应用场景

我们在没有明确的类别的情况下会使用聚类。聚类模型试图将数据集中的样本划分为若干个通常是不相交的子集，每个子集称为一个“簇”（cluster），簇所对应的概念语义需由使用者来把握和命名。

1. 聚类模型在金融方面的应用

例如为了给金融商品画像做准备而进行的聚类分析。

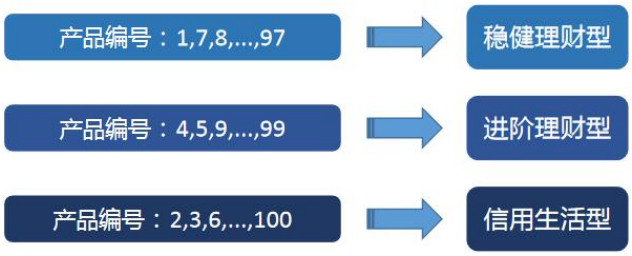
某金融投资公司有 100 样理财产品，如下所示：

编号	类型	售价	用途	产品期限	对象	单位净值
1	个人养老保障产品	1000元起购	养老保障	灵活申赎	非老年人	
2	指数型产品	100元起购	定投优选	29天	所有人群	
3	实物黄金	1万元起购	避险	便携自提	中产阶级	
...	...	...	...	...	...	
100	趋势投资混合型	500元起购	定投优选	3个工作日	所有人群	

可以看到，产品特征向量为 X=(类型、售价、用途、产品期限、对象、...)，根据上表中公司每个产品的特征向量去聚类，最终聚成了 3 类，即该公司的理财产品大致可区分成 3 类。

为此我们不仅可以通过每类的总体特征，判断出该公司的主营产品、擅长产品等，同时结合每类产品的特征进行产品战略设定。

例如经过聚类后的 3 类产品：



分别设定为稳健理财型、进阶理财型、信用生活型，下一步递交给品牌宣传部或市场部进行公司经营战略制定。

2. 聚类模型在电商方面的应用

例如为了给电商公司的店铺画像做准备而进行的聚类分析。

某电商公司有 100 间入驻店铺，如下所示：

店铺编号	店铺名称	类型	商品数	商品平均价格	月均销量	月均新增收藏量
1	海风潮ORIGINAL	男装	364	183.16	752	5413
2	喔喔食品旗舰店	零食	30	14.32	2418	4894
3	飞鸟户外用品	户外	16	75.4	22	357
...	...	...	...	...	...	...
100	山里遇农夫	生鲜	118	16.72	265	1081

可以看到，产品特征向量为 X=(类型、商品数、商品平均价格...)，根据上表中店铺特征向量去聚类，最终聚成了 6 类，即该电商公司的店铺可区分成 6 类。

为此我们不仅通过每类的总体特征，判别出该电商的行业特点、店铺类型等，同时结合每类店铺的特征安排营销策略的制定。

例如经过聚类后的 6 类店铺：

分别设定为货源优势型、特色优势型、创意优势型、价格优势型、地域优势型、薄利多销型，下一步递交给战略规划部或市场推广部进行店铺经营战略的制定。

3. 聚类模型在教育方面的应用

例如为了给学生画像做准备而进行的聚类分析。

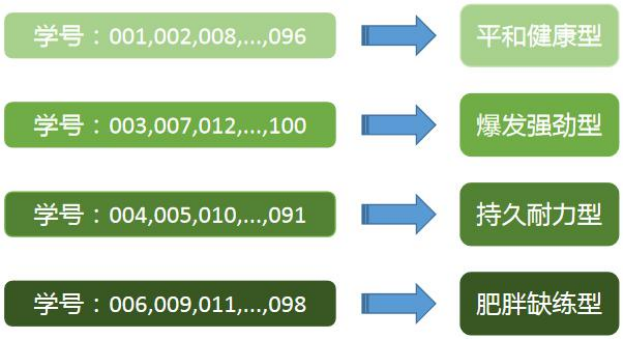
某教室里有 100 个学生，学生信息如下所示：

学号	姓名	性别	年龄	身高	体重	跳高	跳远	肺活量	握力	引体向上
001	陈明	男	22	176	65	71	2.36	4033	29.6	20
002	李华	女	23	165	48	49	1.93	4020	19.2	3
003	王晶	女	22	162	46	60	2.07	3918	19.6	8
...	...	...	...	...	...	...	...	...	...	...
100	赵强	男	21	178	71	82	2.42	6941	31.4	7

