GEO5017 A1
Clustering Urban Objects from AirBorne LiDAR Data

1 Introduction

Airborne LiDAR point clouds are vastly available for many countries[1]. Identifying semantically meaningful objects from the point clouds is key to many real-world applications, such as urban planning, autonomous driving, and robotics, which is usually known as semantic segmentation. That is, answering both 'what are the objects?' and 'where are the objects?' Point cloud segmentation can be deployed at scene level (semantic segmentation), object-level (instance segmentation), and part level (part segmentation). Point cloud segmentation is not a trivial task due to that the data is usually noisy, sparse, and unorganized. Apart from that, the sampling density of points is uneven and the surface shape can be arbitrary with no statistical distribution pattern in data. Moreover, due to limitations in 3D sensors, the background is entangled with the foreground.

In this assignment, instead of solving the semantic segmentation problem, we will focus on classification only. You will be working on grouping a set of pre-segmented individual urban objects into groups using unsupervised learning techniques (i.e., clustering).

2 The tasks

2.1 Task 1: Implement the three clustering methods

Using python, implement k-means, hierarchical, and DBSCAN clustering algorithms introduced in the lectures.

You will need to design a set (≥ 3) of features and use these features to define the 'distance' of two feature vectors. There are no standard features you can directly use. It could be helpful to visualize and observe the point clouds of different types of objects to get some intuition. It can also be helpful to select good features by looking into the distributions of potential features. Be creative! Any quantities that can separate different types of objects can be good features and are worth to look into.

Note: you must implement the algorithm by yourself and it is NOT allowed to simply call the implementation of an algorithm from any existing package or library.

2.2 Task 2: Test your implementation on the provided data

Run your implemented clustering algorithms using the given urban object dataset. To do so, you will also need to write some code to load the point clouds from files, compute the features for each point cloud, and feed the features to your clustering algorithms.

[1] Wikipedia: National LiDAR Dataset

2.3 Task 3: Evaluate and analyze your results

Come up with an idea to evaluate your results. For example, you can look into the ground truth data and measure the percentage of correctly labeled data samples (this metric is usually referred to as 'accuracy' in classification tasks). Since clustering algorithms do not assign labels to the test data, you may consider the label received by the majority of the data samples in a group as the predicted label. You can write some code for such an evaluation. Again, there is no standard solution, and be creative!

3 Test dataset

The test dataset contains 500 point clouds of urban objects. Each object is stored in a single point cloud file of the 'xyz' format, where each line has three floating-point numbers denoting the x, y, and z coordinates of a 3D point. All point cloud files have a 3-digit integer base name followed by the '.xyz' extension, e.g., 000.xyz, 231.xyz, and 499.xyz. For each point cloud, the dataset also provides its ground truth label as follows (based on the base names of the point cloud files):

- 000 - 099: building
- 100 - 199: car
- 200 - 299: fence
- 300 - 399: pole
- 400 - 499: tree

4 Submission (Due: Mar 07, Monday. 23:59)

Please compress all the following into a single archive titled **GEO5017_A1_Group_0X.zip** (where 'X' is your group ID) and submit it to BrightSpace:

- **A report ($\leq$ 3 pages excluding references)**
  - **Introduction**
    Describe the motivation and goal of this assignment. (5%)
  - **Methodology**
    - Describe your features and justify your design choice of the features. (10%)
    - Describe the distance measure between two feature vectors and explain why it is defined in that way. (10%)
    - Describe the clustering algorithms (pseudo code is recommended) (10%)
    - Describe the performance metric (i.e., how the quality of a clustering result is measured). (10%)
  - **Experiment**
    - Describe your experiment and results. (5%)
    - Discuss and analyze your results. (5%)

  - **Conclusion**
    * What conclusion did you draw from this assignment? (5%)
    * What could be done to achieve a better result?(5%)
  - **A short description of "who did what"**. (5%)

- **Source code**
  - The source code, archived in a 'code' subfolder. The code should be able to build, run, and reproduce your results without any modification. (30%)
  - [*optional*] Provide a link to the GitHub repository (only if you use GitHub) in the 'Experiment' section of your report. You are encouraged to collaborate with your teammates on GitHub.