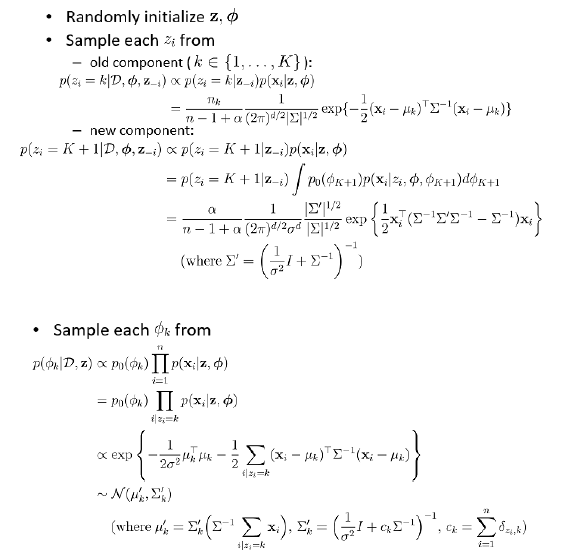
**机器学习 第三次仿真作业**

1. **题目简述**

基于中国餐厅过程(Chinese Restaurant Process，简称CRP)和Gibbs采样实现混合高斯模型(Gaussian mixture model，简称GMM)的参数估计。为了方便起见，GMM各个高斯分布的协方差为单位矩阵，其均值向量为待估计的参数。观察迭代过程中，聚类个数和与样本所属类别的均值向量的距离的变化，并研究CRP的参数alpha改变对聚类过程和结果的影响。

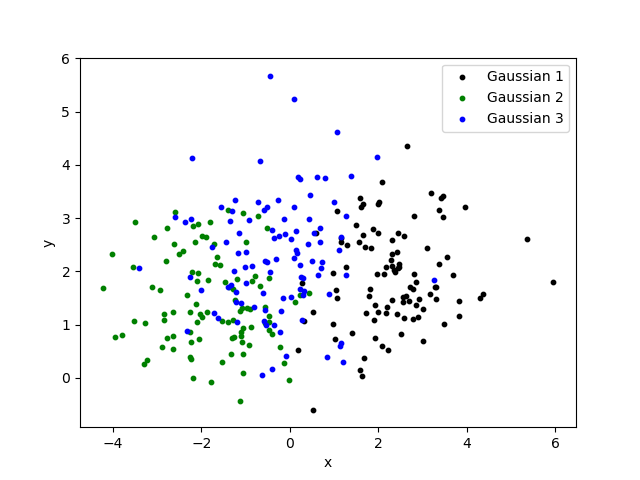
1. **实验过程简述**

原始数据分别由3个二维高斯分布产生，均值向量分别为(2.4, 2)、(-1.8, 1.4)和(-0.2, 2.6)，各100个数据点，一共300。CRP和Gibbs采样的实现过程已经在题目中给出了详细的说明，具体的python代码可参阅附件，有详细的注释。



1. **实验结果**

实验随机生成的数据点如图1所示，从三个高斯分布中产生的数据点分别标注了不同的颜色进行区分。



**图1 从三类高斯分布中随机生成的数据点**

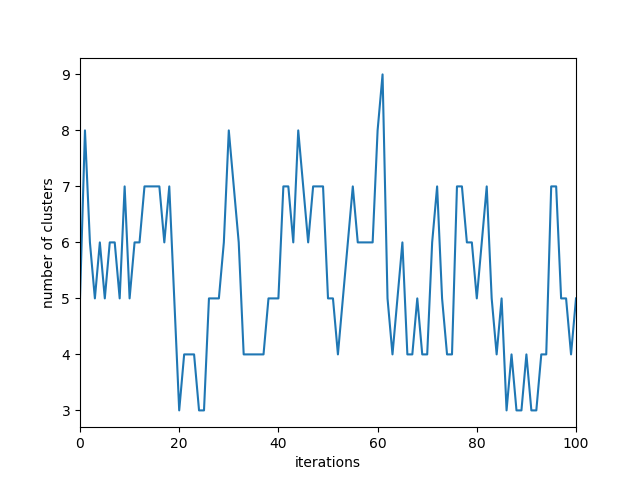
基于上一节描述的过程对数据点进行迭代，每一轮迭代包含了对隐变量**z**和参数的完整采样。实验分别对原始数据进行了30次、100次和300次迭代，各选取一次典型的结果由表1所示。