可以看到，学生特征向量为 X=(性别、年龄、身高、体重...)，根据上表中学生特征向量去聚类，最终聚成了 4 类，即该教室的学生大致可区分成 4 类。

为此我们不仅通过每类的总体特征，判别出该群学生的肥胖情况、健康情况等，同时结合每类学生的特征安排体育课程。

例如经过聚类后的 4 类学生：



分别设定为平和健康型、爆发强劲型、持久耐力型、肥胖缺练型，下一步递交给体育老师或健康专家安排体育课程内容。

k-means 模型

在各种聚类算法中，k-means 算法可以说是最简单的，资源消耗非常小。尽管代价小，看似简单，但 k-means 算法却出奇地高效，运行速度非常快，收敛速度通常也很快。它在很多数据挖掘问题上效果很好，因此成为聚类模型中最为经典和流行的一种，使用最为广泛。本课程的案例中主要选用了 k-means 算法。

1. 思想

- 在样本数据 N 中，样本特征向量为 X，存在 S 类的簇，则每个样本存在标签 T 属于 S 中的一个。然而，S 是未知的，T 也是未知的。
- 给定一个超参数 k，输入某样本特征向量 X，根据模型算法计算得出该样本的 Y，此 Y 即为标签属于 k 中的一个。目的是希望 Y=T。
- 它属于聚类，无监督学习，也就是 Y 是没有参考结果。(代表需要事后根据业务理解去解释结果的优劣，属于定性解释)

注：关于 k-means 模型的详细介绍请查看：数据分析入门课程/数据建模/2.挖掘初始模型/2.5k-means 聚类分析-1

2. 案例应用

在本次课程中，我们一共使用 k-means 模型解决了以下问题：

1) Python 数据分析（入门）

传统的对历代官员类别的区分多是历史专家根据相关史料人为进行梳理，得出结论，如忠臣、奸臣等。一方面，专家的精力有限，这样传统的判别方法无法应用到所有有记载的官员上；另一方面，传统的判别方法多是定性的，很少会定量地考虑不同类别之间的区别。我们是否可以仅仅根据历史官员自身的特征，通过数据分析将这些官员归入不同的类别呢？

在这个案例中，我们获得了中国历代人物传记资料库（CBDB）中收录的宋朝部分官员的政治关系记录和亲属关系记录，其中涉及到的官员特征有:姓名、指数年、性别、社会关系人姓名、社会关系人指数年、社会关系人性别、联系、籍贯、关系人籍贯、地理位置、ID、关系人 ID、亲戚姓名、亲属关系等等。经过对官员特征的一系列探索后，发现官员存在 6 种不同的政治关系（“不合”、“反对/攻讦”、“得到 Y 的支持”、“支持”、“政见趋同”、“遭到 Y 的反对/攻讦”），另外还得到了官员的“亲戚为官数”。通过对官员的这 7 个特征使用 k-means 模型进行聚类，并对 k 的不同取值下聚类结果的可解释性判断，发现聚成 4 类时效果最佳，最终将这些宋朝官员分为 4 类：非活跃型、活跃负面型、活跃争议型、平缓争议型。

通过官员自身的特征对有记载的官员进行聚类分析既可以弥补传统官员判别方法的不足，也是对传统方法的一种验证及补充。

2) 数据分析初级：人物关系案例一

汗青影视制作公司正在筹划拍摄一部关于宋朝官员的历史纪录片，该纪录片的题材主要是针对宋朝党政、流派频繁的朝野生活，但是没有合适的官员人选名单，希望通过分析 CBDB 的传记数据，寻找出一批具有较高政治领导力或者有特色的官员。该如何通过聚类的方法进行筛选呢？

