MathWorks®

应用无监督学习

MathWorks®

何时考虑无监督学习

无监督学习适用的场景是，您想要探查数据，但还没有特定目标或不确定数据包含什么信息。这也是减少数据维度的好方法。

无监督学习技术

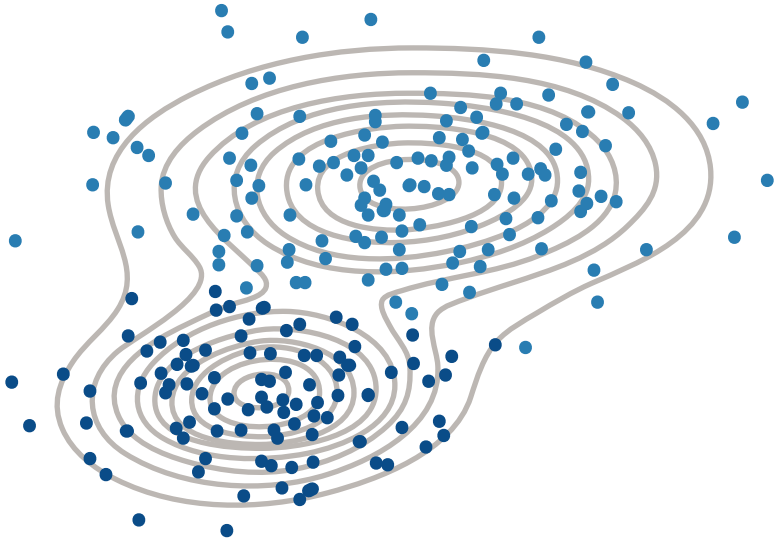
如我们在第 1 部分所见，绝大多数无监督学习技术是聚类分析的形式。

在聚类分析中，根据某些相似性的量度或共有特征把数据划分成组。采用聚类的组织形式，同一类（或簇）中的对象非常相似，不同类中的对象截然不同。

聚类算法分为两大类：

- 硬聚类，其中每个数据点只属于一类
- 软聚类，其中每个数据点可属于多类

如果您已经知道可能的数据分组，则可以使用硬聚类或软聚类技术。



高斯混合模型可用来将数据分成两类。

如果您不知道数据可能如何分组：

- 使用自我组织的特征图或层次聚类，查找数据中可能的结构。
- 使用聚类评估，查找给定聚类算法的“最佳”组数。

常见硬聚类算法

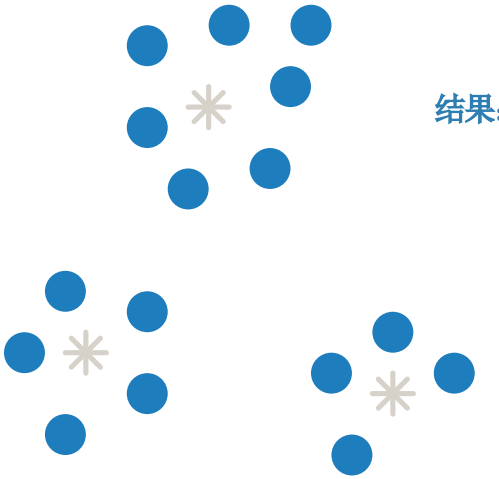
k-均值

工作原理

将数据分割为 k 个相互排斥的类。一个点在多大程度上适合划入一个类由该点到类中心的距离来决定。

最佳使用时机...

- 当聚类的数量已知时
- 适用于大型数据集的快速聚类



结果: 聚类中心

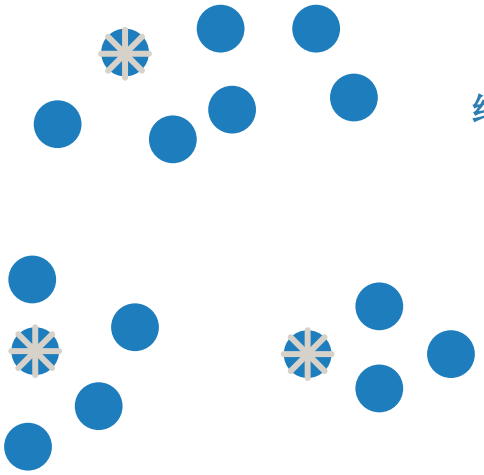
k-中心点

工作原理

与 k-均值 类似，但要求类中心与数据中的点契合。

最佳使用时机...

