Math 425 Fall 2025 Project 1 due: 5PM on Fri October 10

Project Directions

- Include a report on every group member's contribution.
- Submit the group's well commented code used for the project with instructions on how to compile and run.
- Make a **15** to **20** minute video presentation of your results.

The project consists of 3 problems

You are given part of the Wisconsin Diagnostic Breast Cancer (WDBC) dataset[1]. For each patient, you are given a vector **a** giving features computed from digitized images of a fine needle aspirate (FNA) of a breast mass for that patient. The features describe characteristics of the cell nuclei present in the image. The goal is to decide whether the cells are malignant or benign.

Here is a brief description of the way the features were computed. Ten real-valued quantities are computed for each cell nucleus:

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

The mean, standard error (stderr), and a measure of the largest (worst) (mean of the largest values) of each of the features were computed for each image. Thus each specimen is represented by a vector **a** with thirty entries. The domain D consists of thirty strings identifying these features, e.g. `‘‘radius (mean)"`, `‘‘ radius (stderr)"`, `‘‘radius (worst)"`, `‘‘area (mean)"`, and so on. Four files are provided containing data:

- `train.txt`: data for 300 patients
- `train_values.txt`: Indicator for malignant speciment (+1) or benign specimen (-1)

[1](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))

- `validate.txt`: data for 260 points
- `validate_values.txt`: Indicator for malignant specimen (+1) or benign specimen (-1)

Problem 1

(a) Apply k-means clustering with $k = 2$ to the training data. Then use the validation data to assess the accuracy of your clustering. You will need to come up with a scheme to determine the accuracy (i.e. a scheme to determine whether a patient in the validation set has a malignant tumor or a benign tumor based on the clustering).

(b) Embed the data in dimensions $d \in \{5, 10, 20\}$ using Gaussian matrix embedding and rerun k-means on the lower dimensional data set. What is the accuracy of the clustering for each dimension d? What is the computational time averaged over 500 independent runs?

(c) Embed the data in dimensions $d \in \{5, 10, 20\}$ using the sparse random rrojection and rerun k-means on the lower dimensional data set. What is the accuracy of the clustering for each dimension d? What is the computational time averaged over 500 independent runs?

Problem 2

- Read in the data in `train.txt` into a matrix A whose rows correspond to the data for each patient in the data set. The elements in a row correspond to the 30 features measured for a patient.
- Read in the data in `train_values.txt` into a vector $\mathbf{b}$ whose domain is the set of patients and $\mathbf{b_i}$ is 1 if the specimen of patient i is malignant and it's -1 if the specimen is benign.

(a) Use the QR algorithm to find the least-squares linear model for the data.

(b) Apply the linear model from (a) to the data set `validate.txt` and predict the malignancy of the tissues. You will have to define a classifier function

$$C(\mathbf{y}) = \begin{cases} +1 & \text{if the prediction is non-negative} \\ -1 & \text{otherwise} \end{cases}$$

(c) What is the percentage of samples that are incorrectly classified? Is it greater or smaller than the success rate on the training data?

(d) Embed the data in dimensions $d \in \{5, 10, 20\}$ using Gaussian matrix embedding and repeat the work in (a), (b) and (c) for each lower dimension d. What is the computational time averaged over 500 independent runs?

(e) Embed the data in dimensions $d \in \{5, 10, 20\}$ using sparse random projection and repeat the work in (a), (b) and (c) for each lower dimension d. What is the computational time averaged over 500 independent runs?

3. Apply k-means to the class music data `songList.xlsx` and use `Class Roster` to group the class into 8 distinct music clusters.