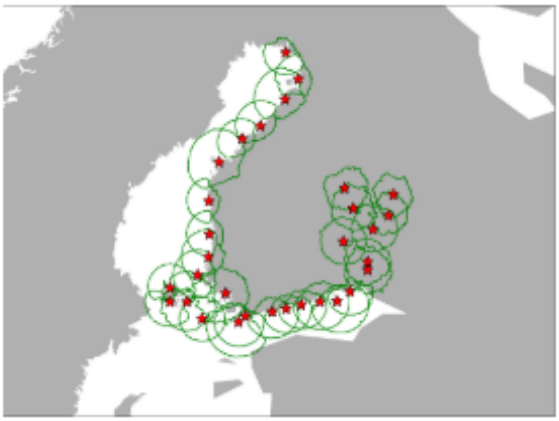
3. demo: Supervised learning

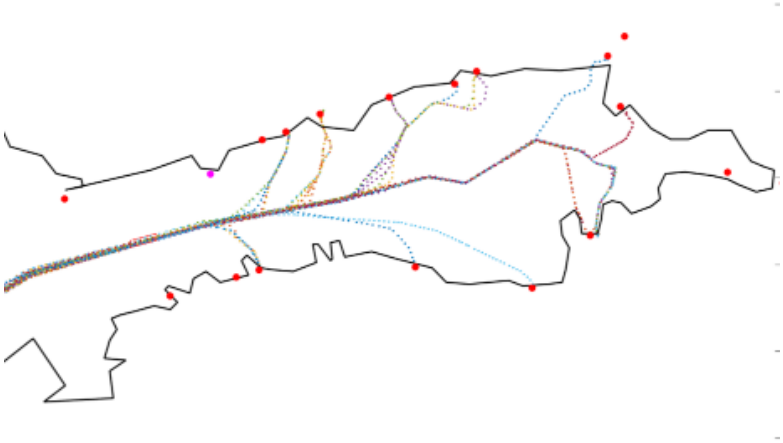
Background for the data

According to IMO SOLAS Convention, all passenger ships, tankers and other ships of 300 tons engaged in international voyages are obligated to use AIS (Automatic Identification System). It is an automatic system for sending information at specified radio frequencies on the location, speed, course, identity and other properties of a ship at frequent time intervals depending on the speed of the ship. It helps other ships and coastal states to identity ships, monitor the ships to prevent collision, and detect atypical behavior. Finnish Transport Agency (Liikennevirasto) has a comprehensive net of ground stations. The data (of ships equipped with A-class transmitters) is freely available via open interfaces.



From https://www.liikennevirasto.fi/avoindata/tietoaineistot/ais-tiedot#.W9r_6JMzZGM

The data for this demo includes a subset of properties collected from AIS data. In the data set **ship_data.txt** there are data for ships that are travelling in the Gulf of Finland. The data includes an identification for the ship (MMSI-number), the average speed over ground (SOG) in knots (through a certain area), the average course over ground (COG) in degrees, the destination harbor, ship type (cargo, tanker, tug), gross tonnage, and the length and breadth of the ship in meters.



Data Analysis and Knowledge Discovery
December 2018

Tasks

1. **Explore the data. Preprocessing.**

Open ship_data.txt. Check, how many different ship types there are, and how many ships are associated with each ship type.

Plot a scatter plot using ship length and gross tonnage, using a different color for each ship type. Are there any evident outliers? If so, delete the outliers (or alternatively you can try to find the correct value from marinetraffic.com). Do you need to use some transformation? If so, make the transformation.

Destination harbor is a categorical variable. Convert it as numerical. You can use get_dummies from pandas to implement onehot coding for categorical features.

The numerical variables have quite different ranges so it is good to make a Z-score standardization. Perform it for speed, length, breadth and gross tonnage.

2. **Predict the *ship type* with the speed, destination, length, breadth and gross tonnage data using kNN classifier with k=3. Find an estimation for the classification accuracy using random training and test sets.**

Divide the data randomly into training (70 %) and test sets (30 %). Should you use stratification? Why?

Repeat the calculation, say, 1000 times, and for each repetition, calculate the classification accuracy, i.e. how many times the classifier predicts the ship type correctly divided by the number of the ships in the test data. Plot the repeated classification accuracy values. Comment your result.

3. **Predict the *ship type* with the speed, destination, length, breadth and gross tonnage data using kNN classifier with k=3. Find an estimation of the classification accuracy using leave-one-out. Find the optimal value for k.**

Divide the data into training and test sets using leave-one-out, i.e. use each ship once as the test data and the remaining ships as the training data, and predict the ship type of the test data using the training data using kNN classifier with k=3. You will get n (=the number of ships in the data) predictions for the ship type. Calculate the classification accuracy, i.e. the how many times the ship type was predicted correctly divided by n. Compare the result with the one you got in task 2. Which method is a better evaluation of the performance of the classifier with this data set? Repeat the calculation in task 3 with values k=1…20. What is the best classification accuracy achieved?

4. Testing with training data (generally, this should NOT be used!)

Predict the *ship type* with the speed, destination, length and gross tonnage data using kNN classifier with k=1..20. Use all ships in the preprocessed data for training. Use the same training data also to test the classifier. Find an estimation for the classification accuracy. Plot the classification accuracy with different k values acquired with leave-one-out and using training data in the same figure. Comment your result. Why shouldn't you test with your training data?

A bonus task: Implement nested-cross-validation for task 3 to estimate the classification accuracy of K-nn method that selects the optimal value k=1… 20. If running leave-one-out inside leave-one-out takes too much time, you can replace either inner or outer CV loop with 10-fold CV. Does the classification accuracy achieved differ from the one obtained earlier?

If you have any questions or feedback concerning the tasks, don't hesitate to ask: pekavir@utu.fi