- 当聚类的数量已知时
- 适用于分类数据的快速聚类
- 扩展至大型数据集



结果: 与数据点契合的聚类中心

常见硬聚类算法（续）

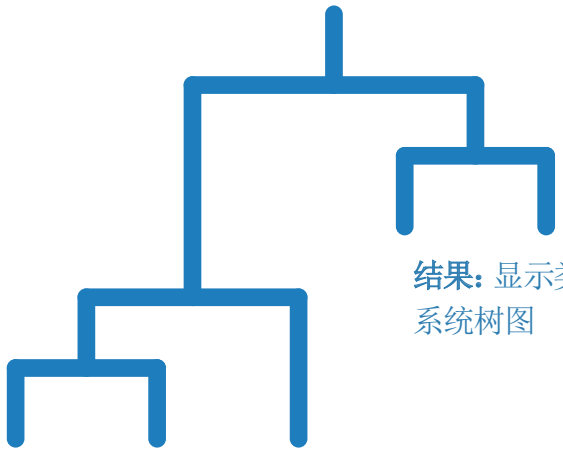
层次聚类

工作原理

通过分析成对点之间的相似度并将对象分组到一个二进制的层次结构树，产生聚类的嵌套集。

最佳使用时机...

- 当您事先不知道您的数据中有多少类时
- 您想要可视化地指导您的选择



结果：显示类之间层次关系的系统树图

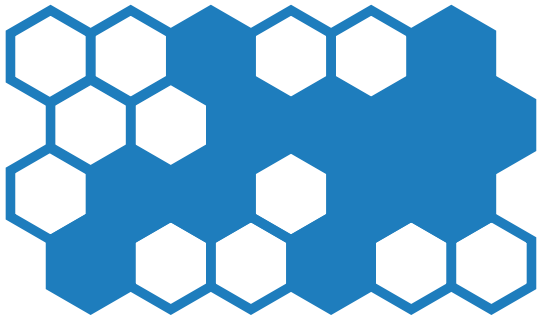
自组织映射

工作原理

基于神经网络的聚类，将数据集变换为保留拓扑结构的 2D 图。

最佳使用时机...

- 采用 2D 或 3D 方式可视化高维数据
- 通过保留数据的拓扑结构（形状）降低数据维度



结果：
低维度（通常 2D）表现形式

常见硬聚类算法（续）

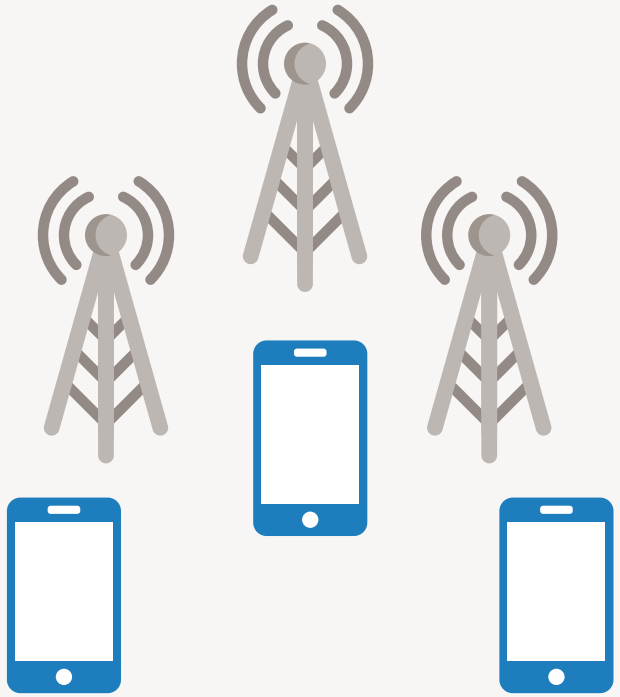
示例：使用 k-均值 聚类为手机信号塔选址

移动电话公司想知道手机信号塔的数量和位置，以便提供最可靠的服务。为实现最佳信号接收，这些塔必须位于人群聚集的地方。

工作流程从最初猜想需要划分多少个人群开始。为了评估这个猜想，工程师采用三个塔和四个塔比较服务效果，查看每种情形下的聚类有多好（换句话说，信号塔提供服务的效果如何）。

