第7次作业 特征选择与特征提取

1.

MDS（Multidimensional Scaling）方法通常被用于在二维或三维上可视化地显示一组复杂样本之间的关系，我们收集到8个城市铁路交通的通勤时间，单位：小时，如表1。这里认为通勤时间矩阵是对称矩阵。根据通勤时间，请用MDS的方法得到城市的二维表示并作图，然后简要分析与真实地图上各个城市相对位置的差异。

表 1 铁路交通通勤时间

城市	武汉	郑州	北京	周口	运城	十堰	汉中	重庆
武汉	0	1.75	4.25	6	15.50	3.66	12.75	6.35
郑州	1.75	0	2.5	3.25	8.33	8.15	16.07	8.25
北京	4.25	2.5	0	11.75	5.57	15.75	27.1	12
周口	6	3.25	11.75	0	17	16.75	24.33	17.07
运城	15.50	8.33	5.57	17	0	9.66	11	14.9
十堰	3.66	8.15	15.75	16.75	9.66	0	6.15	8.7
汉中	12.75	16.07	27.1	24.33	11	6.15	0	11
重庆	6.35	8.25	12	17.07	14.9	8.7	11	0

2.

在提供的手写数据集（MNIST的一部分）上测试PCA、t-SNE、LLE降维方法中的任意两种。具体要求如下：

a. 给出各算法的简要流程；

b. 数据降维到2维的可视化展现；

c. 选择一种合适的分类器和较合适的降维维数，分别对降维前后的数据进行训练和测试，比较降维对分类效果的影响；

d. 分析、对比两种降维方法在该数据集上的降维效果。

t-SNE：https://lvdmaaten.github.io/tsne/

LLE：https://www.cs.nyu.edu/~roweis/lle/code.html

上述链接代码可直接使用。

3.

有一组肿瘤组织(1)和正常组织(0)的数据，feature_selection_X.txt 是特征数据，一行代表一个样本；feature_selection_Y.txt 是标记。请设计特征选择算法，挑选出有利于区分不同组织的特征，比较特征选择前后对分类器分类效果的影响。

要求：

a. 请自行选择合适的可分性判据；

b. 请自行选择合适的分类器；

c. 给出你所挑选特征的个数以及这些特征对应的列数；

d. 请使用十倍交叉验证的方法给出分类器的错误率；

e. 与不做特征选择的分类器错误率做对比，并简要分析结果。

注意：提交的作业中请不要包含原始数据集(数据量太大)。