哈尔滨工业大学计算机科学与技术学院

实验报告

课程名称: 机器学习

课程类型: 选修

实验题目: 实现 k-means 聚类方法和混合高斯模型

学号: 1190202110

姓名: 田雪洋

2021 年 11 月 5 日

一、实验目的

实现一个 k-means 算法和混合高斯模型，并且用 EM 算法估计模型中的参数。

二、实验要求及实验环境

1. 实验要求

目标：实现一个 k-means 算法和混合高斯模型，并且用 EM 算法估计模型中的参数。

测试：

用高斯分布产生 k 个高斯分布的数据（不同均值和方差）（其中参数自己设定）。

（1）用 k-means 聚类，测试效果；

（2）用混合高斯模型和你实现的 EM 算法估计参数，看看每次迭代后似然值变化情况，考察 EM 算法是否可以获得正确的结果（与你设定的结果比较）。

应用：可以 UCI 上找一个简单问题数据，用你实现的 GMM 进行聚类。

2. 实验环境

Windows10; python3.8.6;Pycharm

三、设计思想（本程序中的用到的主要算法及数据结构）

1.k-means 算法

聚类的对象是观测数据的集合。假设有 n 个样本，每个样本由 m 个属性的特征向量组成。样本集合可以表示为矩阵 X:

$$X = [x_{ij}]_{m\times n} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

矩阵的第 j 列表示为第 j 个样本，$j = 1, 2, ..., n$; 第 i 行表示为第 i 个属性，$i = 1, 2, ..., m$。

K-means 算法的聚类过程如下：

(1) 初始化。令 $t = 0$, 随机选取 k 个样本点作为初始聚类的中心 $m^{(0)} = \left(m_1^{(t)}, \cdots m_l^{(t)}, \cdots, m_k^{(t)}\right)$

(2) E 步：对于每一个样本 i，计算其应该属于的类，使得损失函数极小化，结果构成聚类 $C^{(t)}$:

$$\min_C \sum_{l=1}^{k} \sum_{C^{(t)}(i)=l} \left\| x_i - m_l^{(t)} \right\|^2$$

即，在类中心确定的情况下，将样本分到一个类中，使得样本和其所属类的中心之间的距离总和最小。

(3) M 步：计算新的类中心，对于聚类结果 $C^{(t)}$，计算当前各个类中的样本均值，作为新的类别中心 $m^{(t+1)} = \left(m_1^{(t+1)}, \cdots m_l^{(t+1)}, \cdots, m_k^{(t+1)}\right)$:

$$m_l^{(t+1)} = \frac{1}{n_l} \sum_{C^{(t+1)}(i)=l} x_i$$

2. 混合高斯模型

聚类的对象是观测数据的集合。假设有 n 个样本，每个样本由 m 个属性的特征向量组成。样本集合可以表示为矩阵 X:

$$X = [x_{ij}]_{m\times n} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

矩阵的第 j 列表示为第 j 个样本，$j = 1, 2, ..., n$; 第 i 行表示为第 i 个属性，$i = 1, 2, ..., m$。则对于上述 n 维空间中的随机变量 x_j 服从多元高斯分布，其概率密度函数为:

$$p(\mathbf{x} \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right) \tag{1}$$

其中 μ 是 n 维均值向量，Σ 是 $n*n$ 的协方差矩阵。则由其生成的高斯混合混合分布为:

$$p_{\mathcal{M}} = \sum_{i=1}^{k} \alpha_i p\left(\mathbf{x} \mid \mu_{\mathbf{i}}, \Sigma_i\right) \tag{2}$$

该分布由 k 个高斯分布组成，其中 μ_i 和 Σ_i 是第 i 个高斯分布的参数。且 $\aleph_i > 0, \sum_{i=1}^{k}\alpha_i = 1$。令随机变量 z_j 为生成样本 x_j 的高斯混合成分。显然 z_j 的取值未知，但 z_j 的先验概率 $P(z_j = j) = \alpha_j$，那么由贝叶斯定理得 z_j 的后验分布为:

$$p_{\mathcal{M}}\left(z_j = i \mid \mathbf{x_j}\right) = \frac{p\left(z_j = i\right) \cdot p_{\mathcal{M}}\left(\mathbf{x_j} \mid z_j = i\right)}{p_{\mathcal{M}}\left(\mathbf{x_j}\right)} = \frac{\alpha_i \cdot p\left(\mathbf{x_j} \mid \mu_{\mathbf{i}}, \Sigma_i\right)}{\sum_{l=1}^{k}\alpha_l p\left(\mathbf{x_j} \mid \mu_{\mathbf{1}}, \Sigma_l\right)} \tag{3}$$

则，当高斯混合分布已知时，高斯混合聚类将把样本集 X 划分为 k 个簇 $C = C_1, C_2, ..., C_k$。每个样本 x_j 的簇标记 λ_j 为：

$$\lambda_j = \arg\max_i p_{\mathcal{M}}\left(z_j = i \mid \mathbf{x_j}\right) \tag{4}$$