一部电话一次只能与一个塔通信，所以这是硬聚类问题。该团队使用 k-均值聚类，因为 k-均值将数据中的每个观察点视为空间中的一个点。找到了一种分割方法，每个类中的对象尽可能地相互靠近，并且尽可能远离其他类中的对象。

在运行算法之后，该团队能够准确地确定将数据分割成三个和四个类的结果。



常见软聚类算法

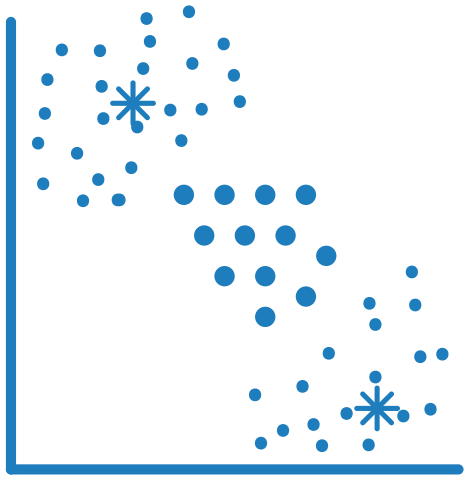
模糊 c-均值

工作原理

当数据点可能属于多个类时进行基于分割的聚类。

最佳使用时机...

- 当聚类的数量已知时
- 适用于模式识别
- 当聚类重叠时



结果: 聚类中心（类似于 k-均值），但有模糊性，所以点可能属于多个类

高斯混合模型

工作原理

基于分割的聚类，数据点来自具有一定概率的不同的多元正态分布。

最佳使用时机...

- 当数据点可能属于多个类时
- 当聚集的类具有不同的大小且含有相关结构时

结果: 一个高斯分布的模型，给出一个点在一个类中的概率

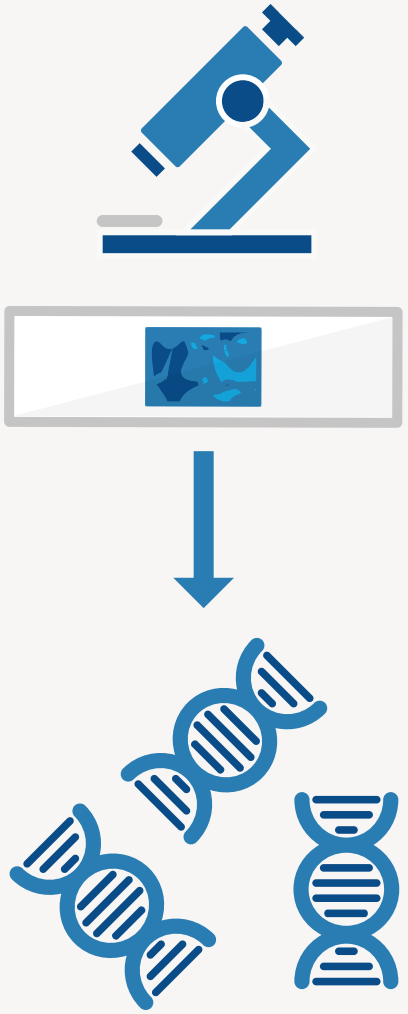
常见软聚类算法（续）

示例：使用模糊 c-均值聚类法分析基因表达数据

一个生物家团队正在通过微阵列分析基因表达数据，更好地了解涉及正常和异常细胞分裂的基因。（如果某个基因积极参与蛋白质生产之类的细胞功能，则称该基因为“已表达”。）

微阵列包含两个组织检体的表达数据。研究人员想要比较检体，确定某些基因表达模式是否与癌细胞增生有牵连。

在对数据进行预处理以消除噪声之后，他们对数据进行聚类。因为相同的基因可能涉及多个生物学过程，没有单个基因可能只属于一类。研究人员对数据运用模糊 c-均值算法。然后，他们对聚集生成的类进行可视化，识别具有类似行为方式的基因组。



