非监督学习与聚类分析

汪小圈

2025-04-14

非监督学习简介

- 非监督学习（Unsupervised Learning）是机器学习的一个分支，其主要特点是训练数据**没有**预先标记的输出标签。
- 与监督学习不同，非监督学习的目标是从数据本身发现隐藏的结构、模式或关系。
- 它试图理解数据的内在分布和特征，而不是预测一个特定的目标变量。
- 在探索性数据分析、数据预处理和发现未知模式方面非常有用。

主要非监督学习技术

非监督学习主要包括三类技术：

- **聚类 (Clustering)**: 将相似的数据点分组。
- **降维 (Dimensionality Reduction)**: 减少数据的维度，保留主要信息。
- **文本分析/自然语言处理的非监督部分**: 从文本数据中发现模式和主题。

降维技术简介

降维是指在保留数据主要信息的前提下，减少数据特征数量的过程。解决高维数据的问题：

- 冗余信息多
- 计算复杂度高
- 可视化困难
- “维度灾难”

主要降维算法

- **主成分分析 (PCA)**:
  - 最常用的线性降维方法。
  - 寻找数据方差最大的方向（主成分）进行投影。
- **t-SNE**:
  - 非线性降维方法。
  - 特别擅长高维数据可视化，保留局部结构。
- **自编码器 (Autoencoders)**:
  - 基于神经网络的非线性降维方法。
  - 通过编码器压缩数据，再通过解码器重构。

降维的金融应用

- **风险因子识别**: 从大量市场指标中提取主要的风险因子。
- **数据可视化**: 将高维金融数据降维到二维或三维空间进行可视化分析。
- **特征工程**: 减少模型输入的特征数量，提高训练效率和泛化能力。
- **资产定价**: 减少影响资产价格的因素数量，构建更简洁的定价模型。

非监督文本分析简介

- **词嵌入 (Word Embeddings)**:
  - 如 Word2Vec, GloVe。
  - 将词语表示为低维稠密向量，捕捉语义关系。
- **主题建模 (Topic Modeling)**:
  - 如潜在狄利克雷分配 (LDA)。
  - 从大量文档中自动发现隐藏的主题结构。
- **文本聚类**: 将相似的文本（如新闻、报告、评论）分组。

非监督文本分析的金融应用

- **舆情分析**: 分析新闻、社交媒体中的市场情绪或对公司的情感倾向。
- **信息提取**: 从财报、公告、合同等非结构化文本中自动提取关键信息。
- **文档摘要与分类**: 自动生成研报摘要，或将金融文档按主题分类。
- **风险信号挖掘**: 从新闻或监管文件中识别潜在的风险事件或趋势。

聚类分析的概念

- 聚类分析是一种非监督学习技术，其目标是将数据集中的样本根据它们的相似性划分为若干个组（簇）。
- 聚类的核心思想包含两点：
  - **簇内相似性最大化：** 同一个簇内的数据点应该尽可能相似。
  - **簇间差异性最大化：** 不同簇之间的数据点应该尽可能不相似。

聚类与分类的区别

- **分类（有监督学习）**：
  - 有预先定义的类别标签。
  - 目标是学习将新样本分配到已知类别的规则。
  - 评估基于预测标签与真实标签的比较。
- **聚类（无监督学习）**：
  - 没有预先定义的类别标签。
  - 目标是发现数据中的自然分组。
  - 评估基于分组的内部结构特性（如组内紧密度、组间分离度）。

K 均值聚类 (K-Means) 算法原理

K-Means 是一种迭代算法，旨在将数据划分为预先指定的 K 个簇。算法步骤：

1. 随机选择 K 个初始质心。
2. **分配步骤：** 将每个数据点分配给距离其最近的质心，形成 K 个簇。
3. **更新步骤：** 重新计算每个簇的质心（通常是簇内所有点的均值）。
4. 重复步骤 2 和 3，直到质心不再发生显著变化或达到最大迭代次数。

K-Means 算法的优缺点

优点: * 算法简单，容易理解和实现。* 计算效率高，处理大数据集速度较快。

缺点: * 需要预先指定簇的数量 K，而 K 值的选择往往比较困难。* 对初始质心的选择敏感，可能陷入局部最优解。* 对非球状簇、不同大小和密度的簇效果不佳。* 对异常值（Outliers）比较敏感。

K-Means 聚类评估方法：肘部法则

原理： * 对不同的 K 值运行 K-Means 算法，计算每个 K 值对应的惯性（簇内平方和，WCSS）。* 随着 K 值的增加，惯性总体呈下降趋势，但下降速率会逐渐变缓。* 寻找图中的”肘部”，即曲线下降速率发生明显变化的点。该点对应的 K 值被视为较佳选择。

