CSDN
不 止 于 代 码

4.5 聚类（Clustering）

CSDN学院

2017年11月

大纲

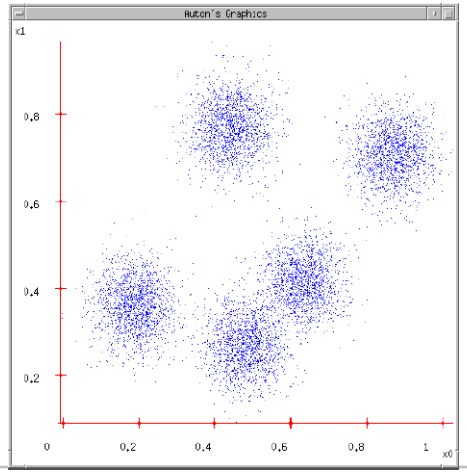
- 常用聚类算法
- 聚类性能评估
- 案例分析

非监督学习

- 在监督学习任务中，我们有一系列标签的数据 $\mathcal{D}=\{\mathbf{x}_i, y_i\}_{i=1}^{N}$，我们需要据此拟合一 个函数，使得这个函数能表示输入$\mathbf{x}$与输出y之间的关系。
- 在非监督学习中，我们只有一系列点 $\mathcal{D}=\{\mathbf{x}_i\}_{i=1}^{N}$，却没有标签的数据$y$。
- 在非监督学习中，我们需要将一系列无标签的训练数据，输入到一个算法中，这个算法将寻找这个数据的内在结构。

聚类

- 聚类的输入是一组未被标记的样本，根据数据中样本与样本之间的距离或相似度将样本划分为若干组 / 类 / 簇（cluster）。
- 划分的原则：类内散度最小、类间散度最大

