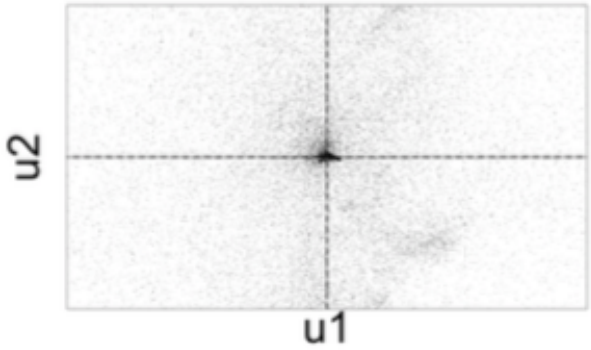
大 数 据 分 析

2021 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) = UΣV^T, U 分解后每一列 $\mathbf{u_i}$ 有 m 个元素对应 m 个点，记为 u_{i*}. 我们画出 $\mathbf{u_1}$- $\mathbf{u_2}$的坐标下的 m 个点的散点图。如图：



我们把这个图称为：Spectral Subspace Plot of $\mathbf{u_1}$and $\mathbf{u_2}$。

数据集 1：https://github.com/shenghua-liu/HoloScope/blob/master/testdata/
下面的 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 K=10
3）考虑稀疏矩阵的 svds 工具
数据集 2：自由一个选择不少于 10, 000 x 10,000 个 id 的数据集
链接：http://konect.uni-koblenz.de/networks/
http://jmcauley.ucsd.edu/data/amazon/
http://snap.stanford.edu/data/index.html
或自己领域的大规模数据

作业提交结果：

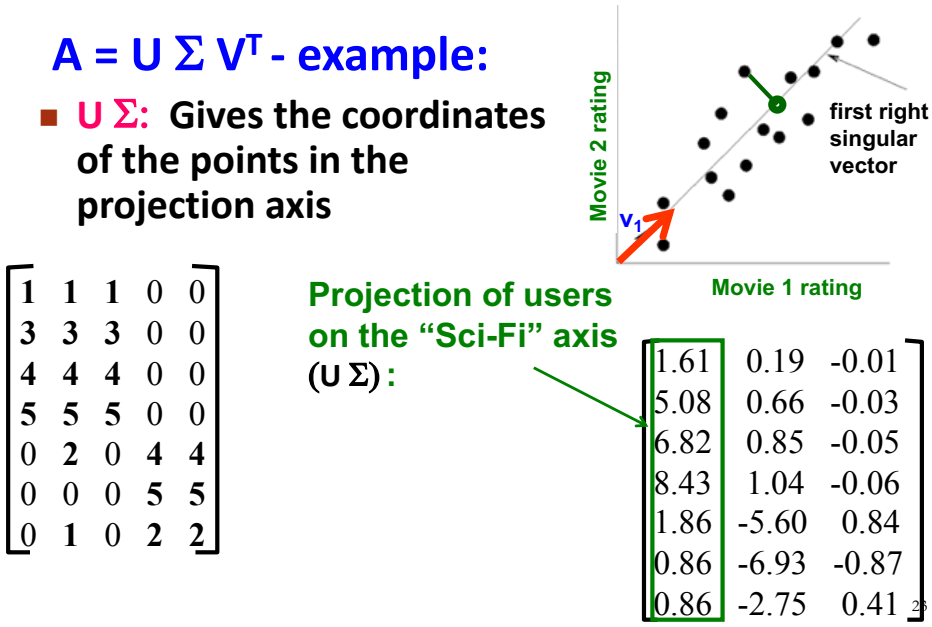
两个数据集的结果，包括

1）每一个数据集的 Spectral Subspace Plots，$\mathbf{u_1}$-$\mathbf{u_2}$，$\mathbf{u_2}$-$\mathbf{u_3}$,….，$\mathbf{u_9}$-$\mathbf{u_{10}}$，以及 $\mathbf{v_1}$-$\mathbf{v_2}$，$\mathbf{v_2}$-$\mathbf{v_3}$,….，$\mathbf{v_9}$-$\mathbf{v_{10}}$ 共 18 副图

2）20 副图片一行两个，贴入 word 中。按图例注明横、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

3. 课程中的 slides



证明或说明为什么 U Σ 中的第一列是矩阵 A 的行坐标点在第一个右奇异向量 v1 上投影的坐标。（提示：空间中一个点 p 在向量 v1 上的投影如何表示）