聚类模型实验报告

任俊彦　520021910546
电子信息与电气工程学院 自动化

目录

1 问题描述

1.1 问题设定

本次作业选择在 UCI 机器学习数据库中的鲍鱼数据集 (Abalone Dataset), 数据集链接 https://archive.ics.uci.edu/ml/datasets/Abalone。该数据集共 4177 例样本，每例样本含有 7 种特征，其样本数与特征数较适合于直接进行聚类而不必进行降维等操作。设定任务为：依据鲍鱼样本的各易测量的形态特征（长度特征、质量特征等），对鲍鱼进行聚类，以期将鲍鱼按成熟情况聚类，并探究聚类后各类与其实际成熟情况（数据集中给出）的关系。本次作业编写高斯混合模型对数据集进行聚类。

1.2 问题分析

聚类算法为无监督算法。该数据集给出了鲍鱼的实际成熟情况（年龄），在聚类过程中该指标不参与聚类而作为聚类性能的评价指标。在对该数据集设定如上节所述的问题时假定三类鲍鱼年龄差异足够可观，该假定是否正确，问题的设定是否足够合理，可以由聚类结果分析得到。数据集中各鲍鱼样本具有雌性、雄性的标签，聚类时选择全聚类或是依据该指标将两种鲍鱼分别聚类，需要经过尝试及分析确定。

2 数据清洗

在对数据代入模型求解前，需要处理空数据并剔除不合理数据。经分析，鲍鱼数据集中不存在空数据且数据数值区间均合理。

3 算法策略

3.1 矩阵运算

高斯混合模型采用 EM 算法极大化似然概率。在 M 步更新各成分的均值及协方差矩阵时，原始公式如下：

$$\boldsymbol{\mu}_i = \frac{\sum_{j=1}^{n} \gamma_{ji} \mathbf{x}_j}{\sum_{j=1}^{n} \gamma_{ji}} \tag{1}$$

$$\boldsymbol{\Sigma}_i = \frac{\sum_{j=1}^{n} \gamma_{ji} \left(\mathbf{x}_j - \boldsymbol{\mu}_i\right) \left(\mathbf{x}_j - \boldsymbol{\mu}_i\right)^T}{\sum_{j=1}^{n} \gamma_{ji}} \tag{2}$$

在代码实现中需要实现遍历求和以更新均值及协方差，直接使用循环的方式计算效率较低，引入哈达玛积（*numpy* 中的 $*$）后并将其写为矩阵运算可减少运算时间，表达式改写成

$$\boldsymbol{\mu}_i = \frac{\sum \boldsymbol{\gamma}_i * \mathbf{X}}{\sum_j \gamma_{ji}} \tag{3}$$

$$\boldsymbol{\Sigma}_i = \frac{\left(\boldsymbol{\gamma}_i * \left(\mathbf{X} - \boldsymbol{\mu}_i\right)\right) \left(\mathbf{X} - \boldsymbol{\mu}_i\right)^T}{\sum_{j=1}^{n} \gamma_{ji}} \tag{4}$$

其中 $\mathbf{X}$ 为输入数据集，具体实现代码见 $algorithm.GussianModel.\_Mstep()$.

3.2 参数初始化

高斯混合模型的参数初始化即为类分布概率及各类高斯分布的均值及方差的初始化。各类分布概率初始化一般以均匀分布作为策略。而各类高斯分布的均值及方差有多种初始化方式，如先采用简单的 kmeans 或其变种对均值中心进行初始化，或随机初始化。采用 kmeans 及其变种进行初始化可以减小初始点选取在稀疏区域或产生奇异协方差矩阵的可能性；而使用随机初始化则在实现上简洁高效。本次作业对均值采用随机初始化，选取各特征分量在数据集上最小值及最大值之间的随机值作为该特征分量的均值初始值。

3.3 奇异性和正则化

在高斯混合模型迭代的过程中，协方差矩阵可能出现行列式较小甚至为 0 的奇异情况，这是由于初始点选取不当或聚类数目而导致的。如下图所示[1]，当初始点选取不当或聚类数目过多时，某一类的样本点数目过少，导致其协方差矩阵难以估计，例如当某一类内只有一个点时，其协方差矩阵为零矩阵，即其为奇异阵，会造成其高斯概率无穷大，无法继续迭代。或当类内仅有少量点且其距离较近时，其协方差矩阵行列式仍接近于 0，可能导致数据过小而超出特定数据类型能表示的范围，造成溢出，尤其是当特征空间维数较高且某一数据点离群程度较大时可能产生该问题。解决此问题的一种方式是给协方差以一个较小的正则化项 $\lambda\mathbf{I}$，λ 一般取 10^{-6} 或更小的数，使在对其行列式值影响较小的情况下保证其行列式不为 0.

