大 数 据 分 析

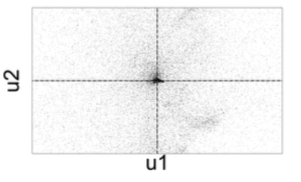
2019秋

Homework #1

1. 比较说一下 Random Forest 与 GBDT的共同点与区别？

提示：从两个方法的目标函数，学习参数、预测等方面阐述

2. 写程序利用SVD分解对数据进行降维，并画出降维后各个节点在”新维度“上值的分布。例如对于矩阵 A(mxn) = , U 分解后每一列 有m个元素对应m个点，记为. 我们画出**-**的坐标下的m个点的散点图。如图：



我们把这个图称为：Spectral Subspace Plotof and**。**

数据集1：<https://github.com/shenghua-liu/HoloScope/blob/master/testdata/>

下面的[yelp.edgelist.gz](https://github.com/shenghua-liu/HoloScope/blob/master/testdata/yelp.edgelist.gz" \o "yelp.edgelist.gz)

1）数据的格式是：用户id，饭店id，1

2）所有的id都已转换到从0开始的整数

提示：1）降维的矩阵是：用户x饭店，需要先构造矩阵；考虑稀疏矩阵，否则内存可能会溢出

2）降维的维度数选择10

3）考虑稀疏矩阵的svds工具

数据集2：自由选择不少于10000x10000的数据集。

或者公开数据集：<http://konect.uni-koblenz.de/networks/>

<http://jmcauley.ucsd.edu/data/amazon/>

<http://snap.stanford.edu/data/index.html>

作业提交结果：

两个数据集的结果，包括

1）每一个数据集的Spectral Subspace Plots，**-, -,…., -, 以及 -, -,…., -** 共10副图

2）20副图片一行两个，贴入word中。按图例著名横纵坐标、以及例子的名字和链接（如果数据集2是网上公开的）。

3）相应的代码