首先官员的政治领导力体现在机构领导力、事件领导力和个人领导力三个方面，通过进一步分解，分别划分为议题树中的子问题：官员角色等级（机构）、机构级别、官员角色等级（事件）、事件级别和个人数量。除了个人数量可以通过在关系表中对关系人的 ID 进行分组计数得到，其余的四个字段都需要结合官职树编码表中的官僚机构级别(office_l1、office_l2、office_l3)、宋朝事件表中的事件中文名称(event_name_chn)

和扮演角色（role），以及查询宋朝历史政治文化和咨询相关专家的方式来人为赋值得到。最终得到官员的机构领导力、事件领导力和个人领导力。然后通过对这三个特征进行聚类分析，在结合了轮廓系数评价指标和聚类结果的可解释性评判后，将官员聚成了4类。先从4类中选择“高领导力”官员，再从其中筛选机构领导力取到最大值，个人领导力排名前15的官员有：秦桧、蔡京、文天祥、王安石、章惇、赵普、韩琦、张浚、史弥远、张商英、司马光、韩绛、文彦博、富弼、张邦昌，为记录片的制作提供参考。具体的人数可以根据纪录片的内容继续拓展。除此之外，还可以考虑纳入事件领导力最高的官员，即章堉。

对于汗青影视制作公司而言，通过聚类分析，在尽可能符合历史事实的基础上，高效、低成本地获取纪录片中的官员名单，也保证了素材内容有深度、有吸引力。

3) 数据分析初级：人物关系案例二

在得到了政治领导力较为突出的官员后，汗青影视制作公司还希望通过分析CBDB的传记数据，寻找出一批具有较大政治权力的官员，来作为纪录片的制作参考。该如何筛选呢？

官员的政治权力体现为直接权力和间接权力。直接权力即为官员自身的官职权力，我们通过咨询专家建议根据官员所在官僚机构级别（office_l1、office_l2、office_l3）进行人为赋值；间接权力又称裙带权力，体现为家族裙带权力、婚姻裙带权力、朋友裙带权力和师生裙带权力。其中，家族裙带权力和婚姻裙带权力我们分别通过对亲戚表和姻亲表中为官亲戚的官职权力加和求得；朋友和师生关系都在社交表中，对相应的为官关系人的官职权力求和，分别得到朋友裙带权力和师生裙带权力。

随后对官职权力和4种裙带权力进行聚类分析，在结合了轮廓系数评价指标和聚类

结果的可解释性评判后，确定聚成 5 类时结果最佳。5 类里有 3 类的官职权力都很高，但是裙带权力各有突出。在官职权力最高、4 种裙带权力均为 0 的一类中筛选了共 111 位，称为“自我奋斗型”官员；在官职权力和婚姻裙带权力最高的一类中筛选了有吕公著、张士逊、苏颂等 10 位官员，称为“婚姻裙带型”官员；在官职权力和家族裙带权力最高的一类中筛选了有吕蒙正、宋祁、韩绛等 10 位官员，称为“家族裙带型”官员；除此之外，还可以考虑纳入 5 种权力最高的官员。排除重复的人物，再添加朱熹（他的师生裙带权力最大），为纪录片的制作提供参考。

4) 数据分析初级：人物关系案例三

汗青影视制作公司筹划组希望将思路放宽，不仅局限于政治领域，从社会层面上去讲述有突出影响力的官员故事。希望数据分析人员可以继续利用 CBDB 寻找出宋朝哪些政治官员的社会影响力较大。

