CSCI316 Big Data Mining Implementation and Techniques

Laboratory 2

Objective

- To a simple kNN classifier
- To develop an end-to-end big data project in Scikit-Learn

1. Implement a simple kNN classifer

Review the kNN classifier taught in Lecture 2, and understand the maths behind. Based on the sample code in the lecture slides, implement this classifier in Python using Jupyter Notebook.

Given a testing tuple, the kNN classifier presented in the lecture note returns a hard decision (e.g., either class A or class B). Modify the kNN classifier to return the frequencies of class labels in the k-nearest neighbours in each class. For example, the output is a printing statement such as: "x% in class A and y% in class B".

2. Develop a big data project with Scikit-Learn

In many real-world big data programs, the delivery is in the form of an end-to-end big data project. To obtain some hands-on experience in implementation the big data project lifespan in Scikit-Learn, walk through and reproduce the sample code of big data project in Lecture 2.

The house dataset is available on Moodle.

Group Forming (reminder)

Besides working on the tasks of this laboratory, please also form groups for the two group-based assignments. Group size is limited to 2-3 members (not smaller than 2 not larger than 3). Members in the same group must be either all full-time students or all part-time students.