# 软件学院 数据分析/挖掘课程作业之2

1. 利用作业#1提供的1000条栅格化轨迹数据,通过sklearn k-mean算法聚类; 通过Silhouette系数https://en.wikipedia.org/wiki/Silhouette\_(clustering)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=1000为轨迹数量，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. 将最优k值的轨迹聚类结果，投影到百度地图进行可视化，不同聚类的轨迹使用不同的颜色进行标记区分；

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)：

1. 通过sklearn DBSCAN算法进行如上的1000条栅格化轨迹数据, 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. 将最优k值的轨迹聚类结果，投影到百度地图进行可视化，不同聚类的轨迹使用不同的颜色进行标记区分；
   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1000条栅格化轨迹数据, 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);

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)

1. 提交日期：2017/04/09日 23：59PM