官员的社会影响力主要从时间影响力、空间影响力和互动影响力三方面来体现。在构建议题树和绘制了 E-R 图后，我们将官员的仕途年限作为官员的时间影响力，地点跨度与地点数量乘积作为官员的空间影响力，同时从政治关系、学术关系、社交关系和著述关系四个方面来反映官员在社会中的互动影响力。时间影响力和空间影响力可以根据现有字段转换得到，但互动影响力需要对四种互动影响力进行降维，刚好可以利用聚类的方式，将这四种互动影响力降维到一个字段上。随后对三种社会影响力使用 k-means 模型进行聚类，发现聚成两类时效果最佳，一类三种社会影响力取值都较小，另一类都较大。将其中三种社会影响力都较大的一类作为社会影响力较大的官员名单，提供给纪录片制作人员，包括范仲淹、欧阳修、曾巩、苏轼、范祖禹等 24 名宋朝官员。

在这个案例中，就充分体现了聚类的作用，它不仅可以在不破坏数据隐含内容的前

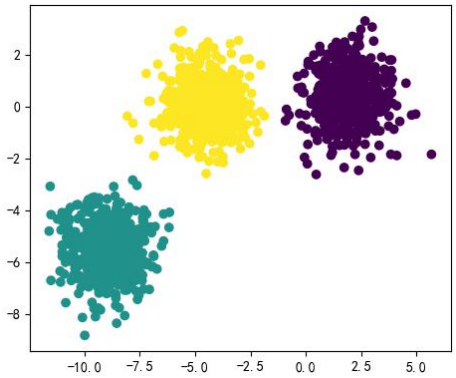
提下，将多维的数据降维到一维上，同时还可以基于多个角度筛选出具有高质量的纪录片素材内容，供汗青影视制作部门人员参考。

