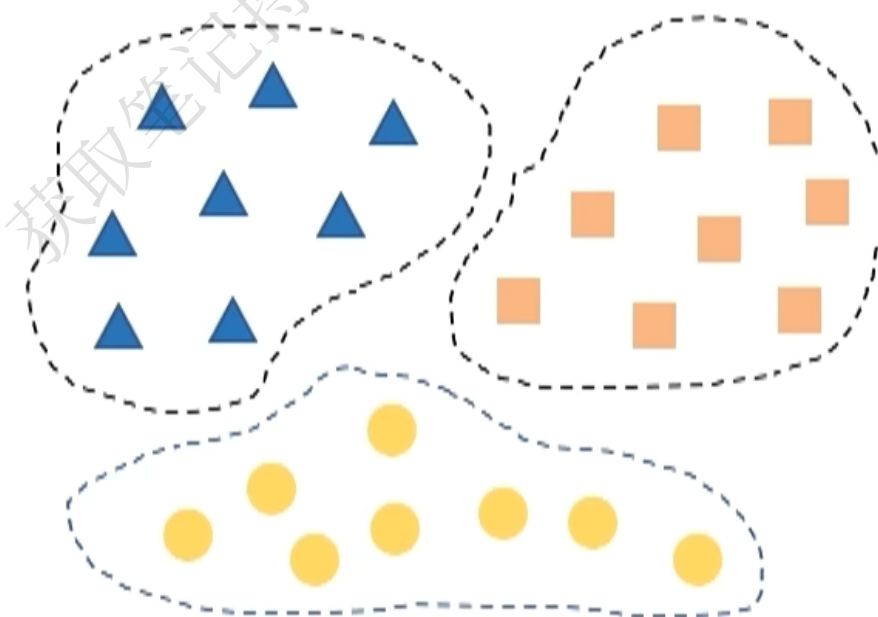


Lecture 9: 聚类

聚类基本概念

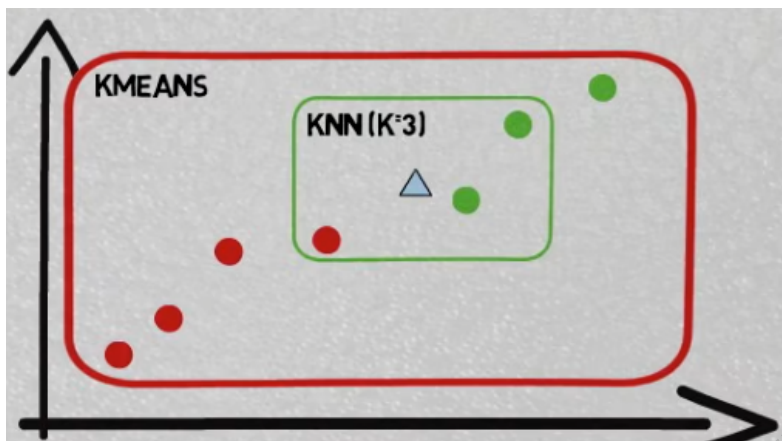
概念：聚类基本概念

- 引入：聚类算法与无监督学习
 - 在“无监督学习”任务中研究最多、应用最广；
 - 在有监督学习情况下，分类是把离散的数据分类到几个序列，回归是有实值的；
 - 无监督中的聚类，对应于有监督中的分类；
 - 无监督中的密度估计，对应于有监督中的回归；
- 概念：如何定义相似
 - 用两个点在欧氏空间中的距离，定义定义为它们之间的相似度；
 - 为了定义两个样本是否属于一类，定义了簇（`cluster`）的概念
- 目的：聚类算法的目的
 - 将数据样本划分为若干个通常不相交的“簇”（`cluster`）；
 - 既可以作为一个单独过程（用于找寻数据内在的分布结构），也可作为分类等其他学习任务的前驱过程；
 - 比如要对一个通讯软件上的客户进行分类时，可以先使用聚类的技术，将他们分成几个类，然后再对每个分类当中进行有监督的学习；
- 特点：
 - （1）非监督类的算法，不需要样本的标注信息；
 - （2）对于 K 的取值，也就是你认为数据集中的样本应该分为几类，这个参数的设置极为敏感；
- 图示：



