什么是聚类

聚类，顾名思义，把数据样本分堆儿，相似的样本聚在一起，成为一堆儿，如图 1 所示。

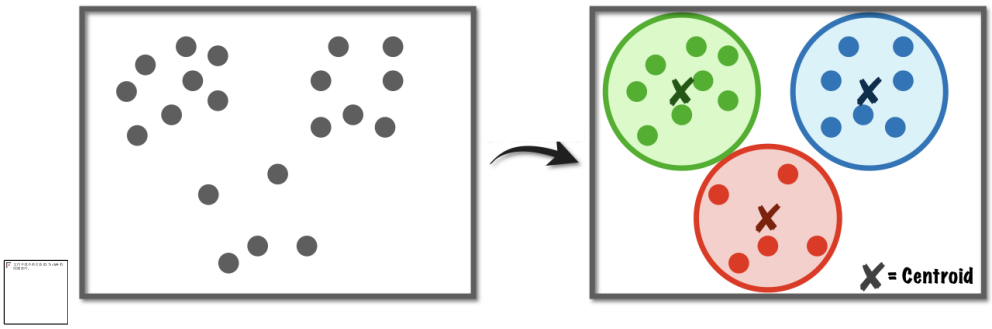
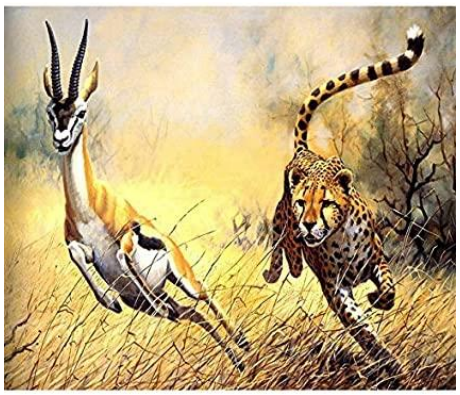


图 1：聚类任务将相似的样本聚在一起[1]

聚类是机器学习里的一项基本任务。那么，聚类有什么用处呢？

1. 人头脑里的聚类

我们先看看人头脑里的聚类。比如我们的祖先在打猎时，看到老虎和羚羊，虽然还不知道他们的名字，但从体形、牙齿、叫声等很容易将他们区分开，形成两类不同的动物。于是，他们给这两类动物贴上了标签：一类叫“羚羊”，不咬人，可以猎捕；另一类叫“老虎”，很凶猛，除非万不得已不要招惹。

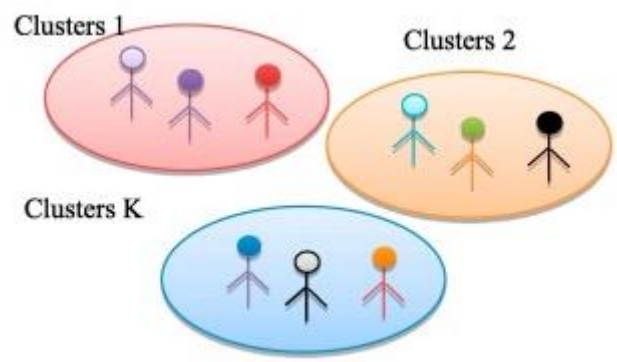


这里有两点有意思的地方：一是人们通过观察和接触，把看到的动物聚成了两类，之后才形成了“羚羊”和“老虎”的概念。二是人们形成了这两个概念，并赋予了他们各自意义之后，再遇到羚羊或老虎的个体出现，就可以把他们归到各自的类别中，从而知道是该出手抓捕，还是马上逃跑。可以看到，聚类体现的是人类的归纳能力，是获得知识的第一步。

2. 机器用聚类做什么？

机器可以用聚类做同样的事情：如果将同类样本聚在一起，就可以形成一个群体概念；对于一个新样本，我们可以将其归类到某一概念下，从而可以快速得到这一新样本的特点和属性。

以一个购物平台为例来说明，平台有很多注册用户，每个用户购买一些商品。如果我们将购买的商品作为用户的“特征”，基于这个特征就可以将有用户聚成若干类，同一类用户具有相似的购买习惯，如“奢侈品专好型”，，“时尚便捷型”，“不打折不买型”，等等。有了这些用户类，商家就可以做很多事，例如可以对不同用户群推荐不同的商品，采取不用的促销策略，快速判断一个新用户的消费习惯，等等。



3. 机器如何聚类

研究人员开发出了很多种聚类方法，其中 K-均值聚类应用的最为广泛。如图 2 所示，首先确定一个类别个数（例子中是 3 类）并初始化这些类的中心向量为随机值。依据数据样本到这些类中心的距离，将数据进行归类，不同颜色代表按当前类中心得到的归类结果。归类之后，利用每一类包含的样本重新计算类中心向量。得到新的类中心后，再次对数据进行归类，并更新类中心。这一过程迭代进行，直到类中心趋于稳定。

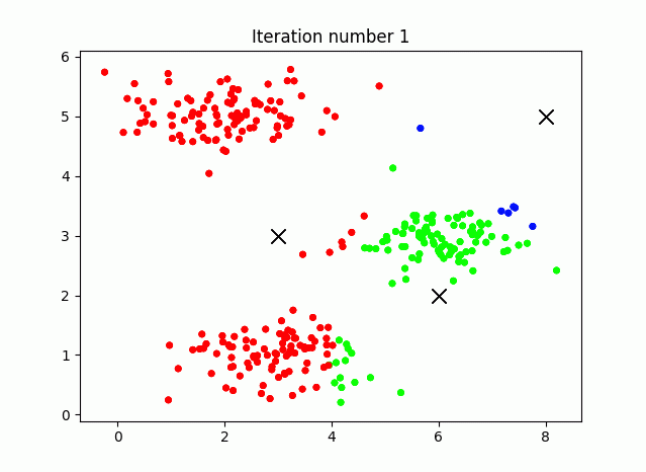


图 2：k-means 聚类过程【动图】[3]

K-均值聚类简单快捷，在实际系统中有广泛应用，但对分布不规则的数据（如月牙状），可能无法得到合理的结果。研究人员开发了很多其他聚类方法，包括基于连接的方法、基于密度的方法、基于概率模型的方法等。这些方法基于不同的数据分布假设，在实际应用中应结合数据的特点进行合理选择[2]。

[1]Alan Jeffares， K-means: A Complete Introduction，
https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c
[2] 王东，机器学习导论，清华大学出版社，2021.1.
[3] Coursera Machine Learning を Python で実装 - [Week8]k-Means, 主成分分析（PCA）
https://qiita.com/koshian2/items/3910bfe8e32e683d9046