# 同济大学软件学院 2018年春季数据分析/挖掘课程作业之1

1. 使用Pythong lshash 0.0.4dev包 (https://pypi.python.org/pypi/lshash/0.0.4dev)处理利用piazza中reco\_data.rar的trade文件，查找某个输入用户的knn结果。
   1. 将trade文件中的记录按照vipno分组，汇总每个vipno购买pluno的总记录数；将以上汇总的vipno分组转为一个矩阵，其中**矩阵行**代表为pluno，**矩阵列**代表为vipno，**矩阵元素**为该vipno所购买pluno的总金额（取整，四舍五入）；
   2. 如上处理之后，将每个矩阵列通过Pythong lshash进行索引处理，hash\_size为全体vipno(非重复)总数的0.01, 0.05, 0.1, 0.2, 0.3, 0.5; 任意选择一个vipno, 然后输出该vipno对应knn的输出vipno (k = 1, 2, 3, 4, 5).
2. sklearn k-mean算法聚类; 通过Silhouette系数评价k-means聚类质量，以求解最佳的聚类数量k。初始值k设置为k ≈√n/2，其中n为vipno总数，通过不断调整k的值，计算kmeans聚类结果的Silhouette系数，求得最有的聚类数量k。要求：
   1. 以k值为横轴、Silhouette系数值为y轴，画出Silhouette系数值-k值的函数图；
   2. 验证上述lsh的knn查询结果是否与输入vipno所在同一个簇，并进行讨论分析；

可以参考[SKLearn kmeans demo](http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html#sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py)： 以及Silhouette系数的wikipedia说明内容：https://en.wikipedia.org/wiki/Silhouette\_(clustering)

1. 通过sklearn DBSCAN算法进行如上的vipno聚类操作，变化eps的值，并计算Silhouette系数，以寻求最优eps值。要求：
   1. 以k值为横轴、Silhouette系数值为y轴，画出Silhouette系数值-k值的函数图；
   2. 验证上述lsh的knn查询结果是否与输入vipno所在同一个簇，并进行讨论分析；
   3. 比较DBScan和kmeans二者的最优聚类结果，并加以讨论和比较。

可以参考[SKLearn DBScan demo](http://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html#sphx-glr-auto-examples-cluster-plot-dbscan-py)

1. 通过SKLearn的GaussianMixture模型(GMM)聚类如上的vipno聚类操作, 并比较GMM和KMeans、DBSCan的聚类结果；在与Kmeans进行比较时，GMM的n\_components可与Kmeans的最佳k值相同；在与DBScan进行比较时，GMM的n\_components可与DBScan的最佳epsk值对应的簇个数相同。求解：
   1. 假定以Kmeans作为真实的聚类结果，计算GMM的准确率(accuracy);
   2. 假定以DBScan作为真实的聚类结果，计算GMM的准确率(accuracy);
   3. 验证上述lsh的knn查询结果是否与输入vipno所在同一个簇，并进行讨论分析；

可以参考[SKLearn GMM demo](http://scikit-learn.org/stable/auto_examples/mixture/plot_gmm_covariances.html#sphx-glr-auto-examples-mixture-plot-gmm-covariances-py)

提交日期：2017/04/08日 23：59PM，提交内容发送至tongjidam18@163.com，提交内容包括：

1. 每个作业提交内容以学号+hw1.zip作为命名方法；其中包括4个子目录，命名方式分别为q1,q2,q3和q4，每个子目录包括对应目的代码和word报告。其中报告包括1）代码运行结果屏幕拷贝；2）讨论分析部分；3）性能比较图表