**中作业2：智能电表数据分析**

1. 作业概况

智能电表数据分析可以帮助电力公司理解负荷需求，辅助需求侧管理或电力营销工作。聚类是数据分析中常用的无监督学习算法，可以将相似的样本聚为一类，被广泛用于用户画像等工作。本作业带领大家体验负荷曲线聚类的过程：根据所给的聚类分析流程和所提供的数据，编写聚类程序对不同用户的负荷曲线进行聚类并展示结果。

1. 作业内容
2. 数据集说明

此次作业提供了1000条单日（96点）负荷数据，见data\_set.xlsx。需要注意的是，由于各种原因，数据可能出现缺漏，需要进行一定的数据预处理，例如：直接以前后两个点的平均值替代缺省值。

1. 聚类分析流程

聚类流程通常包括特征提取、距离衡量和聚类算法三步。首先，需要选择有效的数据特征集（由多个特征组成），以便将不同的用户区分开，比如最大负荷、负荷率等，可以参考ref\_1（表1）；其次，采取合适的距离衡量方法，衡量不同用户特征之间的相似度，比如欧氏距离、余弦距离等；最后是选择合适的聚类算法，给用户标记不同的标签，常用的有算法有k-means、DBSCAN等，MATLAB中有相关包可以调用，请自行学习使用方法。

1. 用户特征选取与权重配置

特征选取可以选取一些有实际物理意义的特征，如负荷率、最大负荷等，也可以通过降维算法如主成分分析法，将原有的高维特征压缩为低维特征，获取一些没有实际意义的特征。为了简单起见，默认可以使用原始的、具有物理意义的特征。感兴趣的同学也可以使用特征权重配置方法来使得用户的重要特征得到体现，可以参考文献ref\_2。

1. 具体任务
2. 给出所选择的用户负荷曲线的特征集（有实际意义的要给出每个特征的公式，没有实际意义的要给出对应的降维算法），选取的特征不少于6个。
3. 给出使用的衡量距离、聚类算法名称和聚类算法对应输入参数的确定过程（使用某些聚类算法如k-means时，需要输入聚类数目，此时应采用合适的判据来确定聚类数，给出最优的聚类数目；而有些算法可以自动确定聚类数目，如DBSCAN，但也需要给出合适的距离参数，此时也应当采用合理的判据，给出最优的距离参数）。
4. 给出测试表中1000条负荷曲线的聚类结果，要求以图表等形式清晰展示。
5. 附加题：如果想利用智能电表数据评估用户参与需求响应的潜力，以便从大量用户中选取适合开展需求响应的用户，应当如何选择特征集？请大家阅读文献（可以参考ref\_3中2.1节内容及所列文献）并结合自己的思考，分析面向需求响应的智能电表数据特征选择和潜力评估的方法以及各自的优缺点。视完成情况，可以有1~5分的额外加分。