用降维的方法改进模型

机器学习是一种发现大数据集内部规律的有效方法。但较大的数据增加了复杂度。

随着数据集越来越大，您经常需要减少特征或**维度**的数量。

示例：EEG 数据减缩

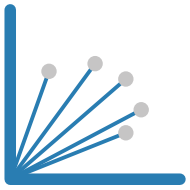
假设您有捕获脑电活动的脑电图 (EEG) 数据，您想使用此数据预测未来的癫痫发作。使用许多导线捕获数据，每根导线对应原始数据集中的一个变量。每个变量都包含噪声。为使您的预测算法更稳健，您使用降维技术生成数量较少的特征。由于这些特征是从多个传感器计算出来的，所以不太容易受单个传感器中的噪声影响，如果您直接使用原始数据，则噪声的影响会非常明显。

常见降维技术

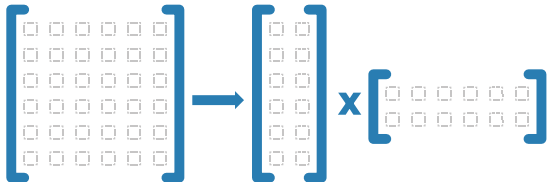
三个最常用的降维技术是：

主成分分析 (PCA) — 对数据执行线性变换，使您的高维数据集中的绝大多数方差或信息被前几个主成分捕获。第一个主成分将会捕获大部分方差，然后是第二个主成分，以此类推。

因子分析 — 识别您的数据集中各变量之间潜在的相关性，提供数量较少的未被发现的潜在因子或公共因子的一种表现方式。



非负矩阵分解 — 当模型项必须代表非负数（比如物理量）时使用。



使用主成分分析

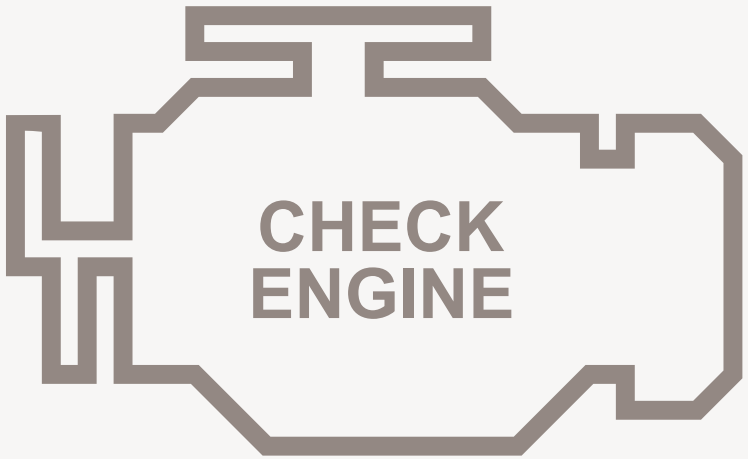
在有许多变量的数据集中，变量组经常一起移动。PCA 充分利用这种信息冗余，通过原始变量的线性组合生成新变量，使少数新变量能够捕获大多数信息。

每个主成分都是原始变量的线性组合。因为所有主成分互不相关，所以没有冗余信息。

示例：发动机健康状况监测

您有一个数据集，包括了对发动机上不同传感器的测量（温度、压力、排放等）。尽管大量数据来自健康的发动机，传感器也会捕获来自需要维护的发动机的数据。

查看任何一个传感器，可能看不出任何明显的异常。发动机异常，通过应用 PCA，您可以变换此数据，使传感器测量中的大部分变动被少数的主成分捕获。与观察原始传感器数据相比，通过检查这些主成分来区别健康和不健康的发动机比较容易。



使用因子分析

您的数据集可能包含重叠的已测变量，意味着这些变量相互依赖。通过因子分析，可将模型拟合到多元数据来评估这种相互依存关系。

在因子分析模型中，已测变量依赖数量较少的未发现（潜在）因子。因为每个因子都可能影响多个变量，所以称为公因子。假定每个变量都取决于公因子的线性组合。

示例：跟踪股价变动

在 100 个星期的时间里，对十家公司记录了股价的百分比变化。这十家公司，有四家是科技公司，三家从事金融业，还有三家从事零售业。假设相同行业的公司股价将随经济环境的变化而一同变化，这似乎很合理。因子分析可以提供数量证据来支持这一假定。

