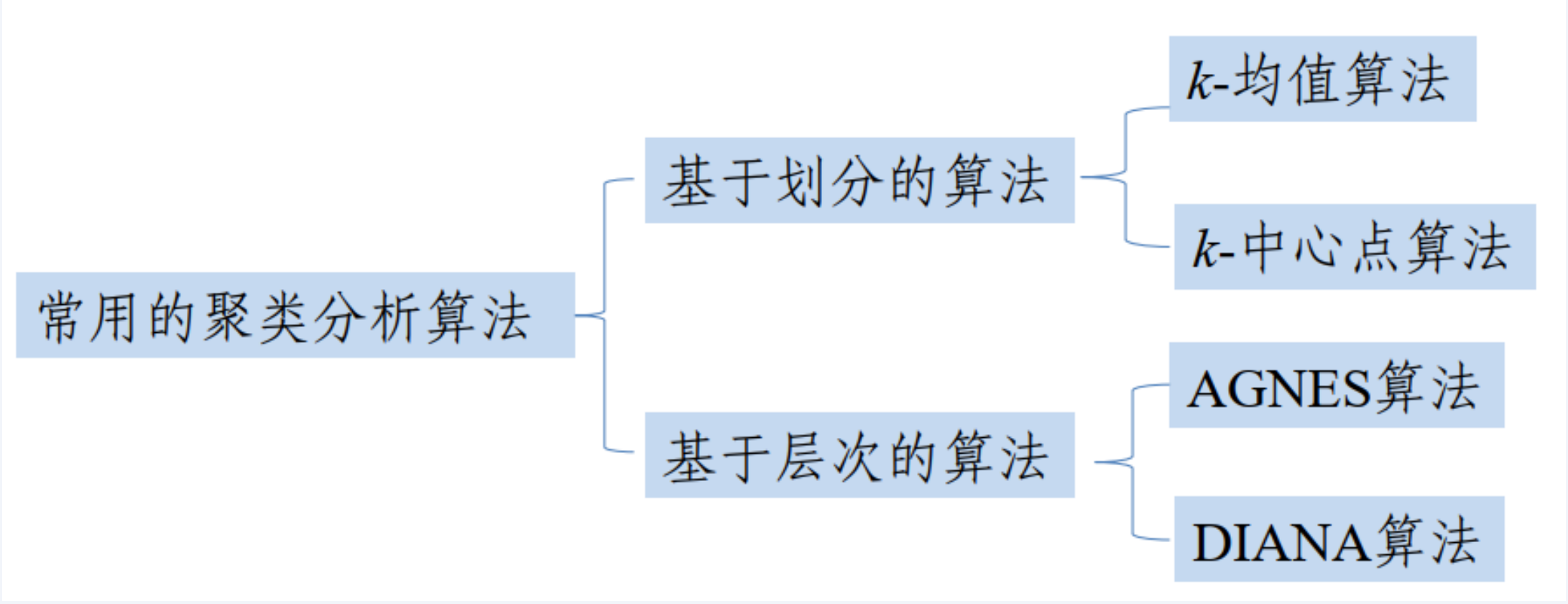
	$C_1=\{x_1\}$	$C_2=\{x_2\}$	$C_3=\{x_3\}$	$C_4=\{x_4\}$	$C_5=\{x_5\}$	$C_6=\{x_6\}$
$C_1=\{x_1\}$	0					
$C_2=\{x_2\}$	1	0				
$C_3=\{x_3\}$	4	5	0			
$C_4=\{x_4\}$	9	4	13	0		
$C_5=\{x_5\}$	18	13	10	9	0	
$C_6=\{x_6\}$	25	18	17	10	1	0

聚类分析

（3）中间距离

	$C_1=\{x_1\}$	$C_2=\{x_2\}$	$C_3=\{x_3\}$	$C_4=\{x_4\}$	$C_5=\{x_5\}$	$C_6=\{x_6\}$
$C_1=\{x_1\}$	0					
$C_2=\{x_2\}$	1	0				
$C_3=\{x_3\}$	4	5	0			
$C_4=\{x_4\}$	9	4	13	0		
$C_5=\{x_5\}$	18	13	10	9	0	
$C_6=\{x_6\}$	25	18	17	10	1	0

聚类分析



基于划分的算法

K-Means

K-Medoids

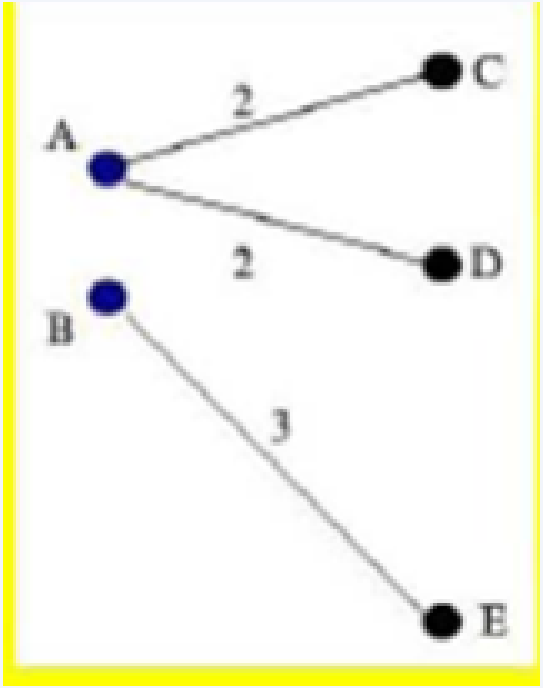
聚类分析

利用K-平均值算法对{2,4,10,12,3,20,30,11,25}进行聚类，K=2，初始聚类中心为2和4。

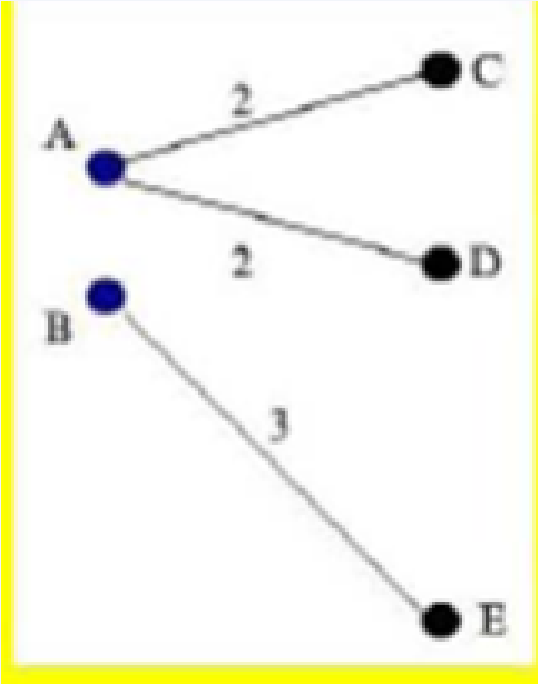
聚类分析

利用K-中心点算法对{A，B，C，D，E}进行聚类，各点之间距离如图所示，K=2，初始聚类中心为A和B。请计算第一轮尝试替换聚类中心产生的代价。

样本点	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0



聚类分析



基于层次的算法

AGNES算法——自底向上

DIANA算法——自顶向下

聚类分析

例子：有如下表所示的数据集，使用DIANA算法对该数据集进行分裂层次聚类。

序号	属性 1	属性 2
1	1	1
2	1	2
3	2	1
4	2	2
5	3	4
6	3	5
7	4	4
8	4	5

对于所给的数据进行DIANA算法，(设n=8,用户输入的终止条件为2个类)，初始类{1,2,3,4,5,6,7,8}。

聚类分析

0						
1	0					
1	1.4	0				
1.4	1	1	0			
3.6	2.8	3.2	2.2	0		
4.5	3.6	4.1	3.2	1	0	
4.2	3.6	3.6	2.8	1	1.4	0
5	4.2	4.5	3.6	1.4	1	1	0

