**CS 412: Introduction to Machine Learning**

# Homework 3

*Prof.: Sathya N. Ravi* Assigned: 02/04/2022, Due: 02/24/2022

Few reminders:

- Homework is due at the end of day on the designated date on Blackboard.
- No homework or project is accepted in mailbox of instructor.
- You may discuss homework with classmates and work in groups of up to **two**. However, you may not share any code, carry out the assignment together, or copy solutions from any other groups. Discussions between groups should be minimal, verbal during lecture or if appropriate, on Campuswire only. The submitted version must be worked out, written, and submitted by your group alone.
- **Important:** Each question (or subpart) should have a **lead** who has finalized the submitted solutions after discussions within group. The lead should be *explicitly* indicated in the submissions. There can be two leads for any question. The submitted solution is assumed to be verified by the other person, and so the grade is assigned equally.
- All solutions must be typeset (I recommend Latex, Markdown, or Word, please do not use nonstandard formats) with code attached in Python format whichever is appropriate. Code in Jupyter Notebook converted to Pdf for Gradescope and more experimental results including figures should be provided in the report.
- Your final submission will be made in two places: (i) zip folder with a PDF file containing solutions for each question including text, sample figures, calculation, analysis, and general writeup submitted to Blackboard; and (ii) one PDF submitted to Gradescope. **Everyone** will make these two submissions.
- Submitting someone else's work (outside of the group) as your own is academic misconduct. Such cheating and plagiarism will be dealt with in accordance with University procedures (see the page on Academic Misconduct at this link).

Answer the following questions as completely as possible. We use $\mathbb{R}^d$ to denote $d$ dimensional vector spaces over the reals $\mathbb{R}$. Similarly, $\mathbb{Z}_+ =$ Set of positive integers, $\mathbb{Z}_{\geq 0} =$ set of nonnegative integers.

1. **Derivative Calculations.** Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is a function.

   (a) Given $c \in \mathbb{R}^n$, write down the gradient function $\nabla_w f$ where $f(w) = (c^T w)^2$. Given $w$, how would you implement the function $f(w)$, and its derivative $\nabla_w f(w)$?

   (b) Given $A \in \mathbb{R}^{m\times n}$. Write down the gradient function $\nabla_w f$ where $f(w) = (\|Aw\|_2)^2$. Given $w$, how would you implement the function $f(w)$, and its derivative $\nabla_w f(w)$?

   (c) Given $A \in \mathbb{R}^{m\times n}$. Write down the gradient function $\nabla_w f$ where $f(w) = (\|Aw\|_1)^2$. Given $w$, how would you implement the function $f(w)$, and its derivative $\nabla_w f(w)$?

2. **Predicting with Optimal Parameters.** Suppose we collect data for a group of students in our CS 412 class with variables $X_1 =$ hours studied, $X_2 =$ undergrad GPA, and $Y =$ receive an A grade. We fit a logistic regression using a iterative algorithm that produces the optimal coefficients/parameters for our logistic model given by $\beta_0^* = -6, \beta_1^* = 0.05, \beta_2^* = 1$.

   (a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A grade in the class.

   (b) How many hours would the student in part (a) need to study to have a $50\%$ chance of getting an A in the class

3. **Comparing Linear vs Logistic Regression by Iterations.** For this problem, we will work with the dataset called Optical Recognition of Handwritten Digits dataset. You can download the dataset at this url: `https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits`. We will consider the binary classification setup by restricting our dataset to the two (classes of) digits 0 and 1.

   Please read and understand the data preprocessing code provided in the starter code. Importantly, note that we add a coordinate with the value of 1 to model the fixed bias. For centering mean and scaling variance, we will use off-the-shelf `scikit-learn` function called `StandardScaler`. **Please** read through the documentation and understand how it works since it will be very useful for your course project. Similarly, we will use the function `train_test_split` to create train, and test splits for our analysis.

   Recall that $w \in \mathbb{R}^d$ are unnknown parameters, $X \in \mathbb{R}^{n \times d}$ is the matrix containing features/covariates ($d$ should be equal to 65 including the coordinate that we allocated for the fixed bias term for this dataset), and $y \in \{0, 1\}^n$ is the binary vector of labels.

   (a) Fill in the code to implement `sigmoid` function. Now implement the prediction function `logistic_regression` function that outputs the probabilities/logits of samples in digit 1 class given data $X$, and $w$.

   (b) Fill in the `logistic_loss`, function that returns the logistic loss function (derived in class), and the gradient of the logistic loss function evaluated using the dataset $X, y$ at a given parameter value $w$.

   (c) Fill in the `square_loss`, function that returns the squared loss function (derived in class), and the gradient of the squared loss function (aka linear regression) evaluated using the dataset $X, y$ at a given parameter value $w$.

   (d) Fix total number of iterations $T$ to be say 1000. Fill in the code for training, plotting, and testing the model. Finally, print the test accuracy of the model. For plotting, the x-axis corresponds to iterations, and y-axis corresponds to the loss value. You will have to also to pick the step size or learning rate $\eta$, use $\eta = 0.1$.

   (e) We will now try a few combinations $T \in \{100, 200, 500, 1000\}, \eta \in \{0, 1.0.2, 0.5, 0.8\}$, so totally there are 12 combinations. Report which choice of the pair $(T, \eta)$ resulted in the best accuracy in both formulations, and comment which of the two is overall better.

**Important Note.** Follow the same instructions as Assignment 1 and submit on Gradescope.