CSE455/CSE552 – Machine Learning (Spring 2018)

Homework #1

Handed out: 8:00pm Tuesday February 27, 2018.

Due: 11:55pm Tueday March 13, 2018.

Hand-in Policy: Via Moodle. No late submissions will be accepted.
Collaboration Policy: No collaboration is permitted.
Grading: This homework will be graded on the scale 100.

Description: Experiments with KNN and SVM on two well known classification data sets (IRIS - https://archive.ics.uci.edu/ml/datasets/iris and Leaf - https://archive.ics.uci.edu/ml/datasets/Leaf) data. The data is available on the class site. You can also use Python libraries to read these data files.

This project is expecting you to write four different functions to test your solutions to the problem. You are expected to use the Python language. You will prepare a Jupyter Notebook including your code and results.

- Part 1: Build a classifier based on KNN (K=5 for testing) using Euclidean distance.
  - You are expected to code the KNN classifier by yourself.
  - Report performance using an appropriate k-fold cross validation using confusion matrices on both datasets.
  - Report the run time performance of your above tests.
- Part 2: Build a classifier based on KNN (K=5 for testing) using Manhattan distance.
  - You are expected to code the KNN classifier by yourself.
  - Report performance using an appropriate k-fold cross validation using confusion matrices on both datasets.
  - Report the run time performance of your above tests.
- Part 3: Build a classifier based on linear SVM.
  - You may use an available implementation of SVM in Python.
  - Report performance using an appropriate k-fold cross validation using ROC curves and confusion matrices. Find the best threshold for the SVM output as described in the note by Fawcett.
  - Report the run time performance of your above tests.
- Part 4: Build a classifier based on polynomial SVM.
  - You may use an available implementation of SVM in Python.
  - Report performance using an appropriate k-fold cross validation using ROC curves and confusion matrices. Find the best threshold for the SVM output as described in the note by Fawcett.
  - Report the run time performance of your above tests.
- Part 5 (optional): Improve your search procedure in Part 1 and Part 2 using an advanced search algorithm such as kd-trees.

What to hand in: You are expected to hand in one of the following

- **HW1_lastname_firstname_studentnumber_code.ipynb** (the Python notebook file containing the code and report output).

Your notebook should include something like the following:

Part 1:

Code:

Results:

Comments: