**课后练习**

**一、简答题**

1. 如何测度样品和变量间的相似性？计算样品之间的距离有哪些公式？它们各有什么特点？

2. Q型聚类法和R型聚类法有什么异同？

3. 简述系统聚类法的基本思想及主要步骤。

4. 简述系统聚类分析的优缺点。

5. Q型系统聚类法包括哪几种方法，有什么特点。

6. 简述动态聚类法的基本思想与步骤。

**二、计算题**

1.设有五位销售员，他们的销售业绩由二维变量表示，

|  |  |  |
| --- | --- | --- |
| **销售员** | **销售量（百件）** | **回收款项（万元）** |
| **1** | 1 | 0 |
| **2** | 2 | 1 |
| **3** | 3 | 2 |
| **4** | 4 | 2 |
| **5** | 3 | 5 |

（1） 用欧氏距离计算样品之间的距离，结果用距离矩阵表示。

（2） 采用最短距离法、最长距离法对这5位销售员进行分类。写出聚类过程，并画聚类谱系图。

2. 某公司打算选拔市场营销部门的负责人，现考察5个候选人在10个方面的具体表现，通过评分的方式，经计算得到如下的欧氏距离矩阵D：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **②** | **③** | **④** | **⑤** |
| **①** | 5.2 | 8.6 | 28.2 | 26.1 |
| **②** |  | 43.6 | 27.4 | 36.2 |
| **③** |  |  | 7.5 | 15.1 |
| **④** |  |  |  | 6.5 |

要求：

（1）采用最短距离法对这5个候选人进行分类。写出聚类的计算过程，并画聚类谱系图。

（2）若根据（1）的结果，将5位侯选人分为两类，请给出分类结果。并在此基础上，采用重心法计算两大类之间的距离。

3. 5个地区的两个创新指标数据如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **指标** | **①** | **②** | **③** | **④** | **⑤** |
| **R&D强度** | 5.54 | 3.12 | 1.96 | 1.18 | 1.65 |
| **每万人发明专利授权数** | 7.88 | 3.90 | 1.74 | 0.68 | 0.86 |

规定用欧氏距离计算样品间的距离，经计算得到距离矩阵如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **②** | **③** | **④** | **⑤** |
| **①** | 4.66 | 7.11 | 8.42 | 8.03 |
| **②** |  | 2.45 | 3.76 | 3.38 |
| **③** |  |  | 1.32 | 0.93 |
| **④** |  |  |  | 0.25 |

已知5个地区按创新能力的两个指标，可分成以下两类：

G1＝｛①、②｝，G2＝｛③、④、⑤｝

请分别用最短距离法、最长距离法、重心法、类平均法计算G1和G2之间的距离（提示：按照上述各距离的定义计算）。

**三、上机分析题**

1.数据集EXE3\_1是反映全国30个地区高质量发展的相关指标，请利用聚类分析对我国各地区高质量发展状况进行分析。

|  |  |  |  |
| --- | --- | --- | --- |
| **一级指标** | **二级指标** | **三级指标** | **属性** |
| 经济高质量发展 | 发展动力 | 每万人年R&D人员数(人年/万人） | + |
| R&D经费投入强度（%） | + |
| 每百家企业拥有网站数（个） | + |
| 每万人年R&D人员发明专利授权数（件/万人年） | + |
| 人均技术市场成交额（元/人） | + |
| 有研发机构的规上工业企业单位数占比（%） | + |
| 大学本科及以上学历就业人数占比（%） | + |
| 发展效率 | 劳动生产率（元/人） | + |
| 亩均GDP (亿元/万亩) | + |
| 单位电能创造的GDP（亿元/千瓦小时） | + |
| 结构优化 | 高技术产业主营业务收入占比（%） | + |
| 高新技术企业营业收入占比(%) | + |
| 规上工业企业新产品销售收入占主营业务收入的比重(%) | + |
| 发展效益 | 人均地区生产总值（元/人） | + |
| 居民人均可支配收入（元） | + |
| 居民人均消费支出（元/人） | + |
| 绿色发展 | 单位GDP产生的二氧化硫及氮氧化物排放量（吨/亿元） | — |
| 城市污水处理率（%） | + |
| 城市生活垃圾无害化处理率（%） | + |
| 建成区绿化覆盖率（%） | + |

2. 数据集EXE3\_2是全球GDP排名前30名国家的低碳社会发展状况，请利用该数据集对各国的低碳社会发展状况进行聚类分析。

|  |  |  |  |
| --- | --- | --- | --- |
| **一级指标** | **观测指标** | **单位** | **指标类型** |
| 碳生产发展水平 | X1 人均碳排放 | 吨/人 | 逆指标 |
| X2 单位GDP碳排放 | 吨/万元 | 逆指标 |
| X3 碳排放强度 | 吨/千克 | 逆指标 |
| 社会发展水平 | X4 第三产业增加值占GDP比重 | % | 正指标 |
| X5 基尼系数 | % | 适度指标 |
| X6 人均森林面积 | 平方米/人 | 正指标 |
| X7 教育经费占GDP比重 | % | 正指标 |
| X8 高等教育入学率 | % | 正指标 |
| X9 每千人医院床位数 | 张/千人 | 正指标 |