

大数据分析

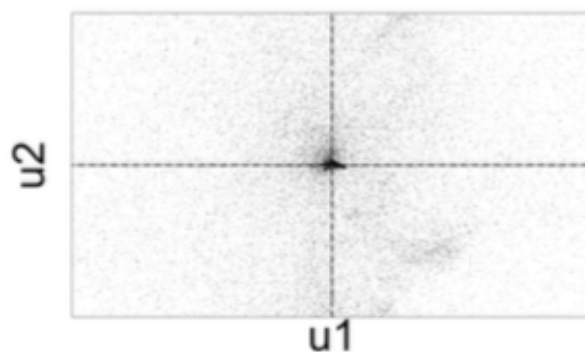
2021 秋

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 $A(m \times n) = U\Sigma V^T$, 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 \times 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 \times 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, $u_1-u_2, u_2-u_3, \dots, u_9-u_{10}$, 以及 $v_1-v_2, v_2-v_3, \dots, v_9-v_{10}$ 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

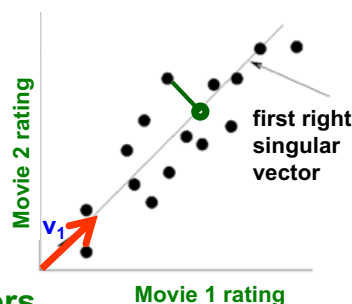
$A = U \Sigma V^T$ - example:

- **$U \Sigma$:** Gives the coordinates of the points in the projection axis

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix}$$

Projection of users
on the “Sci-Fi” axis
($U \Sigma$):

$$\begin{bmatrix} 1.61 & 0.19 & -0.01 \\ 5.08 & 0.66 & -0.03 \\ 6.82 & 0.85 & -0.05 \\ 8.43 & 1.04 & -0.06 \\ 1.86 & -5.60 & 0.84 \\ 0.86 & -6.93 & -0.87 \\ 0.86 & -2.75 & 0.41 \end{bmatrix}$$



证明或说明为什么 $U \Sigma$ 中的第一列是矩阵 A 的行坐标点在第一个右奇异向量 v_1 上投影的坐标。（提示：空间中一个点 p 在向量 v_1 上的投影如何表示）