序号1的平均距离（就是1距离其它各个点的距离长度之和除以7）
s1=(1+1+1.1414+3.6+4.47+4.24+5)/7=2.96;
序列2的平均距离 s2=(1+1.414+1+2.828+3.6+3.6+4.24)/7=2.526;
序列3的平均距离 s3=(1+1.414+1+3.16+4.12+3.6+4.27)/7=2.68;
序列4的平均距离 s4=(1.414+1+1+2.24+3.16+2.828+3.6)/7=2.18
序列5的平均距离 s5=2.18;
序列6的平均距离 s6=2.68;
序列7的平均距离 s7=2.526;
序列8的平均距离 s8=2.96;

聚类分析

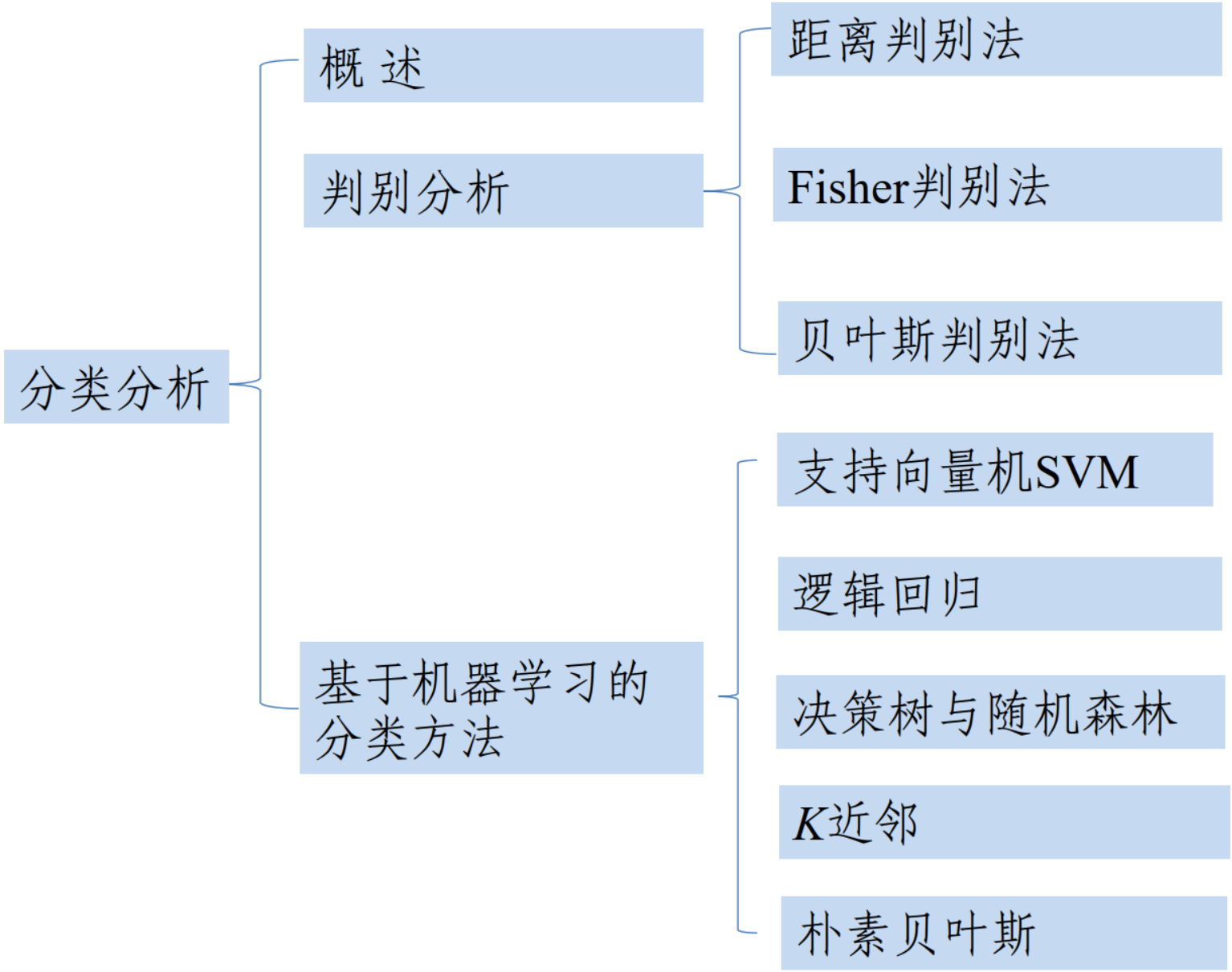
0					
1	0				
1	1.4	0			
1.4	1	1	0		
3.6	2.8	3.2	2.2	0	
4.5	3.6	4.1	3.2	1	0
4.2	3.6	3.6	2.8	1	1.4
5	4.2	4.5	3.6	1.4	1

