HW 3: ECE 601 Machine Learning for Engineers

Important Notes:

(a) When a HW question asks for writing a code, you would need to include the entire code as well as the output of the program as well as any other analysis requested in the question.

(b) Don't panic about the length of the HW assignment. HW assignments are treated as opportunities for improving learning and understanding, so I might include some extra text to help you better understand the concepts or learn about a point that was not covered during the class. The actual work needed from you is indeed manageable.

(c) Combine your solutions and code in **one** zip file called `homework3_UMassUSERNAME.zip`.

1. Remember from the lecture notes that we wrote the cost function for logistic regression as

$$J(\underline{w}, b) = \frac{1}{m} \sum_{i=1}^{m} l^{(i)},$$

where

$$l^{(i)} = -\log(\sigma(\underline{w}.\underline{x}^{(i)} + b)) \qquad \text{if } \; y^{(i)} = 1$$

and

$$l^{(i)} = -\log(1 - \sigma(\underline{w}.\underline{x}^{(i)} + b)) \qquad \text{if } \; y^{(i)} = 0.$$

Also remember that

$$\underline{w}.\underline{x}^{(i)} = \sum_{k=1}^{n} w_k.x_j^{(i)}.$$

Using these definitions, show the following statements:

(a) Consider the sigmoid function

$$\sigma(u) = \frac{1}{1 + e^{-u}}.$$

Show that $\sigma(x)$ and its derivative $\sigma'(x)$ satisfy the following equations:

$$1 - \sigma(u) = e^{-u}\sigma(u),$$

$$\sigma'(u) = e^{-u}\sigma(u)^2 = \sigma(u)(1 - \sigma(u)).$$

(b) Find the gradient terms for gradient descent, that is, show that

$$\frac{dJ}{db} = \frac{1}{m}\sum_{i=1}^{m}(\sigma(\underline{w}.\underline{x}^{(i)} + b) - y^{(i)}),$$

$$\frac{dJ}{dw_j} = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)}(\sigma(\underline{w}.\underline{x}^{(i)} + b) - y^{(i)}), \quad j = 1, 2, \cdots, n$$

2. In this question, we will apply Logistic Regression (LR) to a real dataset containing the health data of a population. It includes a column at the end indicating whether a person has heart disease or not. Our goal is to predict if a person has heart disease using the other columns in the dataset with LR. Follow the provided template named `question2_template.ipynb`. It consists of three main parts:

   (a) Load and Evaluate The Data Set: In this step, the goal is to learn how to visually understand the provided dataset. Different types of plots are discussed.

   (b) LR with no categorical data: In this step, we don't map categorical data to numerical data. To apply LR, we first drop categorical information from the dataset, and next, we apply and evaluate its performance.

   (c) LR with Categorical data: In this step, we use *label encoding* to convert categorical data into numerical format. Next, we apply LR to the revised dataset and evaluate its performance.

   Instructions for each part are provided in corresponding cells. Replace the placeholder `'write your code here'` in the template with your code. The dataset for this question, named `heart.csv`, is located in the same directory as the template.

3. In this question, you will implement a Multiple Linear Regression (MLR) to a real, clean dataset. You need to use a template named `question3_template.ipynb`. Instructions for each part are provided in corresponding cells. Replace the placeholder `'write your code here'` in the template with your code. The dataset for this question, named `50_Startups.csv`, is located in the same file.

4. In this question, you will implement a MLR to a real dataset. In this case, MLR is applied to a more realistic dataset, which includes missing values and redundant information. Given the larger size of the dataset, additional preprocessing steps are necessary before applying MLR. You need to use a template named `question4_template.ipynb`. Instructions for each part are provided in corresponding cells. Replace the placeholder `'write your code here'` in the template with your code. The dataset for this question, named `Car_DS_original_v3.csv`, is located in the same file.