# Homework 7

## AMATH 352, Fall 2022

**Due on Nov 21, 2022 at midnight.**

## Directions, Reminders and Policies

**Read these instructions carefully:**

- **You are required to upload a PDF report to Canvas along with a zip of your code.**
- The report should be a maximum of 3 pages long with references included. Minimum font size 10pts and margins of at least 1inch on A4 or standard letter size paper.
- Do not include your code in the report. Simply create a zip file of your main scripts and functions, without figures or data sets included, and upload the zip file to Canvas.
- Your report should be formatted as follows:
  - Title/author: Title of report, your name and email address. This is not meant to be a separate title page.
  - Sec. 1. Introduction and overview of the problem.
  - Sec. 2. Theoretical background and description of algorithms.
  - Sec. 3. Computational Results
  - Sec. 4. Summary and Conclusions
  - References
- I suggest you use LaTeX(Overleaf is a great option) to prepare your reports. A template is provided on Canvas under the Syllabus tab. You are also welcome to use Microsoft Word or any other software that properly typesets mathematical equations.
- I encourage collaborations, however, everything that is handed in (both your report and your code) should be your work.
- Your homework will be graded based on how completely you solved it as well as neatness and little things like: did you label your graphs and include figure captions. **The homework is worth 10 points. 5 points will be given for the overall layout, correctness and neatness of the report, and 5 additional points will be for specific technical things and computational results that the TAs will look for in the report itself.**

## Problem Description

Your goal in this HW is to classify politicians based on their voting records using the 1984 congressional voting data set available at (`https://archive.ics.uci.edu/ml/datasets/congressional+voting+records`).

- NumPy users: Download the file `voting_data_train_test.npz` from Canvas and load it using `numpy.load`. Be sure to familiarize yourself with this function.
- MATLAB users: Download the file `voting_data_train_test.mat` from Canvas and load it using the `load` command.

The provided data sets contain 4 arrays:

- `X_train` is a $300 \times 16$ matrix where each row contains the voting record of a single politician over 16 bills for a total of 300 politicians. A yes vote is (+1), no vote is (-1) and abstain/missing is (0).
- `Y_train` is a vector of size 300 and contains the party affiliation/label of each politician (+1 or -1).
- `X_test` is a $135 \times 16$ matrix while `Y_test` is a vector of size 135. These are akin to their `_train` counterparts but with fewer rows.

As the names suggest, the pairs (`X_train`, `Y_train`) constitute your training data while (`X_test`, `Y_test`) will be used for validating your classifier and computing MSEs. Henceforth we write $X_{\text{train}}, Y_{\text{train}}, X_{\text{test}}, Y_{\text{test}}$ to denote these matrices and vectors.

1. Let $N = 16$ be the number of features (columns of $X_{\text{train}}$) that we wish to consider in our classifier model and train a linear function

$$f(\underline{x}) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_N x_N.$$

   Take $y(\underline{x}) = \text{sign}(f(\underline{x}))$ as your classifier.
2. Report the training and test MSEs of your classifier trained on the provided training data.
3. We now wish to train our classifier using fewer features, that is for $N < 16$. Take $N = 2, 3, 4$ and repeat steps 1 and 2 above and report the training and test MSEs of your classifiers. Provide a table reporting these values for each choice of $N$. *(Note that here we are simply using the first $N$ columns of $X_{train}$ and $X_{test}$ and ignoring the rest.)*
4. Discuss your results and observations. How does the number of features affect the quality of your classifier? Do you need all 16 columns of $X_{\text{train}}$?