CICS 397A – Fall 2019

Homework 3

Due Thursday, November 7th at 11:59pm

You are encouraged to discuss the assignment in general with your classmates, and may collaborate closely with 1-2 other students on the design and logic of your solutions. If you choose to do so, you must indicate with whom you worked. In addition, the code you submit must be entirely your own; two students submitting the same code will be considered plagiarism.

Code must be written in a reasonably current version of Python (>3.0), and be executable from a Unix command line. You are free to use Python's standard modules for data structures and utilities.

Part 1: k-Won't you be my neighbor?

For this assignment, you will implement a k-nearest neighbors classifier to be run on a dataset generated from the voting records of the United States House of Representatives. For a given representative, the classifier must predict how they voted on an arbitrary bill, based on how their "neighbor" representatives voted. In this context, neighbors are representatives with similar voting records.

The data for your classifier can be found in `congress_data.csv`. The file contains a row for each representative, with columns containing some descriptive attributes (name, state, district, party), along with all of their votes from 2018.

Some stub code can be found in `congress_knn.py`. While the supplied code will take care of the data import and creation of training and test sets, the functions that handle the details of the k-nn algorithm must be filled in (see the comments in the code for more instructions).

You will invoke the classifier code from the command line, passing in the path to the data file, the name of the vote column to predict (e.g., "Vote42"), and a value for k, the number of neighbors to consider when making predictions. For example:

```
python congress_knn.py /my/files/congress_data.csv Vote119 13
```

When completed, your code should have a prediction accuracy of 90-95% for a value k=13.

Part 2: What about your friends?

While `congress_knn.py` is set up to be run as a script from the command line, you can also use its functions analyze the data in different ways, by importing it into another script or an interactive Python session. Once you've implemented all the functionality in Part 1, use your code to answer the following:

1. How does the value of k affect the accuracy of your classifier? Provide a table or plot that shows how accuracy varies with numerous values of k. How do you explain these results?

2. Which 10 members of Congress are the most similar to Representative Jim McGovern, who represents Amherst as part of Massachusetts's 2nd Congressional District? Explain how you determined this, including your choice of k.

3. The fourth column of the has the party affiliation of each member. Using your code for determining nearest neighbors, determine the top 3 Democrats and top 3 Republicans who are the most dissimilar from others in their party. Explain how you scored them to determine the ranking.

Grading

We will run your modified `congress_knn.py` and examine the output for correctness. Your grade will be determined by how many of the exercises achieve the correct output, with partial credit being awarded.

What to Submit

You should submit:

- A modified `congress_knn.py`, containing your code
- A readme.txt, containing
  - Your name
  - The answers to the questions in Part II
  - The name of the worst movie you've ever seen
  - Anyone with whom you worked with on the assignment (see note above)
  - Notes or warnings about you code (what you got working, what is partially working, and what is broken)