Gaussian Mixture Model Classification

0130339024 罗志一

2014 Fall

1 Initialization of GMM

混合高斯模型的参数包括{μ, Σ, π, K},各参数的初始化过程如下：

- 首先，在训练数据中随机选取K个点作为**centroids**（中心点），其中K为*GMM* 模型的*component* 个数。
- μ, 使用*centroids* 的特征值作为数据的均值μ 的初始值。
- Σ, 计算各数据点到各*centroids* 的距离，将各个数据点分给最近的*centroids* 所属的*component* 中去（这步称为*hard assignment*）。然后计算每个*component* 的方差，作为Σ 的初始值。
- π, 在*hard assignment* 时将各数据分到了各自的*component* 中去，计算各*component*中包含数据点的数目占所有数据数目的比例，即可得到参数π 的初始值。
- K, 该参数的设置需要开发数据*dev.txt*。在实验中，分别将类别1和类别2的GMM模型的K设置为{1,2,3,4,5}在开发集上测试后发现$K = 4$时效果最好。后将训练集*train.txt*的二维数据打印出来观察发现两个类别的混合高斯模型均取$K = 4$ 最为合理（图见Figure 1）。

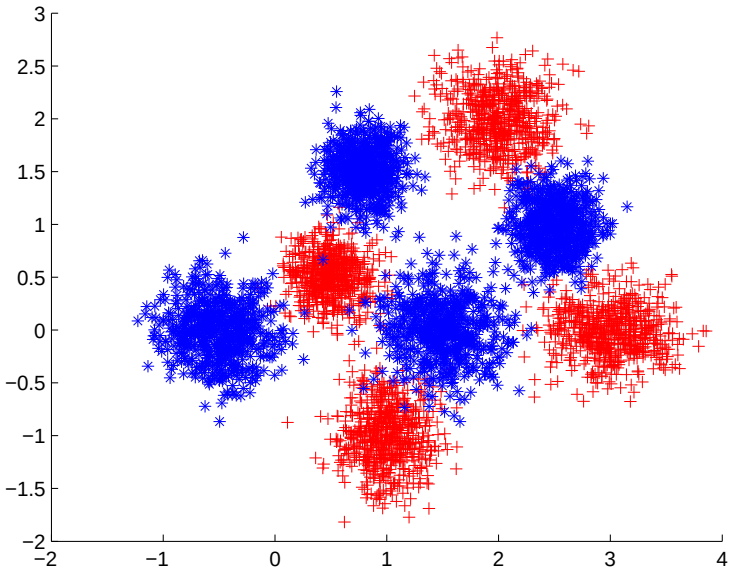

Figure 1: plotting train data

2 GMM Parameter Tuning

实验使用*EM* 算法进行GMM模型的参数调整。

- 实验使用的GMM模型*log-likelihood*函数如下:

$$\sum_{i=1}^{N} log\{\sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k)\}$$

- 实验将收敛的*threshold*(阈值) 设为$1e-7$, 并将迭代次数上限设为200。
- 最后在*dev.txt* 上测试的分类精确度为97.75%. （由于在参数初始化时*centroids*是随机选取的，所以每次运行结果有细微差别，但精度都在95% 以上）

3 Analysis

这里讲一下对实验中遇到的问题的一些分析:

- 为调试代码，我自己编写了一个测试集，只有5个数据点.就在这个小的测试集上运行代码时就出现了*Singular Matrix*这样的错误。后经分析发现原因可能是由于数据不足，采用了正则化的方法来处理。即在每次算完Σ矩阵后，都给它加上一个对角矩阵λI，其中λ为很小的数。在本实验中λ取$1e-3$。具体的实现可参见代码*gmm3.py* 中的*regular*函数。

4 Conclusion

说明及总结:

- 程序在测试集*test.txt*的运行结果为*result/result.txt*。其他说明参见*README.txt*文件，就不在报告中累述。
- 通过本次实验掌握了高斯混合模型在分类问题中的具体应用，并实现了*EM*算法，收获很大。