**表1 30、100、300次迭代后的聚类结果（alpha=1）**

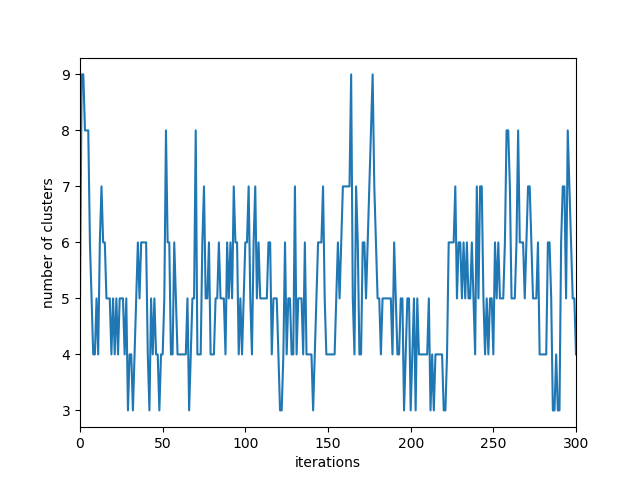
|  |  |  |  |
| --- | --- | --- | --- |
| **迭代次数** | **聚类数** | **每一类的参数** | **每一类的数据点数** |
| 30 | 6 | **(-0.364, 2.608)**  **(2.286, 1.994)**  **(-2.185, 1.248)**  (1.002, 2.094)  (1.616, -0.013)  (2.410, 0.015) | **100**  **94**  **82**  22  1  1 |
| 100 | 5 | **(1.810, 1.350) (2.418 , 1.943)**  **(0.014, 2.723)**  (-1.783, -0.081)  (1.396, 1.424) | **112**  **96**  **89**  2  1 |
| 300 | 4 | **(2.449, 2.040)**  **(-1.866, 1.550)**  **(-0.351, 2.571)**  (-0.738, -0.815) | **109**  **96**  **94**  1 |

从上表中可以看出，迭代后数据点最多的3个聚类同时也是与产生数据点的高斯分布的均值向量最接近的三个类，以下称之为**主类**，其参数和对应的数据点数在表格中加粗，其它的聚类称之为**旁类**。30次迭代时数据点还没有完全聚类，有相当部分的数据点没有归到主类中去。100次迭代时数据点基本都聚到了主类中，但主类的数据点数还不均衡。300次迭代时达到较为理想的结果。

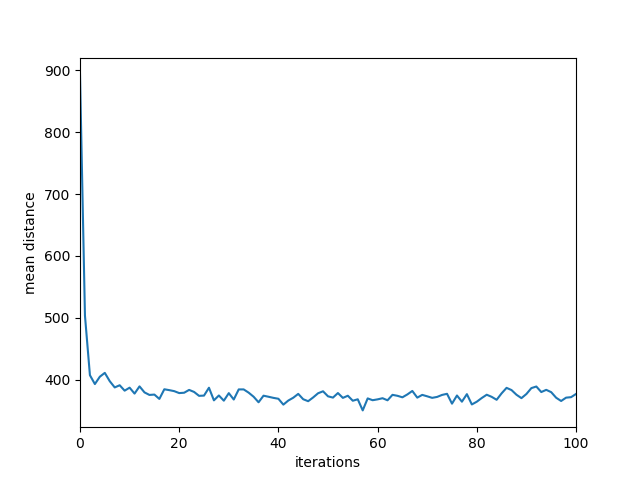
聚类的过程可以由聚类数以及与均值向量的距离两个量的变化进行可视化，如图2至图5所示。