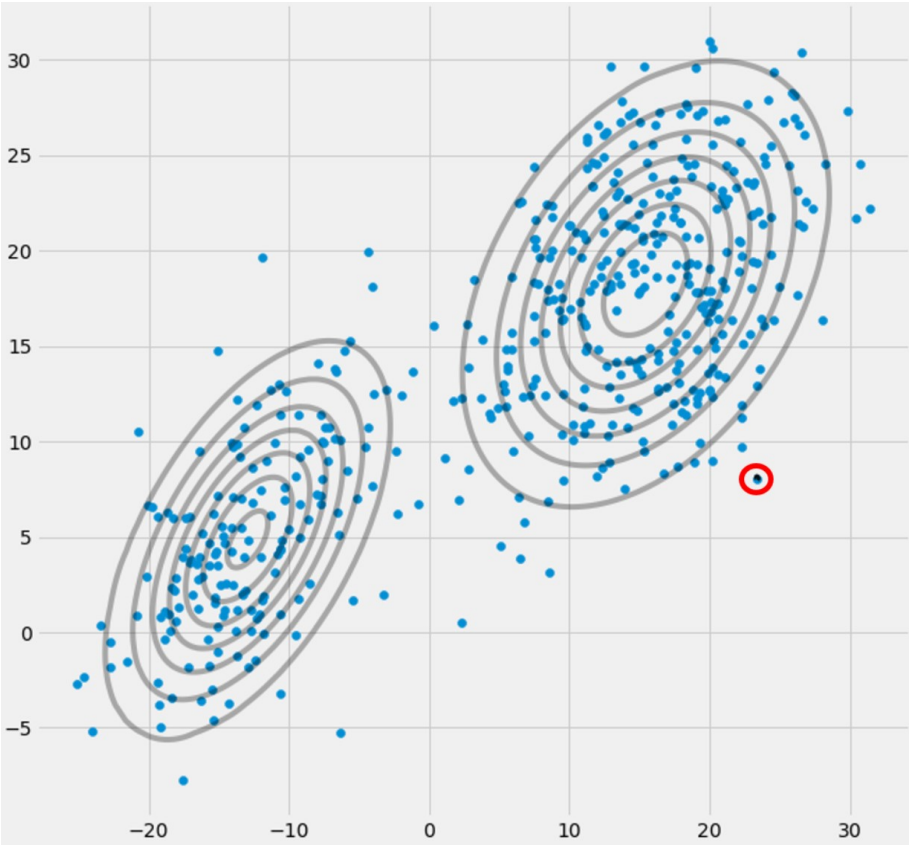


图 1: 奇异点图示，红圈为第三类高斯

[1] https://stats.stackexchange.com/questions/219302

另一方面，在本次作业的代码调试过程中还发现 4177 例数据点中的某几个离群数据点在各类高斯分布下的概率密度都极小，以至于造成下溢出的情况。经分析，应当是因为在特征空间维数较大数据时，高斯分布的概率密度函数可能在远离均值中心的区域内数值极小（如假设各维度独立，一维情况下的 0.01 概率密度对应七维情况下的 10^{-14}），当数值均极小时可能造成溢出。本次作业解决此问题的方法为，当某一离群数据点在各类高斯分布下的概率密度均极小（小于 10^{-30}）时，不计算该点的后验分布而直接将均匀分布作为其后验分布并继续迭代，以防止溢出。值得注意的是，处理这个问题的时候我对 sklearn 的源码进行了参考，发现 sklearn 居然直接采用忽略的方式处理了下溢出的异常抛出（然而 sklearn 对概率的计算始终使用对数运算，其一系列操作使少数离群点的各类后验概率均为 0 而非 nan，详见 $sklearn.mixture.\_base.\_estimate\_log\_prob\_resp()$）

3.4 聚类数及聚类性能

对于聚类数目及聚类性能的判定，在无标签的情况下有多种评判方式，如肘方法、轮廓系数、BIC/AIC 指标及交叉验证等。由于该数据集含有标签，并在设定任务时设定了针对该标签的任务，故在本次作业中直接以聚类结果在该标签（年龄）上的分布情况验证该模型在给定数据集及任务下的性能，并确定相应的聚类数目。

3.5 EM 迭代终止条件

本次作业中模型终止条件设定为达到最大迭代次数或各类高斯均值点变化距离小于 10^{-3}.

4 代码实现

4.1 代码结构

高斯混合模型的类实现在 *algorithm.py* 中，*main.py* 给出示例程序，可以直接运行得到本文中所有展示的结果，欢迎尝试。

4.2 类 API

高斯混合模型类接口定义如下：

1. 创建实例：gmm = GaussianModel $(data, num\_class, max\_iter = 30)$
2. 求解模型：gmm.Solve ()
3. 获得结果：gmm.Inference $(data = None, prob = False)$ -> array: class indices

