深蓝学院
shenlanxueyuan.com

第三章作业思路讲解

主讲人 张点堃

纲要

- 第一部分：思路分析
- 第二部分：作业常见问题及解决思路

KMEANS——思路分析

- Kmeans的实现思路比较简单，直接参照PPT上的步骤即可。

- Fit函数
  - 找到所有点的所归属的聚类
  - 根据每个聚类中的点计算对应的均值，更新参数
  - 重复上述步骤，直到收敛/最大迭代数
- Predict函数
  - 直接通过聚类中心与数据点的距离判断点的类别归属

GMM——思路分析

深蓝学院 shenlanxueyuan.com

- GMM的实现略比Kmeans复杂，但也可以参照PPT上的步骤。

- Fit函数
  - E-step：计算每个数据点属于每个高斯分布的似然概率
  - M-step：根据上述似然概率更新各高斯分布的参数（均值和协方差由上述似然概率加权得到）
  - 重复上述步骤，直到收敛/最大迭代数
- Predict函数
  - 计算每个点属于每个高斯分布的似然概率，取最大作为此数据点的分类结果

谱聚类——思路分析

- 谱聚类由于不能对新加进来数据进行单独的分类，所以Fit和Predict函数其实是一样的（或者不设置predict函数）。
- Fit函数
  - 建立邻接矩阵，选择合适的亲和度度量（全1，欧式距离倒数，或者高斯核函数）
  - 计算对应的度矩阵D和拉普拉斯矩阵L
  - 对拉普拉斯矩阵做归一化（可选）
  - 对拉普拉斯矩阵做特征值分解，取前K小的特征值对应的特征向量，组成N*K的矩阵
  - 以上述矩阵的行作为数据点的表示，做Kmeans。得到分类结果。

KMeans——作业常见问题

- Kmeans的实现相对简单，从作业提交来看基本都没有问题

- 有些同学的Kmeans结果中各类别的大小差异比较大，这往往是由于随机初始化均值得到的均值不理想导致的。可以尝试使用FPS（最远点采样）等方法得到空间上均匀分布的初始化均值点。

GMM——作业常见问题

深蓝学院 shenlanxueyuan.com

- GMM比KMeans更复杂一些，同学们遇到的问题多一些。
- GMM同样存在受到初始化影响的问题。与Kmeans不同的是，GMM初始化参数多了一个协方差。均值初始化点可以同样使用FPS采样，而协方差可以使用数据总体的协方差来进行初始化，这样初始化的高斯分布就会和数据总体的分布更接近，更容易拟合。尤其是作业中的第四张图。
- GMM很多实现中遇到的问题大多由于理论没有搞懂，概念不清晰实现很容易出错
- GMM中大部分运算都可以使用矩阵/向量运算，这样可以显著提高程序运行速度，而不要逐个数据点的去跑循环
- 在程序还没跑对的情况下，可以先不考虑停止条件的问题，使用较大的，固定次数的迭代求解。这样更容易找到程序的问题，在程序确保正确后，再引入停止条件加快求解速度。

谱聚类——作业常见问题

深蓝学院 shenlanxueyuan.com

- 谱聚类是作业问题最多的部分。
- 邻接矩阵的建立：推荐使用KNN来建立数据点之间的边，全连接在数据点数量增多的情况下会失效，KNN的近邻数K不宜太大或者太小，这点大家在作业中可以慢慢调试。
- 亲和度度量：大多数同学使用距离倒数的方式，一般来说，使用高斯核函数会有更好的效果。
- 拉普拉斯矩阵归一化：是否使用归一化，使用哪种归一化都可以得到正确的结果。需要注意的是，Lrw不是对称矩阵，不能使用eigh方法进行分解，此方法是对对称矩阵的快速特征值分解方法。而L和Lsys都是对称矩阵。
- 在完成作业过程中，一定先保证自己的Kmeans写对了。因为谱聚类中要调用自己写的Kmeans。

在线问答

深蓝学院
shenlanxueyuan.com

Q&A

深蓝学院
shenlanxueyuan.com

感谢各位聆听！

Thanks for Listening