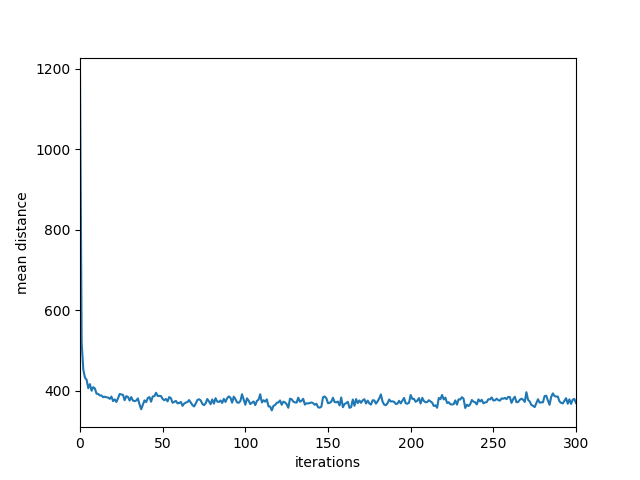
**图2 100次迭代过程中聚类个数的变化（alpha=1）**



**图3 300次迭代过程中聚类个数的变化（alpha=1）**



**图4 100次迭代过程中与均值向量距离的变化（alpha=1）**

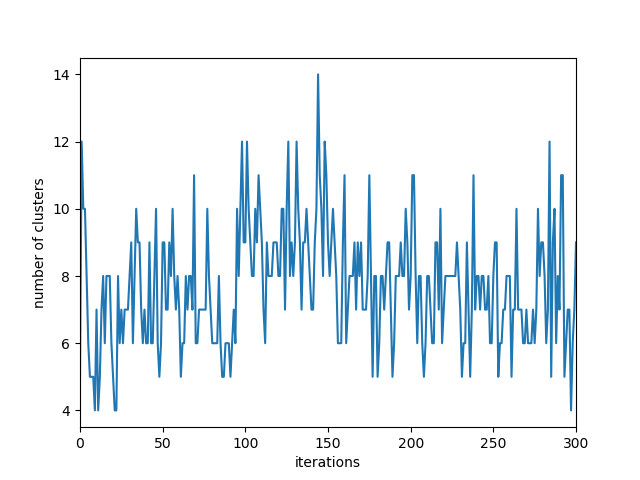
****

**图5 300次迭代过程中与均值向量距离的变化（alpha=1）**

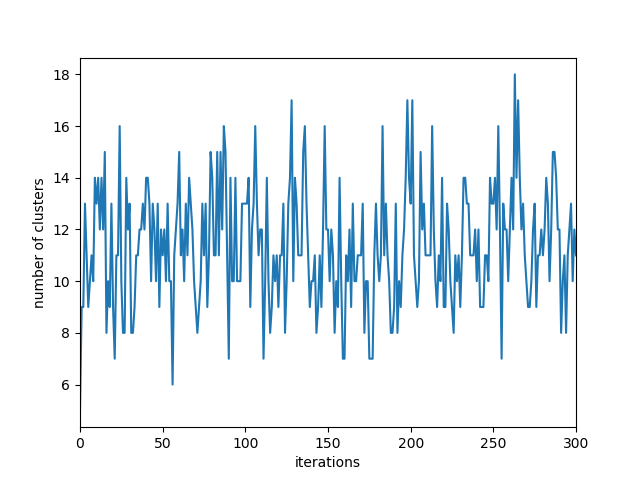
增大CRP的alpha值，将使得每次对**z**进行采样时产生新聚类的概率增大，导致聚类过程中聚类的个数产生更大的波动，聚类结果中数据点在各个聚类中的分布也将更加分散。当alpha值为3时，旁类的数量明显增加，但都呈散点分布，还不至于影响到主类的参数估计。而alpha值增大到5时，旁类的数量继续增加，且明显影响到了主类的数据点数，也导致了主类参数估计的偏差。

