CSCI 353: Assignment 1

The following policy applies to this and all subsequent assignments:

- This assignment is **due *at the beginning of class* on the assigned date,** listed in the schedule in Course Information on the class website. It is **subject to the late penalties** described in the syllabus.
- Arrange your material so it is easy for me to give you credit for each standard and section.
- If you do the above-standard material (e.g., +3 or +4 or +5), please label it clearly as such.
- Only **printed** (not handwritten) reports will be accepted. Diagrams of trees, networks, and other structures may be drawn by hand.
- **If you code to calculate or use a spreadsheet**, you must note that in your hardcopy AND send me email by the **same date and time** that the hardcopy is due. Every code file, output file, folder or spreadsheet should be named Assignment x.YourLastName.YourFirstName where x is the assignment number.
- On this, and all your work, I am looking for *thinking.* You can be wrong (and I'll disagree), but you need to try to analyze what is happening.
- **Language shapes your thinking.** That's why I actually read what you write.
- **Save a tree**: *Please* print on both sides of the paper.
- **Look before you submit**: spreadsheets with the crucial answers off the page are not helpful.
- **Programming:** you don't have to code to do this assignment but I did to get the answers. You may use a spreadsheet app is fine.
- **Precision***: 4 decimal points at a minimum.* I usually carry 8.

Standard for this assignment will be 80; that means that if you meet each of the 4 standards you receive an 80. If you exceed a standard, you earn a +5. If you fall below standard, you earn a –5. If there is no evidence that you reasonably attempted the standard, you earn a -10. For example, if you exceed 1 standard, fall below 1, and ignore 1, and you score: 80 + 1 * 5 + 1 * (–5) – 10 = 70.

Purpose: This assignment introduces you to the role of normalization and distance metrics in the *k*NN algorithm, as well as the vagaries of real-world data.

Problems

Standard 1. The following artificial dataset has 4 numeric features and a nominal class label (A, B, or C).

```
5.1,3.5,1.3,0.2,A
5.7,3.4,1.3,0.2,C
4.7,3.1,1.6,0.2,A
5.0,4.6,1.4,0.1,B
5.9,3.3,4.0,1.3,C
6.5,2.7,4.6,1.5,A
5.7,2.8,4.4,1.3,B
6.3,3.3,4.7,1.4,C
4.7,2.4,3.2,1.0,A
7.7,3.6,6.7,2.2,B
7.7,2.6,6.5,2.3,C
6.0,2.2,5.0,1.4,A
6.9,3.2,5.7,9.3,B
5.6,2.2,4.6,2.0,C
```

Answer in hardcopy: Explain **how** *k*NN would classify the new example 5.0, 2.8, 4.6, 0.7 **and why** for $k = 1$, 4, and 6. Do this once using Euclidean distance and then again using Manhattan distance. *If there is a tie, either for the i^{th} neighbor or between classes, state how you resolved the tie and why you did it that way.*

To receive any credit: Show the computation that underlies your answers. There are 3 ways to do this. Choose one of the following and submit by the same deadline:

a. Calculate by hand and submit typed (not handwritten) computations with your report.

b. Use a spreadsheet app, name your file *Assignment 1 yourLastName yourFirstName*, and email it to susan.epstein@hunter.cuny.edu with subject: *Assignment 1 yourLastName yourFirstName*.

c. Write a program in any language of your choice, submit *your code and your output* in a folder named *Assignment 1 yourLastName yourFirstName* and email it to susan.epstein@hunter.cuny.edu with subject: *Assignment 1 yourLastName yourFirstName*.
To earn a +5: Look up and then apply to this problem any one of the following metrics: Minkowski norms, Levenshtein and Hamming distances, Mahalanobis distance, Chebyshev. How do your results differ?

Standard 2. Normalize the data and then repeat the work for Standard 1. (Remember that normalization**:** is *with respect to the feature*, not the instance.) Explain what differences you see (if any).
To earn a +5: Look up and then apply to this problem any one of the following metrics: Minkowski norms, Levenshtein and Hamming distances, Mahalanobis distance, Chebyshev. How do your results differ?

Standard 3. In Weka, go to Explorer, select Preprocess, and select Open file.
• Navigate to the folder called "data" (as downloaded when you installed Weka on your computer, or see the class website for paths if you are working in the Linux lab) and select a data file NOT encountered in Chapter 1. This is your A file.
• Navigate to the folder called "BiggerData" in the Linux lab (see the class website for paths) and select a data file NOT encountered in Chapter 1. This is your B file. For each of the A and B files do all of the following in ***clear***, ***concise***, *grammatical, and properly-spelled* English:
• For each file, provide full *metadata* (data about a dataset) as follows:
Dataset name
What it is about (1 sentence is fine, e.g., "Data from US communities: socio-economic data from the '90 census, law enforcement data from the 1990 Law Enforcement Management and Admin Stats survey, and the 1995 FBI UCR.")
Repository source
Task (classification or regression or clustering or other)
instances
attributes
Attribute name, type, % missing, # distinct values, # unique values (omit if > 10 attributes)
If numeric: mean, median, max, min, standard deviation
If boolean: # yes, # no
If classification: # classes, # instances in each class
If available: provenance (who created it, who donated it), field of study
• **Look at the data.** Describe what you think is interesting / important about each file. Copy and paste followed by a one-sentence answer suggests that you didn't do much, and your grade will reflect that.
• *Use machine-learning terminology covered thus far: in class*, in the reading and on the first four sets of slides.
To earn a +5: Search online for a published paper that compares at least two machine learning methods on the same data set. Provide full bibliographic data for the paper, describe what they did, how they were compared, and which method "won."

Standard 4. You will find your C file at https://data.cityofnewyork.us/Education/Statistical-Summary-Period-Attendance-Reporting-PA/hrsu-3w2q/data

For the C file, provide full *metadata* (data about a dataset) as follows:

Dataset name

What it is about (1 sentence is fine, e.g., "Data from US communities: socio-economic data from the '90 census, law enforcement data from the 1990 Law Enforcement Management and Admin Stats survey, and the 1995 FBI UCR.")

Repository source

instances

attributes

If available: provenance (who created it, who donated it), field of study

• **Look at the data.** Describe what you think is interesting / important about it. Copy and paste followed by a one-sentence answer suggests that you didn't do much, and your grade will reflect that.

• *Use machine-learning terminology covered thus far: in class*, in the reading and on the first four sets of slides.

To earn a +5: Write a thoughtful 200-400 words about how the data in this file could be applied. Include a word count.

Extra credit for meaningful work is always available. Feel free to read ahead or in greater depth, and to explore some of what you learn. Feel free to talk / write to me about what you are thinking about doing. More work is not always better work. Using an **existing method**, one that we didn't discuss but is in the literature, is fine *if you give a citation.* Be careful though; just because it is out there doesn't make it valid.