Math 160 (De Loera) **Final Project**
Due date: June 11th 6pm

INSTRUCTIONS

- **This is the final project of the course. You need to submit this prior to coming to the in-class written exam. This final project has 3 parts, one about applications of linear algebra models, one about discrete and network models, and finally one about convex models.**
- **Solutions need to be submitted via CANVAS using the uploading online capabilities. Other methods of submission without prior approval will receive zero points. Submit files that contains all code (with comments), and all pdf latex docs. Dont forget to submit the data you used (specially if you reformatted things).**
- **Make sure the files run the code from simply calling them, e.g., all data files should be included. We will not download files for you or retype or copy paste your code. If it does not run straight out of the box, *you will receive zero points*.**
- **Be organized and use the notation appropriately. Show your work on every problem. Correct answers with no support work will not receive full credit.**

1. **Challenge: Automatic Classification of images via SVD decomposition**

   You will write an algorithm in MATLAB for the classification of handwritten digits. For this you can download the following files from `https://www.math.ucdavis.edu/~deloera/TEACHING/MATH160/guessdigit-project.zip`

   NOTE: Data comes in compressed form, to open you type (after downloading), unzip guessdigit-project.zip. Inside the directory that will open up you will find 5 files:

```
ima2.m        -- Code Displays an image vector in the right orientation

azip.mat        -- the matrix of training digits data

dzip.mat        -- command dzip(i) tells you the (correct) digit you have in column
                           i of the matrix azip

dtest.mat       -- tells you the (correct answer) of test digits

testzip.mat     -- the test digits data
```

Use the training set, and compute the SVD of each class matrix (classes are those matrices that represent the same digit). Use the first few (5, 10, 20) singular vectors as basis of a class and classify unknown test digits according to how well they can be represented in terms of the respective bases (use the relative residual vector in the least squares problem as a measure). Here are some specific tasks.

- Write your code to do classification, it brakes the training data in classes, computes the SVD of each class and uses that to make predictions. It takes in a test data point and makes a prediction.
- Tune the algorithm for accuracy of classification. Give a table or graph of the percentage of correctly classified digits as a function of the number of basis vectors. Graph the situation for 5, 10, 20 basis vectors. Display the results in a table (or tables).
- Check the singular values of the different classes. Is it reasonable to use different numbers of basis vectors for different classes? If so, perform a few experiments to find out if it really pays off to use fewer basis vectors in one or two of the classes (i.e., do you get different/same outcome?).
- Check if all digits are equally easy or difficult to classify. Also look at one of the difficult ones, and see that in many cases they are very badly written. What is the most difficult digit to read for the computer? Does it help to increase the number of singular vectors you used? Write comments at the very end of your code with your thoughts.

2. **Challenge: Models for predicting the quality of wines.**

In this problem we will use two types of convex-linear models to to predict wine quality (as judged by enologists) from chemical measurements. The dataset you need to use is available at `http://thibaut.horel.org/wines.csv`. In each line, the first 11 columns contain the results from various chemical tests performed on the wine, and the last column is the evaluation of how good the wine is (a score between 0 and 10).

First consider a model based on Linear programming. For wine sample i, let us denote by $y_i \in \mathbb{R}$ its score and by $\mathbf{x}_i \in \mathbb{R}^{11}$ its chemical properties. Construct a linear model to predict y_i as a function of $\mathbf{x}_i$, that is, we want to find $\mathbf{a} \in \mathbb{R}^{11}$ and $b \in \mathbb{R}$ such that:

$$y_i \simeq \mathbf{a}^\intercal \mathbf{x}_i + b.$$

The quality of the model will be evaluated using the ℓ_1 norm, *i.e.,* we want to find a solution to this optimization problem:

$$\min_{\substack{\mathbf{a}\in\mathbb{R}^{11} \\ b\in\mathbb{R}}} \frac{1}{n}\sum_{i=1}^{n} |y_i - \mathbf{a}^\intercal \mathbf{x}_i - b|.$$

a. Remember from class that the above problem is equivalent to the following linear program:

$$\begin{aligned} \min_{\substack{\mathbf{a}\in\mathbb{R}^{11} \\ b\in\mathbb{R} \\ \mathbf{z}\in\mathbb{R}^n}} \ & \frac{1}{n}\sum_{i=1}^{n} z_i \\ \text{s.t. } & z_i \geq y_i - \mathbf{a}^\intercal \mathbf{x}_i - b, 1 \leq i \leq n \\ & z_i \geq \mathbf{a}^\intercal \mathbf{x}_i + b - y_i, 1 \leq i \leq n \end{aligned}$$

Explain how to rewrite this problem in matrix form:

$$\min_{\mathbf{d}\in\mathbb{R}^{12+n}} \mathbf{c}^\intercal\mathbf{d}$$
$$\text{s.t. } A\mathbf{d} \leq \mathbf{b}$$

In particular, give the dimensions and definitions of $\mathbf{c}, A$ and $\mathbf{b}$.

b. Use an LP solver (again, we recommend using SCIP) to solve the above problem and report your code as well as the optimal value of the problem. Note that the value of the problem is exactly the average absolute error of the linear model on the dataset. Does it seem to be within an acceptable range?

In this next part, you will re-use the dataset fit a linear model to predict wine quality as a function of the chemical measurements. However we will use least-squares regression instead of ℓ_1-regression.

a. Verify that for this dataset, matrix X as defined in Problem 3 satisfies the condition of Problem 3, part c.

b. Write code to compute the optimal solution to the least-squares regression problem. Report your code, the linear model ($\mathbf{a}$ and b) and the value of function RSS for this model.

c. Now strengthen the model to find the key features of the wine that make it a good wine. What are the top 5 features for that according to a LASSO model?

3. **How to influence the opinion of people on a budget (Facebook?).**

It is in fashion to influence voters in elections. In this problem, you will consider a simplified model of influence in social networks: the social network is represented by an undirected graph $G = (V, E)$ and for a set of nodes $S \subseteq V$, we denote by $N(S)$ the set of their neighbors:

$$N(S) = \{v \in V \mid \exists u \in S,\ (u, v) \in E\}$$

The influence $I(S)$ of a set of nodes is measured by $I(S) = |N(S)|$.

Imagine you work now for facebook and you are given the dataset available at `http://thibaut.horel.org/facebook.txt`. This dataset is in fact a subgraph of the Facebook social graph. Each line in the file contains the id of two users, indicating that these two users are friends on Facebook.

a. As the designer of a marketing campaign to influence the opinion of voters, your goal is to find a subset $S \subseteq V$ of at most K nodes whose influence is maximal. Write a mathematical model to solve this problem. Can you solve the problem for the data you were given using SCIP? If not, can you write a *practical* method to give an approximation to the optimum? Extra bonus if you can prove a guarantee of approximation.

b. Using any of your models/methods from part (a) write a computer function which, given the social network described in the dataset and a budget $K \in \mathbb{N}^+$, returns an approximately optimal set of nodes S for the influence function $I(S)$. The function should return both the users to influence and the value (total amount of influence) obtained.

c. Plot the influence $I(S)$ obtained by your function as a function of the budget K. E.g., what happens when $K = 1$ and $K = |V|$?