CSC459 Database Management Systems (Spring 2024)

Programming Assignment 2: AGNES

Goals for learning

In this assignment, we will:

1. Take a close look at hierarchical cluster analysis, using the AGNES (AGglomerative NESting) algorithm.
2. Practice using the Pandas and NumPy libraries for data analysis.

Submission details

- When is it due: **Monday 04/15, 11:59 PM**
- What is provided to you:

    1. This assignment description.
    2. A "unit test" dataset (unit_test_data.csv).
    3. My results for the "unit test" dataset (agnes_unit_test_given.txt)
    4. The "full" dataset (apartments_for_rent_classified_100.csv)
    5. The template for your script (agnes.py).

- What to submit:

    1. Your implementation of the AGNES algorithm (agnes.py).
    2. Text files containing the output from running your completed agnes.py.
        - "agnes_unit_test.txt"
        - "agnes_single_link.txt"
        - "agnes_complete_link.txt"
    3. A short write-up (exported as PDF) that includes the plots generated by the agnes.py template.

- Where to submit: Upload to the associated assignment in **Brightspace**
- As specified in the syllabus, a 5% penalty will be applied per day late.

Demonstration details

During the next lab session **after the submission deadline** (tentatively Tuesday 4/16), you should be prepared to give a short demonstration of your solution and walk through your code.

About the data

- For our "full" dataset, we will be using a subset of the "Apartment for Rent Classified" dataset from the UC Irvine Machine Learning Repository.
- This data is in a ";"-separated value format, which can be read in by Pandas.
- There are many **attributes** of the apartments that we could use for clustering, but for this assignment we will be focusing on the **latitude** and **longitude** data.

Instructions

1. Download the script template (agnes.py).
2. Download the dataset files (unit_test_data.csv, apartments_for_rent_classified_100.csv)
    - By default, the script template looks in its current directory for the dataset files.
3. Modify the template with your solution:
    - See the section "About agnes.py" below for details.
4. Run the program: "python3 agnes.py"
    - Upload the modified "agnes.py" file to Brightspace.
    - Copy the output files to Brightspace.
        - "agnes_unit_test.txt"
        - "agnes_single_link.txt"
        - "agnes_complete_link.txt"
5. Generate a short write-up and export it as a PDF.
    - Include: The color-coded cluster plots that are generated by the agnes.py template.
    - Reflect on: What are the differences between the plots generated by the single-link and complete-link variants?
    - Reflect on: AGNES is a slow algorithm. What enhancements can you think of that might speed things up? (You do not need to implement your ideas, just brainstorm them).
6. **(After Submitting)** Demonstrate your code in-person:
    - During the week of 4/16, I will be making time during our lab session for in-person demonstrations.
    - Be prepared to talk through your code and your solution.

About agnes.py

- What you need to do:
    - Implement any steps marked with a "TODO" comment.
    - Do not use any libraries or imports beyond what is provided in the template.
    - Make sure that your "unit test" output matches the given output ("agnes_unit_test_given.txt").
- Implementation Hints:
    - You are not obligated to use the existing function outlines.
        - The template is provided as a suggestion. You are not required to use all provided parameters for each function.
        - Feel free to make modifications to function names and/or parameters.
    - You will need to search the Pandas library documentation, especially the DataFrame documentation, to see what functions are available to you.