0						
1	0					
1	1.4	0				
1.4	1	1	0			
3.6	2.8	3.2	2.2	0		
4.5	3.6	4.1	3.2	1	0	
4.2	3.6	3.6	2.8	1	1.4	0
5	4.2	4.5	3.6	1.4	1	1

0						
1	0					
1	1.4	0				
1.4	1	1	0			
3.6	2.8	3.2	2.2	0		
4.5	3.6	4.1	3.2	1	0	
4.2	3.6	3.6	2.8	1	1.4	0
5	4.2	4.5	3.6	1.4	1	1	0

二、分类分析

分类分析


分类分析
概 述
判别分析
距离判别法
Fisher判别法
贝叶斯判别法
基于机器学习的
分类方法
支持向量机SVM
逻辑回归
决策树与随机森林
K近邻
朴素贝叶斯

判别分析:

距离判别法——马氏距离

Fisher判别法

贝叶斯判别法

协方差矩阵不相等的距离判别法：

例子： 已知有两个类G_1和G_2，分别为设备A、B生产的产品。设备A生产的产品平均耐磨度$\mu_1=80$，精度$\sigma_1^2=0.25$；设备B的平均耐磨度$\mu_2=75$，精度$\sigma_2^2=4$。现有一耐磨度为78的产品x，试判断它为哪一台设备生产的。

协方差矩阵相等的距离判别法：

先明确以下概念/公式：

（1）样本离差阵

（2）样本合并组内离差阵

（3）合并样本协差阵

（4）判别函数

分类分析

例1： 记二维正态总体$N_i(\mu^{(i)}, \sum)$为$G_i\,(i=1, 2)$（两总体协差阵相同），已知来自$G_i\,(i=1, 2)$的样本数据为

$$X^{(1)}=\begin{pmatrix}2 & 12\\ 4 & 10\\ 3 & 8\\ 3 & 10\end{pmatrix}, X^{(2)}=\begin{pmatrix}5 & 7\\ 3 & 9\\ 4 & 5\end{pmatrix} \quad \begin{pmatrix}k=2, & m=2\\ n_1=4, & n_2=3\end{pmatrix}$$

（1）试求两总体的样本离差阵S_1, S_2和合并样本协差阵S。

（2）今有样本$x_0=(2, 8)'$，试问按马氏距离准则样本x_0应判归哪一类。

机器学习方法的分类分析:

决策树

KNN

Naïve Bayes

决策树

先明确以下概念/公式：

（1）信息熵

（2）信息增益

（3）增益率

分类分析

利用ID3算法，根据以下数据构建决策树。

表 4-1 高尔夫活动决策表

编号	天气	温度	湿度	风速	活动
1	晴	炎热	高	弱	取消
2	晴	炎热	高	强	取消
3	阴	炎热	高	弱	进行
4	雨	适中	高	弱	进行
5	雨	寒冷	正常	弱	进行
6	雨	寒冷	正常	强	取消
7	阴	寒冷	正常	强	进行
8	晴	适中	高	弱	取消
9	晴	寒冷	正常	弱	进行
10	雨	适中	正常	弱	进行
11	晴	适中	正常	强	进行
12	阴	适中	高	强	进行
13	阴	炎热	正常	弱	进行
14	雨	适中	高	强	取消