对于模型参数 $\{\alpha_i, \mu_i, \Sigma_i \mid i \in \{1, 2, \ldots, k\}\}$ 的求解，则采用极大似然估计，则其极大似然函数为：

$$LL(D) = \ln\left(\prod_{j=1}^{m} p_{\mathcal{M}}\left(\mathbf{x_j}\right)\right) = \sum_{j=1}^{m}\ln\left(\sum_{i=1}^{k}\alpha_i \cdot p\left(\mathbf{x_j} \mid \mu_i, \Sigma_i\right)\right) \tag{5}$$

然后 (5) 式对 μ_i,Σ_i 分别求偏导数，并令偏导数等于 0，得到：

$$\mu_i=\frac{\sum_{j=1}^{m} p_{\mathcal{M}}\left(z_j=i \mid \mathbf{x_j}\right) \cdot \mathbf{x_j}}{\sum_{j=1}^{m} p_{\mathcal{M}}\left(z_j=i \mid \mathbf{x_j}\right)} \tag{6}$$

$$\Sigma_i=\frac{\sum_{j=1}^{m} p_{\mathcal{M}}\left(z_j=i \mid \mathbf{x_j}\right) \cdot\left(\mathbf{x_j}-\mu_{\mathbf{i}}\right)\left(\mathbf{x_j}-\mu_{\mathbf{i}}\right)^T}{\sum_{j=1}^{m} p_{\mathcal{M}}\left(z_j=i \mid \mathbf{x_j}\right)} \tag{7}$$

对于混合系数 α_i，不仅要最大化 (5) 式，还要满足 $\aleph_i>0, \sum_{i=1}^{k} \alpha_i=1$。因此考虑 (5) 式的拉格朗日形式

$$LL(D)+\lambda(\sum_{i=1}^{k} \alpha_i-1) \tag{8}$$

其中 λ 为拉格朗日乘子，且满足 $\lambda=-m$。对该式求偏导得到：

$$\alpha_i=\frac{1}{m} \sum_{j=1}^{m} \frac{p\left(\mathbf{x_j} \mid \mu_i, \Sigma_i\right)}{\sum_{l=1}^{k} \alpha_l \cdot p\left(\mathbf{x_j} \mid \mu_l, \Sigma_l\right)} \tag{9}$$

GMM 算法的聚类过程如下：

(1) 随机初始化参数 $\{\alpha_i, \mu_i, \Sigma_i | i \in\{1,2, \ldots, k\}\}$ 以及 $\mathbf{C_i}=\emptyset$

(2) 开始迭代至带到迭代次数或者是参数值不再发生变化：

E 步: 根据式 (3) 计算每个样本由各个混合高斯成分生成的后验概率

M 步: 根据式 (6)(7)(9) 更新参数 $\{\alpha_i, \mu_i, \Sigma_i | i \in\{1,2, \ldots, k\}\}$

根据式 (4) 确定每个样本的簇标记 λ_j，并将其加入相应的簇 $\mathbf{C}_{\lambda_\mathbf{j}}=\mathbf{C}_{\lambda_\mathbf{j}} \cup\{\mathbf{x_j}\}$

(3) 输出簇划分 $C=\{\mathbf{C_1}, \mathbf{C_2}, \ldots, \mathbf{C_k}\}$

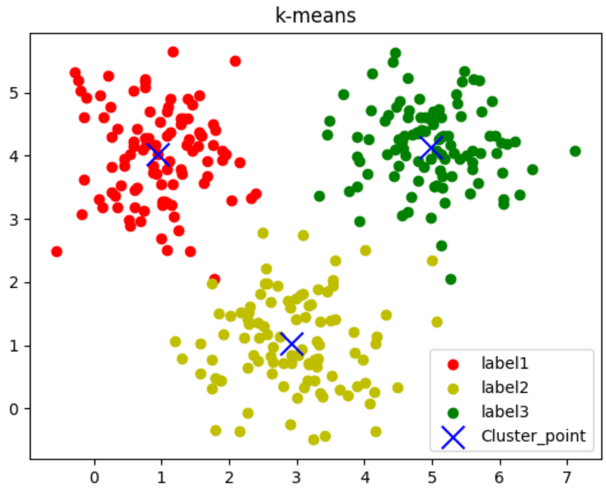
四、实验结果与分析

1. 生成数据

首先，为了方便可视化展现实验结果，本次实验选取二维高斯分布，并生成了三个二维高斯簇，进行测试。在这里，由于本次实验，事先知道每一个点所属的类别，因此，将分类后求出的类别，和已知类别进行对比，可以求出定义聚类准确率。

2.K-means

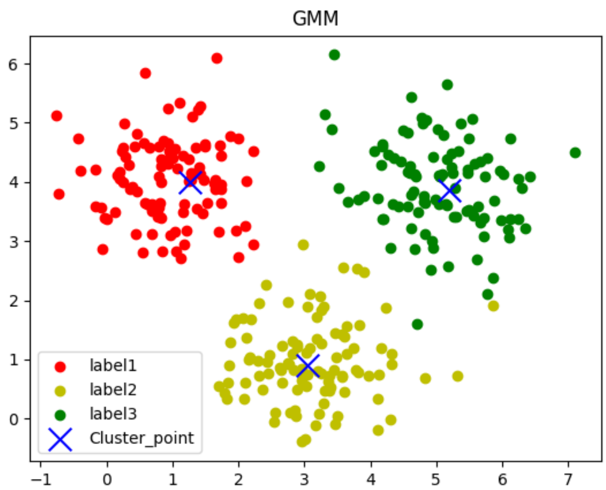
使用 K-means 算法得到的实验结果如下:



可以看到，K-means 算法可以很好地划分出不同的聚类簇，并且求出的簇中心的位置很合理。准确率为：99.3% 。

3.GMM

使用 GMM 算法得到的实验结果如下:



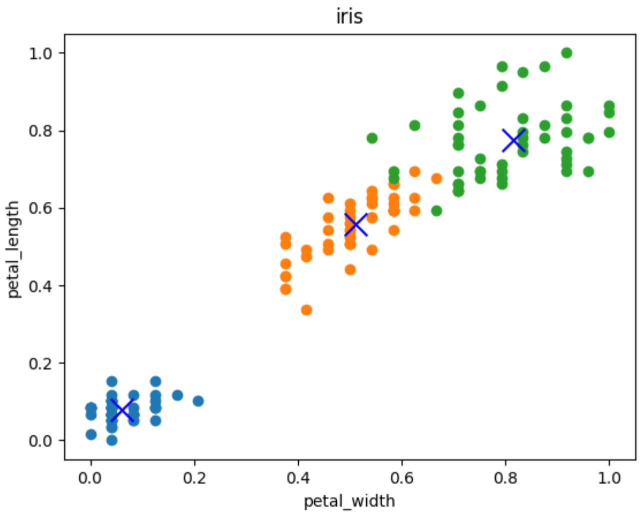
可以看到，GMM 算法可以很好地划分出不同的聚类簇，并且求出的簇中心的位置很合理。准确率为：94% 。

K-means 算法和 GMM 算法法在生成数据上的表现类似，都可以很高效准确地实现聚类。

3.UCI 数据集

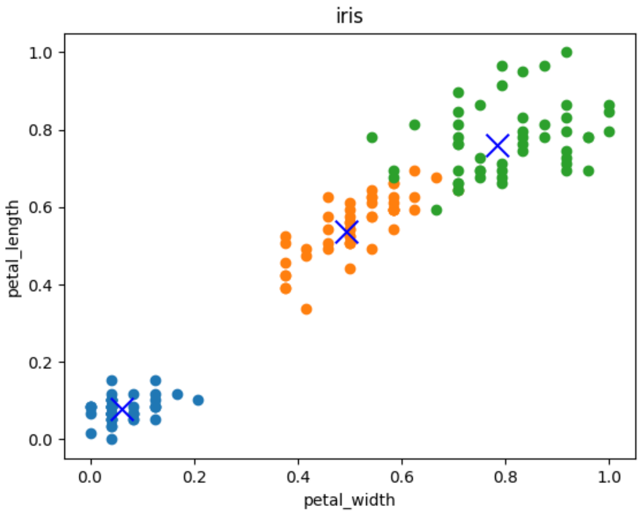
UCI 数据集中的 iris 数据集是简单经典的用于分类任务和聚类任务的数据集。本次实验采用 iris 数据集进行聚类任务测试。本次实验首先对鸢尾花数据集的标签数字化，经分析，选取 petal_length 和 petal_width 这两个特征进行聚类,K-means 算法和 GMM 算法的结果分别如下：

使用 K-means 算法得到的实验结果如下:



可以看到，K-means 算法可以很好地划分出不同的聚类簇，并且求出的簇中心的位置很合理。准确率为：96% 。

使用 GMM 算法得到的实验结果如下:



可以看到，GMM 算法可以很好地划分出不同的聚类簇，并且求出的簇中心的位置很合理。准确率为：98% 。

在实际数据集上，本实验算法依然能够很好地完成聚类任务。

五、结论

- K-means 算法利用欧式距离来衡量样本与各个簇中心的相似程度
- K-Means 的簇中心初始化对于最终的结果有很大的影响，如果选择不好初始的簇中心值容易使之陷入局部最优解
- GMM 使用 EM 算法进行迭代优化，因为其涉及到隐变量的问题，是在不完全数据上进行的聚类。
- EM 算法具备收敛性，但并并不一定找到全局最大值，有可能只能找到局部最大值。

六、参考文献

(1) 周志华著. 机器学习, 北京: 清华大学出版社, 2016.1

(2) 李航著. 统计学习方法，北京：清华大学出版社，2020.6

七、附录：源代码（带注释）

源代码见相关文件

(1) data_process.py: 数据预处理，生成, 计算准确率

(2) GMM.py：实现 GMM 算法

(3) k_means.py：实现 K-means 算法

(4) testOf_GMM.py：利用生成数据测试 GMM 算法，并绘图

(5) testOfK_means.py 利用生成数据测试 K-means 算法，并绘图

(6) iris_GMM.py：利用鸢尾花数据测试 GMM 算法，并绘图

(7) iris_k_means.py 利用鸢尾花数据测试 K-means 算法，并绘图