**机器学习-聚类算法**

## 实现原理

聚类是一种无监督的学习，它将相似的对象归到同一簇中。聚类的方法几乎可以应用所有对象，簇内的对象越相似，聚类的效果就越好。

1. **means:**

K-means算法中的k表示的是聚类为k个簇，means代表取每一个聚类中数据值的均值作为该簇的中心，或者称为质心，即用每一个的类的质心对该簇进行描述。

**LVQ:**

学习向量量化算法(LVQ)和K-means算法类似，是找到一组原型向量来聚类， 每一个原型向量代表一个簇，将空间划分为若干个簇，从而对于任意的样本，可以将它划入到与它距离最近的簇中。特别的是LVQ假设数据样本带有类别标记，可以用这些类别标记来辅助聚类。

**EM:**

混合高斯模型(EM)基于多变量正态分布。类gmdistribution通过使用EM算法来拟合数据，它基于各观测量计算各成分密度的后验概率。高斯混合模型常用于聚类，通过选择成分最大化后验概率来完成聚类。与k-means聚类相似，高斯混合模型也使用迭代算法计算，最终收敛到局部最优。高斯混合模型在各类尺寸不同、聚类间有相关关系的的时候可能比k-means聚类更合适。使用高斯混合模型的聚类属于软聚类方法（一个观测量按概率属于各个类，而不是完全属于某个类），各点的后验概率提示了各数据点属于各个类的可能性。

## 程序结构

**K-means:**

有样本D（x1,x2,...,xn），先随机找K个点作为类别中心，（1）计算每个样本距K个点的距离，将样本分类到距离最近的点，（2）根据每类的样本更新类别中心。

输入：样本集，聚类簇数k

过程：

Repeat

将每个点指派到最近的质心，形成K个簇

重新计算每个簇的质心

Until

簇不发生变化或达到最大迭代次数

效果图为图1：

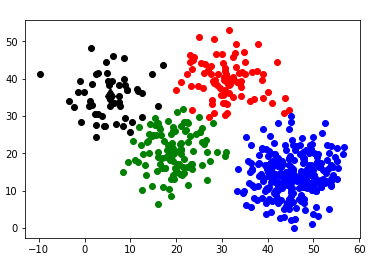


图 1

**LVQ:**

前提：1、假设所有样本是有类别标记的，每个样本Xi，有类别标记Yi

1. 找K个原型向量，（1）随机选一个样本（xi,yi），计算和K个向量的距离，找最近的那个Ki；（2）如果Ki=yi ,更新Ki靠近Xi, 否则更新Ki远离xi

输入：样本集；

原型向量个数q，各原型向量预设的类别标记；

学习率：

过程：

初始化一组原型向量

Repeat

从样本集中选取样本

计算样本与的距离，找出与距离最近的原型向量

if  then 

else 

end if 将原型向量更新为p’

Until 满足停止条件

输出：原型向量

效果图为图2：

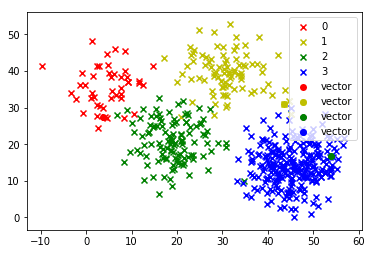


图 2

**EM:**

输入：样本集

高斯混合成分个数K

过程：

初始化高斯混合分布的模型参数

针对每一个高斯分布，随机给其均值和方差进行赋值

Repeat

针对每一个样本，计算其在各个高斯分布下的概率



针对每一个高斯分布，每一个样本对该高斯分布的贡献可以由其下的概率表示，如果概率大，则表示贡献大，反之亦然。这样把样本对该高斯分布的贡献作为权重来计算加权的均值和方差。之后替代其原本的均值和方差。

Until

高斯分布的均值和方差收敛。

效果图二维聚类如图3，3D打印图如图4：

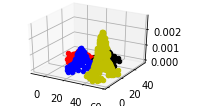
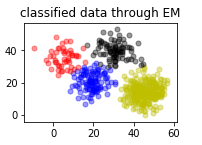


图 3 图 4

## 讨论

K-means，LVQ和EM算法都属于原型聚类，LVQ属于神经网络模型。

聚类的划分还有密度聚类，如：DBSCAN，层次聚类，如AGENS。

高斯混合模型和K-means的比较：

**相同点：**

   分类受初始值的影响

      可能限于局部最优解

      类别的个数只能靠猜测 （有K越大MAP最大后验概率越大的趋势）

**不同点：**

      K-means是硬分类，要么属于这类，要么属于那类，而高斯混合式软分类，一个样本60%属于A，40%属于B。

      多维的时候高斯混合在计算均值和方差时使用了协方差，应用了不同维度之间的相互约束关系。

本次作业用的都是最朴素的聚类，初始值为随机选取，聚类效果还不是很好，不稳定，需要进行优化。