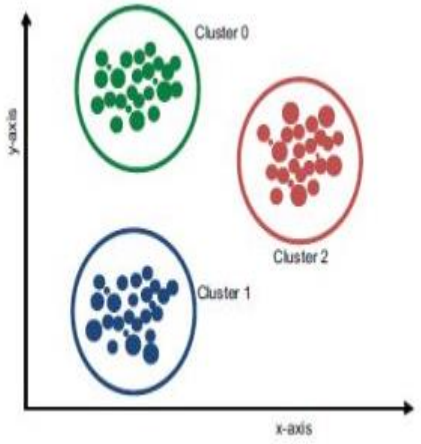
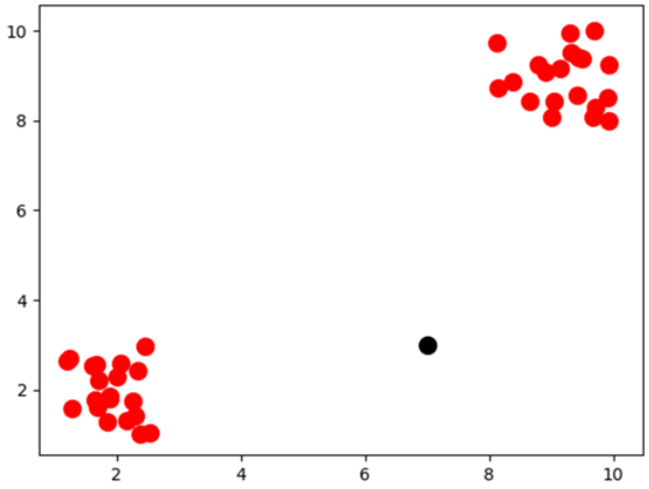
1.聚类算法定义

- 聚类就是按照某个特定标准(如距离准则)把一个数据集分割成不同的簇(cluster)，使得同一个簇内的数据对象的相似性尽可能大，同时不在同一个簇中的数据对象的差异性也尽可能地大。即“物以类聚”。
- 聚类是无监督算法，事先并不知道数据集会被分割成多少个簇及每一个簇的具体含义。



2.聚类算法的用途

- 对于拥有老客户数据的公司，聚类可以将属性相似的客户分到相同的组，称为客户市场划分(customer segmentation)。对不同分组的客户提供特别的服务和产品等，称为客户关系管理(customer relationship management)。同时分组还可以发现离群点，即那些不同于其他客户的客户。
- 离群点可能会形成一块新的市场。



聚类算法的用途

- 一幅图像通常包含大量的像素点，需要存储大量的数据，比如一幅图像的每个像素点以24位的数据来表示，可以表示1600万种颜色，但如果通过聚类只保留其中64种主色调，那么，对于每个像素只需要六位而不是24位。或者将某些像素值相近的点的像素值取这些点的平均值，达到压缩图像的效果。这种转换以丢失图像细节为代价，赢得了存储和传送图像的空间和时间。

聚类算法的用途

原图

聚为10个簇

聚为50个簇

聚为100个簇

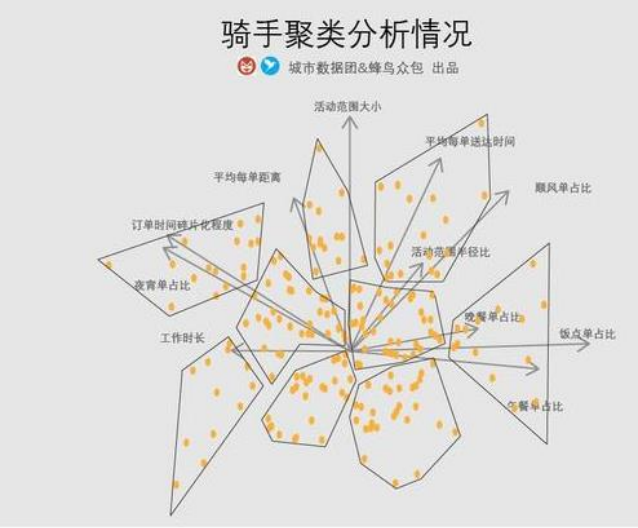
聚类算法的用途

- DNA是生命的蓝图，通过DNA转录RNA，再通过RNA转录蛋白质，机体中的每一个细胞和所有重要组成部分都需要蛋白质参与。生物学领域中的主要研究内容之一就是将一个蛋白质的氨基酸序列与另一个蛋白质的序列进行比对与匹配。由于序列可能很长，许多模板串需要进行匹配，并且可能还会被删节、插入和置换，所以其比对很困难。如果将氨基酸看作是字母、蛋白质看作是句子，并构建一个称为结构域的类似单词的结构(即频繁出现在不同蛋白质中的一串氨基酸)，将聚类应用在其中学习结构域就可以大大降低其难度。

聚类算法的用途

- http://m.toutiaocdn.cn/group/6599781160069366279/?iid=43799390846&app=news_article&timestamp=1536641390&article_category=stock&group_id=6599781160069366279

在此基础上，我们进行了K-means聚类，把骑手工作风格分为9类。如下图所示，图中每个点代表一个骑手，同一多边形内的骑手具有同类风格。



本课程涉及到的聚类算法

- k-means(简单，时间复杂度低，应用广泛)
- 谱聚类(针对图数据的聚类)
- 马尔可夫聚类(这对图数据的聚类)
- 其他聚类算法(简单介绍):基于密度的聚类(DBSCAN、OPTICS)、基于层次的聚类(凝聚型层次聚类、变色龙算法)