可能用到的数据：

$$\frac{2}{5}\log\frac{2}{5}+\frac{3}{5}\log\frac{3}{5}=-0.971$$

$$\frac{2}{3}\log_2\frac{2}{3}+\frac{1}{3}\log_2\frac{1}{3}=-0.918$$

$$\frac{5}{14}\log_2\frac{5}{14}+\frac{9}{14}\log_2\frac{9}{14}=-0.940$$

$$\frac{1}{4}\log_2\frac{1}{4}+\frac{3}{4}\log_2\frac{3}{4}=-0.811$$

$$\frac{3}{7}\log_2\frac{3}{7}+\frac{4}{7}\log_2\frac{4}{7}=-0.985$$

$$\frac{6}{7}\log_2\frac{6}{7}+\frac{1}{7}\log_2\frac{1}{7}=-0.592$$

分类分析

表 4-1 高尔夫活动决策表

编号	天气	温度	湿度	风速	活动
1	晴	炎热	高	弱	取消
2	晴	炎热	高	强	取消
3	阴	炎热	高	弱	进行
4	雨	适中	高	弱	进行
5	雨	寒冷	正常	弱	进行
6	雨	寒冷	正常	强	取消
7	阴	寒冷	正常	强	进行
8	晴	适中	高	弱	取消
9	晴	寒冷	正常	弱	进行
10	雨	适中	正常	弱	进行
11	晴	适中	正常	强	进行
12	阴	适中	高	强	进行
13	阴	炎热	正常	弱	进行
14	雨	适中	高	强	取消

分类分析

例1：给出如表所示的训练样本，目的是判定一个人是否会购买电脑。这个人的属性为X=（年龄<=30，收入=中等，学生=是，信用率=一般）。**使用朴素贝叶斯算法。**

编号	年龄	收入	学生	信用等级	类别：购买电脑
1	<=30	高	否	一般	不会购买
2	<=30	高	否	良好	不会购买
3	31…40	高	否	一般	会购买
4	>40	中等	否	一般	会购买
5	>40	低	是	一般	会购买
6	>40	低	是	良好	不会购买
7	31…40	低	是	良好	会购买
8	<=30	中等	否	一般	不会购买
9	<=30	低	是	一般	会购买
10	>40	中等	是	一般	会购买
11	<=30	中等	是	良好	会购买
12	31…40	中等	否	良好	会购买
13	31…40	高	是	一般	会购买
14	>40	中等	否	良好	不会购买

三、关联分析

关联分析

GBDT树：

GBDT学习例子：训练集: (A, 14岁), (B, 16岁), (C, 24岁), (D, 26岁)。训练数据的均值为：20岁；决策树的个数为：2棵。每个样本的特征有两个：购买金额是否小于1K，经常去百度提问还是回答？

其中：

A:14岁，购物金额<=1K，经常去百度提问；

B:16岁，购物金额<=1K，经常去百度回答；

C:24岁,购物金额>1K，经常去百度提问；

D:26岁,购物金额>1K，经常去百度回答

关联规则分析及Apriori算法

先明确以下概念/公式：

（1）事务、项、K-项集

（2）支持度

（3）频繁项集

（4）置信度

（5）提升度

（6）兴趣因子

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

关联分析

根据下表数据（项按字典序存储），利用Apriori算法进行关联规则分析。支持度阈值为2，置信度阈值为70%，求出频繁相集，最高频繁相集产生的关联规则。

TID	商品ID的列表
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

关联分析

TID	商品ID的列表
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

致谢

- 感谢杨东华老师一学期的辛苦付出
- 感谢同学们的信任与支持

杨明达

2023年12月15日