注意事项： * 肘部法则是一种视觉方法，有时肘部可能不够明显。* 需要结合业务理解和其他评估指标来确定最终的 K 值。

K-Means 聚类评估方法：轮廓系数

计算方法: * 对于数据集中的每个样本点 i：* 计算 a(i): 点 i 与同簇中其他所有点的平均距离（衡量簇内紧密度）。* 计算 b(i): 点 i 与距离最近的其他簇中所有点的平均距离（衡量簇间分离度）。* 轮廓系数 s(i) = (b(i) - a(i)) / max(a(i), b(i))。* 数据集的整体轮廓系数是所有样本点轮廓系数的平均值。

轮廓系数解读: * 值域范围：[-1, 1] * 接近 +1：表示聚类效果好。* 接近 0：表示样本处于两个簇的边界附近。* 接近 -1：表示样本可能被分配到了错误的簇。

DBSCAN 聚类：基本概念

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) 是一种基于密度的聚类算法。

核心概念: * **(Epsilon):** 邻域半径，定义点之间的”近”的概念。* **MinPts:** 邻域内最少点数，用于判断密度。* **核心点 (Core Point):** 在其 邻域内至少有 MinPts 个点的点。* **边界点 (Border Point):** 不是核心点，但在某个核心点的 邻域内的点。* **噪声点 (Noise Point):** 既不是核心点也不是边界点的点。

DBSCAN 聚类：算法步骤

1. 任选一个未被访问的点 p。
2. 标记 p 为已访问。
3. 如果 p 是核心点，创建一个新簇，并将 p 的所有密度可达点加入该簇。
4. 如果 p 不是核心点，标记为噪声点并继续。
5. 重复以上步骤，直到所有点都被访问。

DBSCAN 聚类：优缺点

优点: * 不需要预先指定簇的数量。* 能发现任意形状的簇，不限于球形簇。* 能自动识别和处理噪声点。* 对离群点不敏感。

缺点: * 参数选择（ 和 MinPts）有时较为困难。* 对数据集中密度差异较大的簇效果不佳。* 计算复杂度较高（约为 $O(n^2)$），但通常可以通过空间索引优化。* 不能处理高维空间中的”维度灾难”问题。

聚类算法的金融应用：客户细分

应用场景: * 根据客户的交易行为、人口统计特征、风险偏好等将客户分组。* 应用于精准营销、个性化推荐和风险管理。

实施步骤: 1. 收集客户数据（交易历史、消费习惯、信用记录等）。2. 特征工程（缺失值处理、标准化、降维等）。3. 选择合适的聚类算法（如 K-Means）。4. 确定最优簇数（使用肘部法则或轮廓系数）。5. 解释每个客户群体的特征，制定差异化策略。

聚类算法的金融应用：异常检测

应用场景: * 识别与正常模式显著不同的交易或行为。* 用于欺诈检测、洗钱识别、市场操纵行为识别等。

实施方法: * **基于距离的方法:** 将远离聚类中心的点识别为异常。* **基于密度的方法:** 使用 DBSCAN 自动标记的噪声点作为异常点。* **混合方法:** 结合多种聚类方法，提高异常检测准确率。

聚类算法的金融应用：投资组合构建

应用场景: * 将具有相似风险收益特征的资产聚类，辅助构建多元化的投资组合。

实施步骤: 1. 收集资产历史收益率、波动率、相关性等数据。2. 使用 K-Means 或层次聚类将资产分组。3. 从每个簇中选择代表性资产，构建多样化投资组合。4. 通过这种方式实现更好的风险分散效果。

金融市场应用：股票板块轮动分析

目标： * 利用聚类技术识别股票市场中可能存在的板块轮动现象或隐藏的股票群体特征。

步骤： 1. **数据收集：** 选择股票池，收集历史收益率数据。2. **数据预处理：** 处理缺失值，标准化数据。3. **聚类分析：** 使用 K-Means 对股票进行聚类。4. **结果解读与分析：** * **簇成员分析：** 检查每个簇中包含的股票类型。* **簇表现分析：** 分析不同时期各簇的表现，寻找轮动模式。

聚类在半监督学习中的应用

背景: 在标记数据稀少的情况下，如何充分利用大量无标签数据?

方法: 1. 使用 K-Means 对所有数据进行聚类，找到代表性样本。2. 仅对代表性样本进行人工标记。3. 将标签传播到同一簇中的所有样本。4. 使用扩展后的标记数据训练监督学习模型。

优势: 大幅减少标记工作量，同时保持较好的模型性能。

总结：非监督学习技术比较

- **降维技术**: 减少维度，保留主要信息，解决”维度灾难”问题
  - PCA: 线性降维，计算效率高
  - t-SNE: 非线性降维，可视化效果好，但计算复杂
  - 自编码器: 基于神经网络，可处理复杂非线性关系
- **聚类技术**:
  - K-Means: 简单高效，适合大数据集，但需预先指定簇数
  - DBSCAN: 能处理任意形状簇，不需预先指定簇数，但参数选择困难
  - 层次聚类: 提供多层次结构，直观易解释，但计算复杂度高
- **应用价值**: 发现隐藏模式、简化复杂性，辅助决策制定