使用非负矩阵因式分解

此降维技术基于特征空间的低秩逼近。除了减少特征数量以外，还保证特征为非负数，从而产生遵守诸如物理量非负等特征的模型。

示例：文本挖掘

假设您想要探查多个网页间词汇和风格的变化。您创建一个矩阵，其中每行对应一个网页，每列对应一个单词（“the”、“a”、“we”等）。数据将是一个特定词出现在特定页面上的次数。

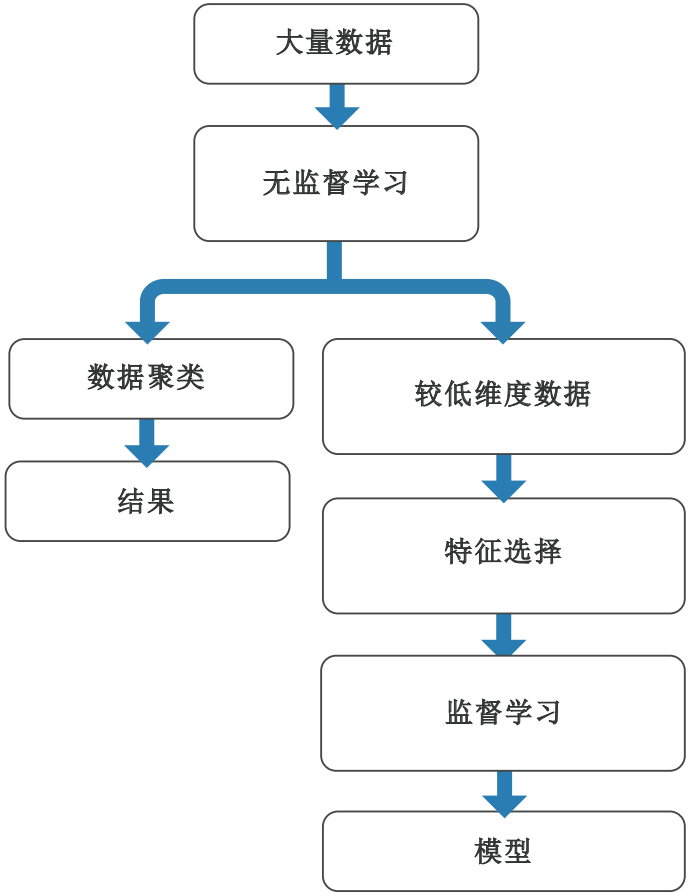
由于英语有一百多万个单词，所以您应用非负矩阵因式分解，创建任意数量的特征，表示高级别概念，而不是一个个单词。运用这些概念，更容易区分新闻、教育内容和在线零售内容。

后续步骤

在本部分中，我们详细介绍了无监督学习的硬聚类和软聚类算法，提供了一些为您的数据选择合适算法的技巧，展示了减少数据集内的特征数量如何改进模型性能。至于后续步骤：

- 无监督学习可能是您的最终目标。例如，如果您做市场研究并想根据网站行为有针对性地划分消费群体，那么，聚类算法几乎肯定能给您想要寻求的结果。
- 另一方面，您可能想使用无监督学习，作为监督式学习的预处理步骤。例如，应用聚类技术得出数量较少的特征，然后使用这些特征作为训练分类器的输入。

在第 4 部分，我们将探索监督学习算法和技术，了解如何通过特征选择、特征减缩和参数调节来改进模型。



了解更多

准备更深入地钻研？浏览以下无监督学习资源。

聚类算法和技术

k-均值

使用 K-均值和层次聚类来发现数据中的自然模式

使用 K-均值和自组织映射进行基因聚类

使用 K-均值聚类实现基于颜色的分割

分层聚类

基于连接的聚类

鸢尾花聚类

自组织映射

使用自组织映射进行数据聚类

模糊 c-均值

使用模糊 C-均值聚类法对拟随机数据进行聚类

高斯混合模型

高斯过程回归模型

对来自高斯分布混合的数据进行聚类

使用软聚类法对高斯混合数据进行聚类

调节高斯混合模型

图像处理示例：使用高斯混合模型检测汽车

降维

使用 PCA 分析美国城市的生活质量

使用因子分析法分析股价

非负矩阵因式分解

执行非负矩阵因式分解

使用减法聚类对郊区通勤进行建模

93116v00