TONGJI UNIVERSITY



**《模式识别》**

**高斯混合概型参数估计的EM算法**

**学 院 电子与信息工程学院**

**专 业 计算机科学与技术**

**学 号 2130750**

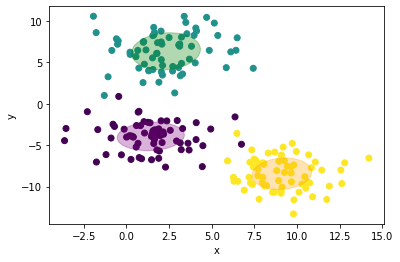
**姓 名 伍 谦**

**日 期 2021-10-20**

1. 前言

本文使用了EM算法对高斯混合概率模型进行参数估计。

1. 数据集介绍

使用了sklearn.datasets包中的make\_blobs方法生成200个样本点，每个样本具有两维特征，共有三类。

1. 基本步骤
   1. 明确需要估计的参数

在该任务中需要估计的参数有：三个高斯分布的均值𝜇1,𝜇2,𝜇3以及协方差𝜎1,𝜎2,𝜎3，高斯混合模型中各类的先验概率𝑝1,𝑝2,𝑝3,其中需要满足p1+p2+p3=1。

3.2初始化需要估计的参数

3.2.1先验概率p

1. import numpy as np
2. p=np.random.rand(classes)  *#根据类别数随机生成p*
3. p=p/p.sum()  *#归一化，保证p之和为1*

根据类别数随机生成先验概率，随后经过归一化来保证它们之和为1

3.2.2 高斯分布均值𝜇

1. means=np.random.rand(classes,dimention)  *# shape [3,2]*

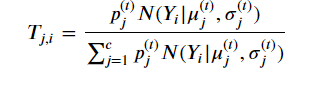
根据类别数和特征维度生成均值，均值的尺寸为[类别数，维度]

3.2.3高斯分布协方差𝜎

1. covs=np.empty((classes,dimention,dimention))
2. for i in range(classes):
3. covs[i]=np.eye(dimention)\*np.random.rand(1)\*10

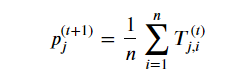
根据类别和均值生成协方差，协方差尺寸为[类别数，维度，维度]

3.3使用EM算法进行一次参数更新的完整过程

3.3.1计算每个样本属于某个类别的后验概率

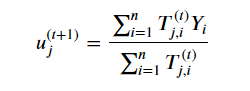
1. posterior= density \* p *#shape [2000,3]*
2. posterior= posterior/posterior.sum(axis=1,keepdims=True)  *# 归一化,使得各样本属于每个类别的概率之和为1*

3.3.2 更新先验概率p

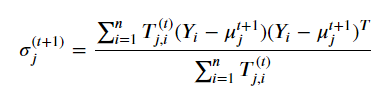


1. p\_hat=posterior.sum(axis=0)
2. p\_hat=p\_hat/num\_of\_data

3.3.3更新均值𝜇

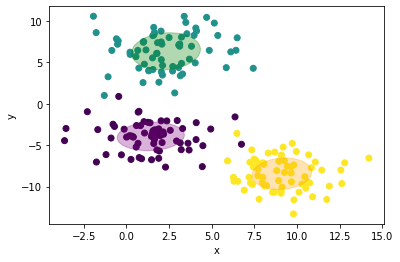


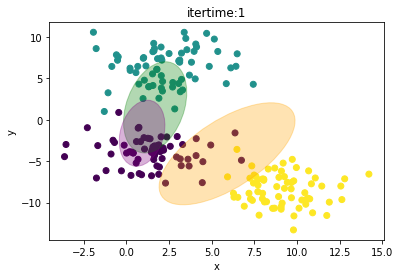
1. mean\_hat=np.matmul(data.T,posterior).T
2. mean\_hat=np.divide(mean\_hat,np.sum(posterior,axis=0,keepdims=True).T)

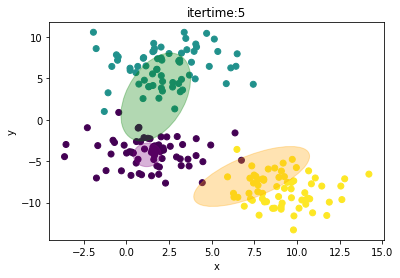
3.3.4更新协方差𝜎

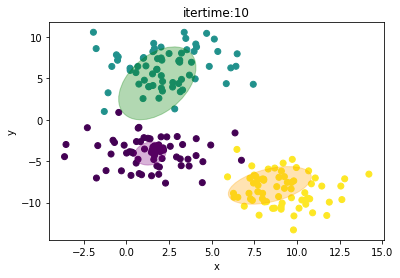
1. cov\_hat=np.empty(covs.shape)
2. posterior\_sum=np.sum(posterior,axis=0)
3. for i in range(classes):
4. tmp=data-mean\_hat[i]
5. cov\_hat[i]=np.dot(tmp.T\*posterior[:,i],tmp)/posterior\_sum[i]

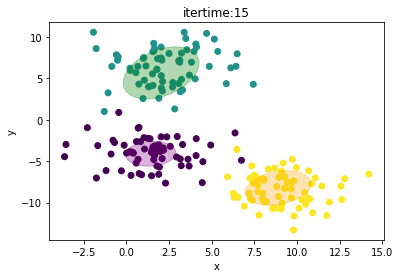
实验结果

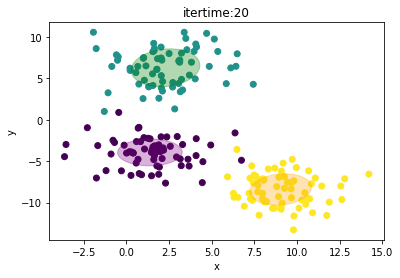
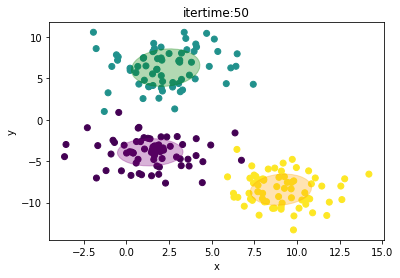
原始数据

EM过程









代码附录