CSC6515 – Machine Learning for Big Data

Project
Nov. 02, 2017
Due date: Sunday, December 10, 2017 – 5:30pm

In this project, you will practice feature engineering and hierarchical classification using Python. You will also practice cross-validation as an evaluation technique and a statistical significance test. Projects are designed for individual effort. Any appearance of plagiarism will be strongly discouraged.

Deliverables:

- A report file in PDF format where you will present all your results. It is important that you also discuss all the decisions you have taken in all the steps of the project. (Maximum 10 pages.)
- Your Python source code.

Submissions:
Please upload your completed project in the Brightspace platform. The filename MUST include your last name and your banner number (e.g. A1_AmilcarSoares_B00444444.pdf).

Dataset:
Use the provided Geolife dataset. The target variable is the last column in the CSV file. Each row in the file labeled with one of the following categories:

1 bus (29%)
2 car (11%)
3 walk (36%)
4 taxi (5%)
5 subway (7%)
6 train (12%)

Other information about the dataset:

- Number of trajectory points: 4.485.796
- Number of classes: 6

Useful Python packages:

- **Numpy:** multidimensional arrays, vector and matrix operations
- **Pandas:** data manipulation and analysis
- **Scikit-learn:** machine learning library for classification, regression, clustering, feature selection and much more

Setup

(a) You don't need to code from scratch anything. What is going to be evaluated is your ability to find the answers for the questions and if your decisions were appropriate to support your answers. This means that you are free to use any packages you want in Python to support your conclusions.

Your task:

A. Feature engineering

1. Group the trajectories by user id and day and compute the following point features: (i) distance traveled (e.g. haversine, in meters); (ii) speed (m/s); (iii) acceleration(m/s2); (iv) bearing (0 to 360 degrees).
2. Create sub-trajectories by class using the daily trajectories and compute the trajectory features as follows:
    a. Discard sub-trajectories with less than 10 trajectory points.
    b. For each point feature, compute the minimum, maximum, mean, median and standard deviation. Those 20 values (5 statistical measures x 4 point features) are the trajectory features you should use for classification.
3. Explore the data and compare the trajectory features values by class. Is it possible to detect similarities or significant differences between the classes? Provide plots to support the conclusions you make.

B. Hierarchical classification

1. After evaluating the trajectory features, propose a hierarchy to classify the data based in your conclusions. Provide an overview of the groups you merged on each layer.
2. Implement the structure you propose. Remember that in the final layer, the output must be the 6 classes from the original dataset.
3. Choose two different classifiers and compare the results using your hierarchical structure with a flat structure (i.e. use the classifier with 6 classes).
    a. Perform a multiclass evaluation and a significance test (e.g. paired t-test) for each classifier. Use a ten-fold cross-validation with stratification.
    b. Report your findings and discuss your results. You are free to use any type of plot that support your conclusions.

Whenever you are asked to compare the results, you have to discuss on the results and provide acceptable reasons that justifies your results.

Also, please feel free to use any kind of plots (e.g. bars, boxplots, …) in order to visualize your results.

For questions regarding the project, contact by email amilcar.soares@dal.ca