最后通过查看历史，发现对 CBDB 数据聚类分析后筛选出来的官员，从不同层面上都是当时社会中举足轻重的人物，极大地减少了专业人士对历史人物的人工评估。

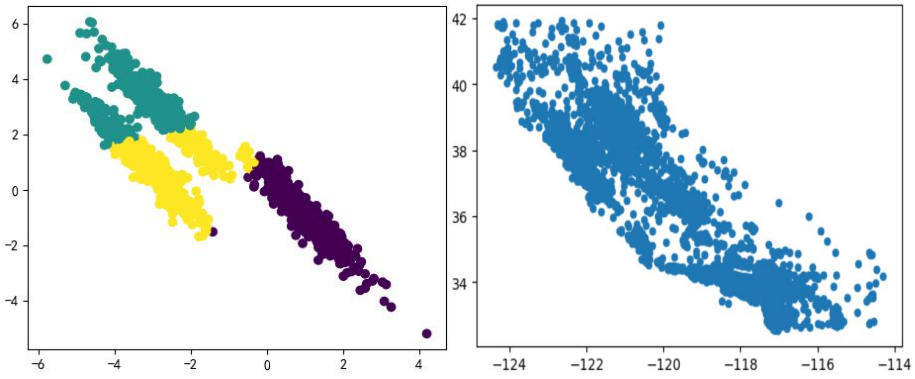
3. 局限

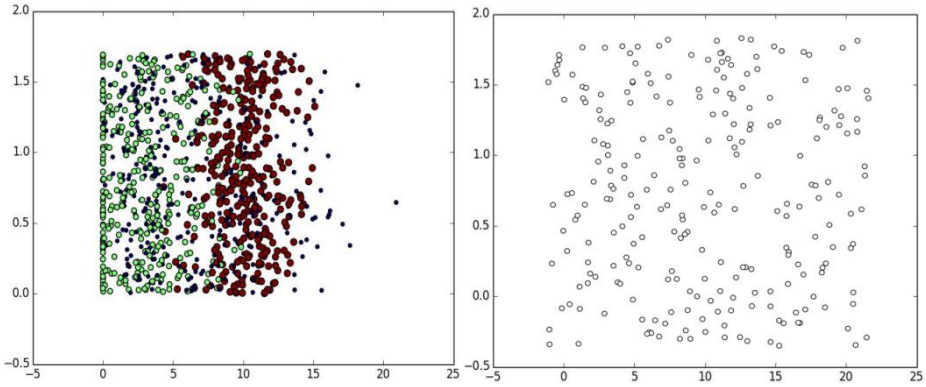
1) 数据的分布

适用于圆形凸状的输入数据分布，例如：



而其他类型的分布，模型效果较差，如下（但不限于下列数据分布形状）：



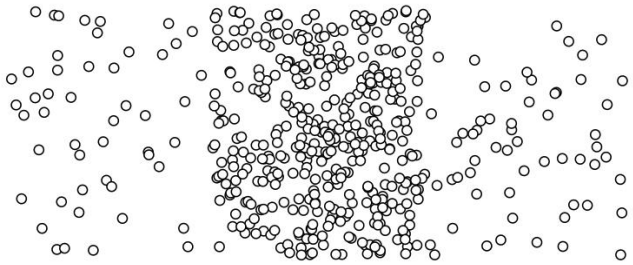


2) 异常值

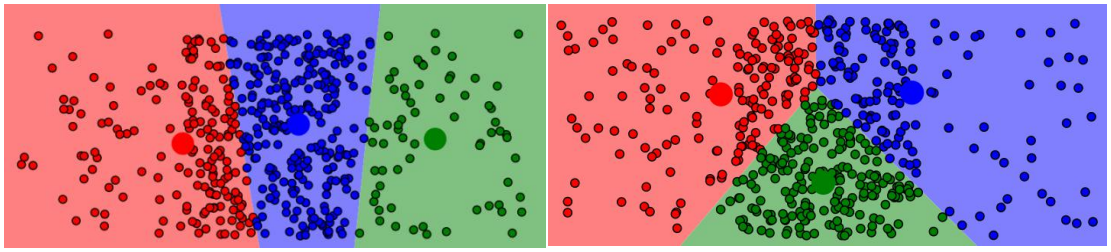
对噪声、孤立点的样本数据敏感，会影响簇的结构。

3) 初始点的选择

初始点的选择对于模型聚类结果也有较为重要的影响，例如下列数据分布：



在两种不同初始点选择下的不同聚类结果：



因此该如何选择初始点？分为以下几步：

- 随机选择一个点，做为第一个中心点；
- 选取离第一个中心点最远的一个点做为第二个中心点；
- 选取离第一个和第二个中心最远的点，做为第三个中心；

- 依此计算后续的中心点。

注：初始点的选择在 KMeans()参数 init 中进行设置。

4) k 值的选择

尽管可以根据评价指标选取最佳 k 值（常用的有肘部法则中的误差平方和、轮廓系数、CH 指数等），但仅根据指标在数值上对 k 值进行调整，对实际场景的意义不大，反而需要人工依据业务理解去评判最佳 k 值下聚类结果的可解释性。

注：

①肘部法则：随着 k 值的增大，平均畸变程度会逐渐减小。k 值增大的过程中，平均畸变程度下降幅度最大的位置对应的 k 值就是肘部，选择此时的 k 值用于聚类比较合适。

----------详见：入门/项目/数据建模的提高题

②畸变程度：每个类的畸变程度是指这一类的中心与其内部数据点距离的平方和，它是衡量类内部成员彼此间紧凑程度的指标。若畸变程度越小，表示聚类的结果中内部成员之间更为紧凑，聚类效果越好。

③轮廓系数：轮廓系数是通过考察群（簇）间的分离情况和群内紧凑情况来评价聚类结果的一种方法，它结合内聚度和分离度两种因素，用来评价不同算法对聚类结果所产生的影响。

----------详见：初级 1/数据建模/1.k-means 聚类/1.4 轮廓系数介绍

③CH 指数：Calinski-Harabasz 指标，它是从簇内的稠密程度和簇间的离散程度来评估聚类的效果好坏，CH 值越大则聚类效果越好。

----------可以通过 metrics.calinski_harabaz_score 实现

4. 应用与调参

- 输入归一化后的 X，选择若干个可能 k 值（例如 2,3,...,10）
- 分别运行 k-means 算法
- 计算得到聚类结果的评价指标
- 根据评价指标选取最佳 k 值
- 人工评判最佳 k 值下聚类结果的可解释性，做适当调整

