AIM 4803 **Machine Learning for AE**
Georgia Tech Homework II Due Feb 21st 11:59pm EST

Homework Instructions

- This homework contains some theoretical problems that you will need to work out algebraically. The important thing is that you **show all steps** you took to arrive at your answer.
- Exercises 1-3 are contained in this document and Exercises 4-6 are in the `Homework1.ipynb` starter-code provided on Canvas.
- Please turn in two files on the Canvas assignment:
  1. `Homework2_FIRSTNAME_LASTNAME.pdf` - a file containing your answers to the theoretical questions.
  2. `Homework2_FIRSTNAME_LASTNAME.ipynb` - a file containing your answers to the programmatic questions.
- For the theoretical questions, the use of LaTeX(the .tex file for this document will be provided so you can use this template if you'd like) is strongly encouraged for this assignment, but not required. You have a Georgia Tech account on http://www.overleaf.com where you can edit and compile LaTeX documents. This is a very useful (and satisfying) skill to create a nice technical documents and is the standard in technical publishing. If you do not want to use LaTeX, you may use Microsoft Word or similar software, just make sure your work is submitted in `.pdf` format. Handwritten responses will not be accepted.
- You must present the answers in order and the submitted documents must be legible and well-organized.
- As always, if you collaborate with others or use online resources, this is allowed, just please outline how you used them. And, of course, directly copying the responses of any entity other than yourself is not allowed.

1 Pseudoinverses and the Normal Equations (15 Points)

In class, we have extensively covered the minimization of the following loss-function to find the optimal least-squares regression weights β:

$$\mathcal{L}(\beta) = ||X\beta - Y||_2^2 \tag{1}$$

The normal equations to find the minimum-norm β when X is not full-rank is given by:

$$\hat{\beta} = (X^\top X)^\dagger X^\top Y \tag{2}$$

where $(\cdot)^\dagger$ indicates a pseudoinverse. Show that by substituting X's SVD into the modified normal equations, the following is true:

$$\hat{\beta} = (X^\top X)^\dagger X^\top Y = X^\dagger Y \tag{3}$$

Hints:

- *For a matrix $A \in \mathbb{R}^{m\times n}$, its pseudoinverse is given by $A^\dagger = V\Sigma^{-1}U^\top$*
- *For a unitary or sub-unitary matrix, $A \in \mathbb{R}^{m\times m}$, $A^\top A = I$.*
- *For a diagonal matrix, $A \in \mathbb{R}^{m\times m}$, $A^\top = A$.*

2 Normal Equations for Ridge Regression (20 Points)

Now let's introduce a new penalty into the least-squares loss-function that also considers the 2-norm of β itself:

$$\mathcal{L}(\beta) = ||X\beta - Y||_2^2 + \lambda||\beta||_2^2 \tag{4}$$

Now our loss-function penalizes both "size" of the model error and the "size" of β. Show that the β which minimizes this new loss-function is given by:

$$\beta_\lambda = \left(X^\top X + \lambda I\right)^{-1} X^\top Y \tag{5}$$

Hints:

- *Start by expanding the loss-function (Recall that for a vector v, $||v||_2^2 = v^\top v$).*
- *Once you have the loss-function entirely in terms of matrix-vector multiplication, find its gradient with respect to β.*
- *Once you have $\nabla\mathcal{L}(\beta)$, set it equal to zero and solve for β.*

3 The Achilles Heel of Linear Regression (10 Points)

In this exercise, we will see how small changes in our X and Y matrices can have *huge* consequences on our computed linear regression weights. This can be the result of noise or corruption in our training data. Let's examine a simple 2x2 example of how regularization works. Consider the following $X\beta = Y$ system:

$$\begin{bmatrix} 1 & 100 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Our matrix X is rank-deficient which means there are infinitely many solutions to this system. The minimum-norm solution to this system is:

$$\beta^\dagger = X^\dagger Y = \begin{bmatrix} 1/1001 \\ 100/1001 \end{bmatrix} \approx \begin{bmatrix} 0.001 \\ 0.1 \end{bmatrix}$$

Now let's consider *perturbing* the matrix X by some small, positive ε.

$$\begin{bmatrix} 1 & 100 \\ 0 & \varepsilon \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1 \\ \varepsilon \end{bmatrix}$$

With the addition of this tiny ε, this system is now full-rank which means there is only one solution. **By hand**, solve for β_ε (the solution to the perturbed system). How does this β_ε compare to the original $\beta^\dagger$? What does this say about the stability of linear regression in the presence of small perturbations?