分析：聚类算法和 KNN 的区别

- KNN :
 - (1) 只关注待测样本的局部分布;
 - (2) KNN 是有监督学习, 需要样本的标注才可以训练模型;
- K-Means
 - (1) 关注样本的全局分布;
 - (2) K-Means 是无监督学习;
- 图示:



距离计算

概念: 距离计算

- 概念:
 - 距离度量 (distance metric) 需满足的基本性质:
 - 非负性: $\text{dist}(\mathbf{x}_i, \mathbf{x}_j) \geq 0$;
 - 同一性: $\text{dist}(\mathbf{x}_i, \mathbf{x}_j) = 0$ 当且仅当 $\mathbf{x}_i = \mathbf{x}_j$;
 - 对称性: $\text{dist}(\mathbf{x}_i, \mathbf{x}_j) = \text{dist}(\mathbf{x}_j, \mathbf{x}_i)$;
 - 直递性: $\text{dist}(\mathbf{x}_i, \mathbf{x}_j) \leq \text{dist}(\mathbf{x}_i, \mathbf{x}_k) + \text{dist}(\mathbf{x}_k, \mathbf{x}_j)$.
 - 有些时候可能需要突破上面四个标志, 违背了某一项, 此时需要非距离计算;
- 实现:
 - 1. 随机从数据集中选取 K 个样本当做 centroids ;
 - 2. 对于数据集中的每个点, 计算它距离每个 centroids 的距离, 并把它归为距离最近的那个 cluster
 - 3. 更新新的中心位置;
 - 4. 重复 2、3, 直到中心的位置不再改变;
- 概念: 常见距离形式

闵可夫斯基距离 (Minkowski distance)

$$\text{dist}_{\text{mk}}(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{u=1}^n |x_{iu} - x_{ju}|^p \right)^{\frac{1}{p}}$$

$p = 2$: 欧氏距离(Euclidean distance)

$p = 1$: 曼哈顿距离(Manhattan distance)

概念: 无序属性和混合属性

- 概念:

对无序(non-ordinal)属性, 可使用 VDM (Value Difference Metric)

令 $m_{u,a}$ 表示属性 u 上取值为 a 的样本数, $m_{u,a,i}$ 表示在第 i 个样本簇中在属性 u 上取值为 a 的样本数, k 为样本簇数, 则属性 u 上两个离散值 a 与 b 之间的 VDM 距离为

$$VDM_p(a, b) = \sum_{i=1}^k \left| \frac{m_{u,a,i}}{m_{u,a}} - \frac{m_{u,b,i}}{m_{u,b}} \right|^p$$

对混合属性, 可使用 MinkovDM

$$\text{MinkovDM}_p(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{u=1}^{n_c} |x_{iu} - x_{ju}|^p + \sum_{u=n_c+1}^n \text{VDM}_p(x_{iu}, x_{ju}) \right)^{\frac{1}{p}}$$

聚类方法概述

概念: 聚类方法不存在绝对好坏标志

- 概念:
 - The goodness of clustering depends on the opinion of the user.
 - 聚类也许是机器学习中“新算法”出现最多、最快的领域, 总能找到一个新的“标准”使以往算法对它无能为力;
- 概念:

□ 原型聚类

- 亦称“基于原型的聚类” (prototype-based clustering)
- 假设：聚类结构能通过一组原型刻画
- 过程：先对原型初始化，然后对原型进行迭代更新求解
- 代表：**k均值聚类**，学习向量量化(LVQ)，高斯混合聚类

□ 密度聚类

- 亦称“基于密度的聚类” (density-based clustering)
- 假设：聚类结构能通过样本分布的紧密程度确定
- 过程：从样本密度的角度来考察样本之间的可连接性，并基于可连接样本不断扩展聚类簇
- 代表：**DBSCAN**, OPTICS, DENCLUE

□ 层次聚类 (hierarchical clustering)

- 假设：能够产生不同粒度的聚类结果
- 过程：在不同层次对数据集进行划分，从而形成树形的聚类结构
- 代表：**AGNES** (自底向上), DIANA (自顶向下)

获取笔记持续更新可加微信: Wand2277