CSDS 340: Introduction to Machine Learning
Spring 2024
Case Study 2

Assigned: April 12, 2024
Due: May 8, 2024, before 11:59 pm.
Total Points: 100

1 Objective

To practice clustering in a real world problem, this case study considers automatic identification system (AIS) data generated by maritime vessels. Please refer to `https://www.navcen.uscg.gov/automatic-identification-system-overview`. Each vessel has a maritime mobile service identity (MMSI) number that uniquely identifies each vessel. A vessel's AIS unit periodically sends a report on its position. The report contains the vessel's MMSI so that the vessel can be tracked over time. In this case study, your job is to track the movements of the different vessels given position reports without the MMSIs. In other words, you must predict which vessel each report came from given multiple reports from each vessel over time.

2 Data

Each row of the data corresponds to an observation of a single maritime vessel at a single point in time. The column descriptions are as follows:

- `OBJECT_ID`: A unique ID assigned to each observation (position report).
- `VID`: Anonymized version of the vessel's maritime mobile service identity (MMSI) number, which uniquely identifies each vessel. ***Note: The VID uniquely identifies a vessel within a single data set but are not consistent across data sets.*** For example, the VID of 100001 in two different data sets, which represent two different time periods, may not refer to the same vessel.
- `SEQUENCE_DTTM`: Time of reporting in the form of hh:mm:ss in Coordinated Universal Time (UTC), where hh is hours, mm is minutes, and ss is seconds. The date information has been removed so that only time information is available.
- `LAT`: Latitude of vessel position in decimal degrees.
- `LON`: Longitude of vessel position in decimal degrees.
- `SPEED_OVER_GROUND`: Vessel speed in tenths of a knot (nautical mile per hour), up to a saturation limit of 1022, which indicates that the speed is 102.2 knots or higher.
- `COURSE_OVER_GROUND`: Angle of vessel movement2 in tenths of a degree, with a range between 0 and 3599 (359.9 degrees). The course over ground is still reported even when the speed over ground is 0 because vessels always have slight movements due to currents and other factors.

This case study will involve 3 data sets, each consisting of AIS data collected from the same area at the same time of day, but over 3 different days. You are provided with "set1.csv" and "set2.csv" with VIDs, and "set3noVID.csv" without VIDs. The final evaluation will take place on data "set3.csv" which is only available for TAs.

3 Models

To implement clustering, using your own algorithms or the functions provided by scikit-learn. Please refer to https://scikit-learn.org/stable/modules/clustering.html for more details.

4 Submissions

Please only submit two files, "groupNumber_CS2.pdf" and "groupNumber_predictVessel.py".

4.1 Experiment Report (50%)

- (15%) Discuss your chosen clustering algorithms and parameters that you will vary for implementing clustering. And what other algorithms you compared against.
- (15%) Your method to solve the problem of unknown number of clusters for "set3noVID.csv".
- (10%) Any feature selection and pre-processing you performed.
- (10%) Insightfulness and clarity of your observations and discussions. (Please be free to add the approaches you tried but failed before arriving at the best solution.)
- The report should be in PDF format and no more than 5 pages. Please use "groupNumber_CS2.pdf" as the file name.

4.2 Code (50%)

4.2.1 Template script

You are provided with a template script named "template_predictVessel.py". In the script, the function "predictor_baseline" is the implementation of the baseline predictor which uses all features except for OBJECT_ID, standardization of data prepossessing and k-means with fixed K=20(the number of unique VIDs of set1). The function "predictor" is where to implement your own predictor and returns your predicted array of cluster numbers of each sample. We will use adjusted Rand index(https://scikitlearn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html) to evaluate the performance of your predictor which should be better than the baseline predictor.

4.2.2 Requirements

- Download "set1.csv", "set2.csv", "set3noVID.csv" and "template_predictVessel.py" (https://drive.google.com/drive/folders/1lmThK7WWfNJOUnDyuzb5PJKjhKrPUt9F) from canvas.

- Add functions that you need and complete the `"predictor"` function with your chosen clustering algorithm in `"template_predictVessel.py"`. Your implementation should solve the problem of unknown number of cluster for `"set3.csv"`.
- To ensure the reproducibility, fixed random seed is required if your code randomly generates values. For example, for `scikit-learn`'s models, pass any integer to the `"random_state"` parameter.
- Please revise the file name to `"groupNumber_predictVessel.py"` before submission.

4.2.3 Rubric

- (10%) Well documented and written code.
- (20%) Complete implementation of assignment requirements.
- (20%) Adjusted Rand index score of your implementation for `"set3.csv"` is higher than baseline.
- 10% bonus for groups with top 10 adjusted Rand index score.