**表2 alpha值分别为1、3、5时的聚类结果（300次迭代）**

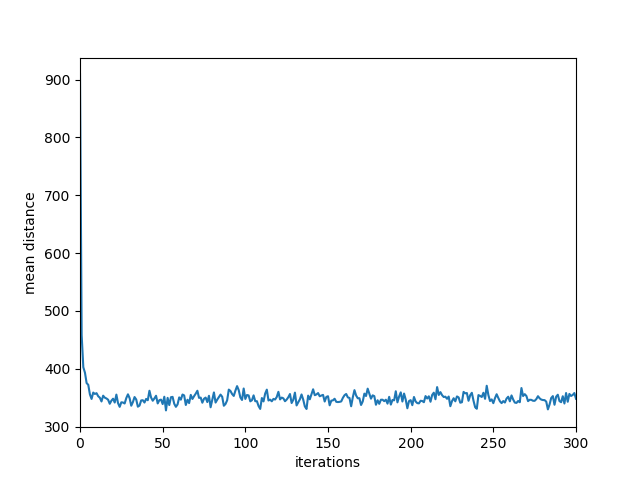
|  |  |  |  |
| --- | --- | --- | --- |
| **alpha值** | **聚类数** | **每一类的参数** | **每一类的数据点数** |
| 1 | 4 | **(2.449, 2.040)**  **(-1.866, 1.550)**  **(-0.351, 2.571)**  (-0.738, -0.815) | **109**  **96**  **94**  1 |
| 3 | 9 | **(-0.191, 2.627)**  **(2.288, 1.753)**  **(-1.843, 1.299)**  (-0.733, 0.945)  (-1.190, 0.628)  (-0.920, 2.506)  (1.258, -0.072)  (-2.338, -0.317) (0.208, -1.659) | **108**  **94**  **91**  2  1  1  1  1  1 |
| 5 | 11 | **(2.451, 1.949) (0.207, 2.258)**  **(-1.748, 1.507)**  (-1.123, 1.497)  (-1.885, 2.558)  (-0.410, -0.187)  (-1.226, -0.711)  (0.436, -1.074)  (-0.016, 0.345)  (-1.847, 0.761) (0.891, 0.524) | **93**  **65**  **64**  41  28  3  2  1  1  1  1 |



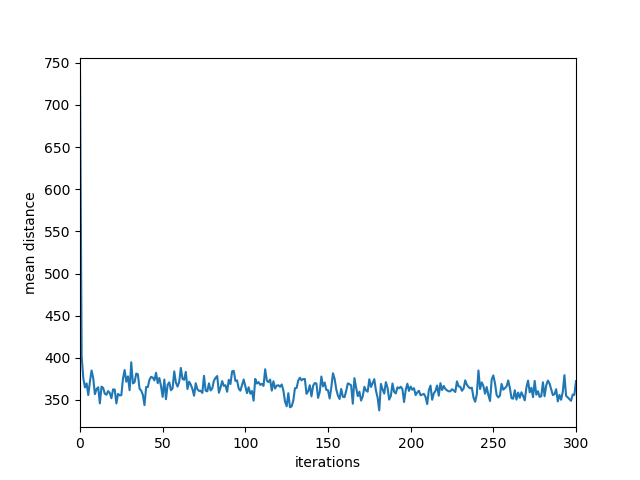
**图6 300次迭代过程中聚类个数的变化（alpha=3）**



**图7 300次迭代过程中聚类个数的变化（alpha=5）**



**图8 300次迭代过程中与均值向量距离的变化（alpha=3）**



**图9 300次迭代过程中与均值向量距离的变化（alpha=5）**

alpha值为3和5时聚类过程中聚类个数均值向量距离的变化如图6至图9所示。可以看出，随着alpha值的增大，聚类个数的波动增大，与均值向量距离的波动也明显增大。结合上文对聚类结果的分析，该场景下alpha取默认值1最合适。