延伸资料

虽然大部分聚类问题都可以使用 k-means 算法，但是它肯定不能解决所有的问题。聚类是一种探索性方法，是用计算机辅助的方法来发现数据集合的结构，它的核心是将数据点分组成簇。在实际中会有很多不同类型的簇，其中圆形凸状簇是比较乐观的情况，可以使用 k-means 算法来解决；除此之外，簇也可能会有比较复杂的形状（甚至是交织簇、嵌套簇），k-means 算法不适合处理这样的簇。为了应对 k-means 算法不能处理的情况，就产生了一些其他划分簇的方法，例如密度聚类、层次聚类等等。不同的聚类算法适用于不同的问题，根据簇的形状和结构而定。

既然有这么多种聚类算法，为什么选择 k-means 算法作为课程内容呢？这不仅仅是因为它的经典，最重要的原因是：k-means 算法是所有聚类算法的本源。一方面，它简单直观，很容易对它进行改进和拓展，衍生出很多变体，例如模糊聚类等；另一方面，为了应对 k-means 算法及其变体不适合处理的复杂数据分布，产生了一些其他聚类方法，例如密度聚类、层次聚类等。

希望你在学习的过程中，重点把握课程中解决聚类问题的思路，k-means 算法的思想和使用方法，尤其要注重学习的方法，能够举一反三。这样，完成课程学习之后，你不仅能掌握 k-means 算法，还掌握了打开聚类算法世界的一把钥匙。顺着课程讲授的方法，对照 k-means 算法，你就能很容易理解 k-means 算法的那些变体和其他聚类方法。只要掌握了最经典最本质的方法，然后在数据分析的实际工作中根据不同的情况稍作调整，不断探索和尝试，最终选择合适的算法解决了实际问题，这样就是一名优秀的数据分析师了。

以下是关于常用聚类方法和 Python 中实现的聚类算法的简要介绍。

1）常用的聚类方法以及对应的主要算法

类别	包括的主要算法
划分（分裂）方法	k-means 算法（k-平均）、k-MEDOIDS 算法(k-中心点)、CLARANS 算法（基于选择的算法）
层次分析方法	BIRCH 算法（平衡迭代规约和聚类）、CURE 算法（代表点聚类）、CHAMELEON 算法（动态模型）
基于密度的方法	DBSCAN 算法（基于高密度连接区域）、DENCLUE 算法（密度分布函数）、OPTICS 算法（对象排序识别）
基于网格的方法	STING 算法（统计信息网络）、CLIOUE 算法（聚类高维空间）、WAVE-CLUSTER 算法（小波变换）
基于模型的方法	统计学方法、神经网络方法

2）Python 主要聚类分析算法

Python 的聚类相关算法主要在 Scikit-Learn 中，主要相关函数如下：

对象名	函数功能	所属工具箱
KMeans	k 均值聚类	sklearn.cluster
AffinityPropagation	吸引力传播聚类，2007 年提出，几乎优于所有其他方法，不需要指定聚类数，但运行效率较低	sklearn.cluster
MeanShift	均值漂移聚类算法	sklearn.cluster
SpectralClustering	谱聚类，具有效果比 k 均值好，速度比 k 均值快等特点	sklearn.cluster
AgglomerativeClustering	层次聚类，给出一棵聚类层次树	sklearn.cluster
DBSCAN	具有噪声的基于密度的聚类方法	sklearn.cluster
BIRCH	综合的层次聚类算法，可以处理大规模数据的聚类	sklearn.cluster

这些不同模型的使用方法是大同小异的，基本都是先用对应的函数建立模型，然后用.fit()方法来训练模型，训练好之后，就可以用.label_方法给出样本数据的标签，或者用.predict()方法预测新的输入的标签。关于以上模型的具体介绍可以查看官方文档：

https://scikit-learn.org/stable/modules/clustering.html#clustering

此外，Scipy 库也提供了一个聚类子库 scipy.cluster，里面提供了一些聚类的算法，如层次聚类等，但没有 Scikit-Learn 那么完善和丰富，这里就不介绍了。