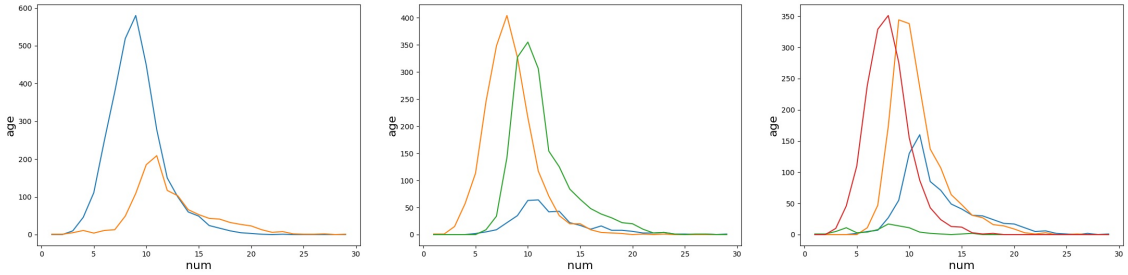
   其中 num_class 为聚类数，max_iter 为最大迭代次数。gmm.Inference() 若不给定数据，则默认返回用于求解模型的数据集的聚类结果；prob 为 True 时返回各数据的后验概率，False 返回各数据的聚类类别，即后验概率最大的类索引。

关于类的具体实现，请参见 *algorithm.py*，文件里写有详细的注释，此处不赘述。

5 模型表现

5.1 直接聚类

将 4177 例鲍鱼样本直接聚类，聚类数为 2-4 类，聚类结果如下。



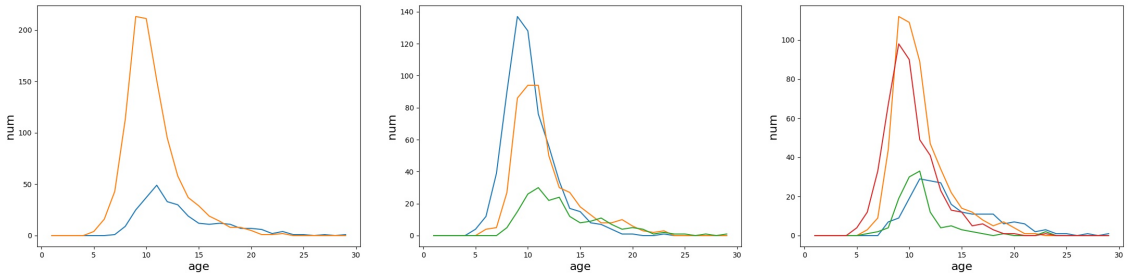
(a) 聚类数为 2　　(b) 聚类数为 3　　(c) 聚类数为 4

图 2: 直接聚类结果

可以观察到聚类结果在年龄特征方向上无显著差异，即聚类结果无法给出年龄维度上的正确分类（各图间颜色无对应关系）。考虑到该数据集的鲍鱼样本有雌雄两类，而雌雄个体在物理形态指标上应当有较大差异，故进一步对该数据集进行分类聚类，再考虑得出结论。

5.2 雌雄分类聚类

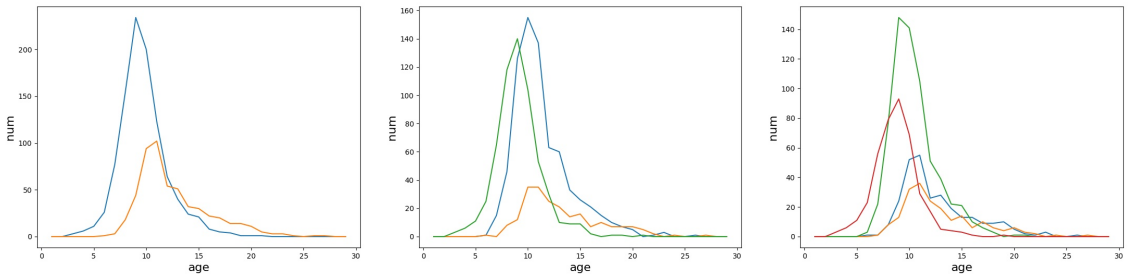
考虑到直接聚类的聚类结果在年龄特征方向上的分布无显著差异可能是由于雌雄鲍鱼存在物化形态指标上的差异而导致的类别混叠，故将鲍鱼按雌雄分开聚类，得到雌性聚类结果如下：



(a) 聚类数为 2　　(b) 聚类数为 3　　(c) 聚类数为 4

图 3: 雌性聚类结果

雄性聚类结果如下：



(a) 聚类数为 2　　(b) 聚类数为 3　　(c) 聚类数为 4

图 4: 雄性聚类结果

可以发现：按雌雄分开聚类并未明显提升聚类结果在年龄特征方向分布上的差异程度，这表明鲍鱼按特定标准对年龄分类后，各类在其余特征空间上的分布与高斯混合分布的假设有较大差异，若继续使用聚类方法以期获得与年龄分类结果相一致的聚类结果，应当考虑其它聚类模式，如谱聚类等；或重新考虑对于年龄聚类的指标选取。对于由高斯混合模型得到的聚类结果，可以用于进一步研究在年龄之外的其它特征方向上的聚类意义。