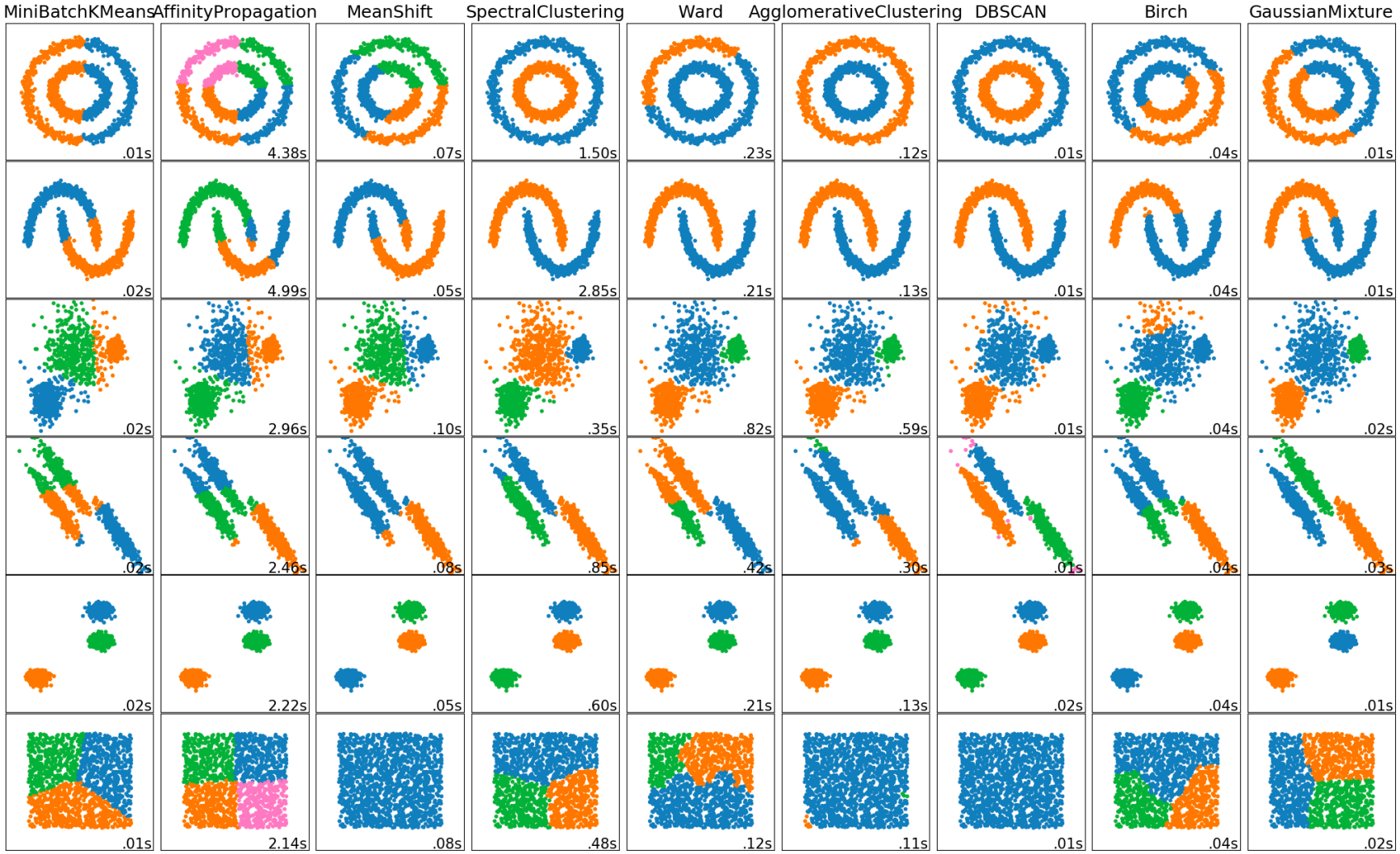


物以类聚、人以群分

► 为什么要做聚类？

- 计算：使用聚类中心而不是原始数据
- 统计：识别/去除离群点（outliers）
- 可视化/理解

例：scikit learning中支持的聚类算法



常用聚类算法

- 基于距离、相似度的聚类算法
  - *K*-means（*K*均值）及其变种（ *K*-centers 、 Mini Batch *K*-Means）
  - Mean shift
  - 吸引力传播（Affinity Propagation ，AP）
  - 层次聚类
    - 聚合聚类（Agglomerative Clustering）
- 基于密度的聚类算法
  - DBSCAN、DensityPeak（密度最大值聚类）
- 基于连接的聚类算法
  - 谱聚类

相似度/距离计算

- 闵可夫斯基（Minkowski）距离
- 余弦相似度(cosine similarity)
- Pearson相似系数
- 杰卡德相似系数(Jaccard)

闵可夫斯基距离

- 闵可夫斯基距离Minkowski
- $d(\mathbf{x},\mathbf{y}) = \left(\sum_{j=1}^{D}|x_j - y_j|^p\right)^{1/p}$
- 当p=2时，为欧氏距离： $d(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{D}\sqrt{\left(x_j - y_j\right)^2}$
- 当p=1时，为曼哈顿距离： $d(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{D}|x_j - y_j|$

余弦相似度(cosine similarity)

- 夹角余弦：两变量**x**与**y**看作*D*维空间的两个向量，这两个向量间的夹角余弦可用下式进行计算

$$d(\mathbf{x},\mathbf{y}) = \frac{\sum_{j=1}^{D} x_j y_j}{\sqrt{\sum_{j=1}^{D} x_j^2 \sum_{j=1}^{D} y_j^2}} = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

相关系数

- 相关系数（亦被称为Pearson系数）经常用来度量变量间的相似性。变量x与y的相关系数定义为

$$r(\mathbf{x},\mathbf{y}) = \frac{cov(\mathbf{x},\mathbf{y})}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}} = \frac{E\left[(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}})\right]}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}}$$

$$= \frac{\sum_{j=1}^{D}\left(x_j - u_{x_j}\right)\left(y_j - u_{y_j}\right)}{\sqrt{\sum_{j=1}^{D}\left(x_j - u_{x_j}\right)^2 \sum_{j=1}^{D}\left(y_j - u_{y_j}\right)^2}}$$

当对数据做中心化后，

$$\mu_{\mathbf{x}} = \mu_{\mathbf{y}} = 0$$

相关系数等于余弦相似度

$$d(\mathbf{x},\mathbf{y}) = \frac{\sum_{j=1}^{D} x_j y_j}{\sqrt{\sum_{j=1}^{D} x_j^2 \sum_{j=1}^{D} y_j^2}} = \frac{\mathbf{x}^T\mathbf{y}}{|\mathbf{x}||\mathbf{y}|}$$

杰卡德相似系数(Jaccard)

- 杰卡德相似系数（交比并，常用于直方图相似度度量）

$$J(\mathbf{x},\mathbf{y}) = \frac{\sum_{j=1}^{D}\left(x_j \cap y_j\right)}{\sum_{j=1}^{D}\left(x_j \cup y_j\right)}$$

CSDN
不 止 于 代 码

THANK YOU

AI100