Homework 1: Linear models **CSCE 633-600**
Due: 11:59pm on September 21, 2023

Instructions for homework submission
a) Please write your code in the provided starter Jupyter Notebook.
b) Your submission should include a Jupyter Notebook that can be run seamlessly and performs all the required steps one after another. Any submission with a runtime error would result in lost points.
c) Make sure to comment your code and complete the required cells in the notebook.
d) Please start early :)
e) Total: 100 points

Part A - Linear Regression

Machine learning with Pokemon GO

Recent studies have found that novel mobile games can lead to increased physical activity. A notable example is Pokemon Go, a mobile game combining the Pokemon world through augmented reality with the real world requiring players to physically move around. Specifically, in the following study, researchers have found that Pokemon Go leads to increased levels of physical activity for the most engaged players!
`https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5174727/`

In this problem, our goal is to predict the combat points of each pokemon in the 2017 Pokemon Go mobile game. Each pokemon has its own unique attributes that can help predicting its combat points. These include:

1. Stamina
2. Attack value
3. Defense value
4. Capture rate
5. Flee rate
6. Spawn chance

Inside the "Homework 1" folder on CANVAS you will find the data file (named "hw1_q1_data.csv") that will be used for our experiments. The rows of these files refer to the data samples (i.e., pokemon samples), while the columns denote the name of the pokemon (column 1), its attributes (columns 2-7), and the combat point outcome (column 8). You can *ignore column 1* for the rest of this problem.

(A-i) (5 points) Data exploration: Plot 2-D scatter plots and compute the Pearson's correlation coefficient between the features and the outcome of interest. Which features are the most predictive of the number of combat points?
Note: The Pearson's correlation coefficient is a measure of linear association between two variables. It ranges between -1 and 1, with values closer to 1 indicating high degree of association between a feature and the outcome. For more details, see this link: `https://en.wikipedia.org/wiki/Pearson_correlation_coefficient`. You can use any available library to compute this metric.

(A-ii) (5 points) Data exploration: Plot 2-D scatter plots and compute the Pearson's correlation coefficient between the features themselves. Which features are the most correlated to each other?

(A-iii) (15 points) Predicting combat points: The goal of this question is to predict the combat points using the considered features. **Implement** a linear regression model using the ordinary least squares (OLS) solution. How many parameters does the model have? To test your model, randomly split the data into 5 folds and use a 5-fold cross-validation. For each fold compute the square root of the residual sum of squares error (RSS) between the actual and predicted outcome variable. Also compute the average square root of the RSS over all folds.
Hint: You will build the data matrix $\mathbf{X} \in \mathcal{R}^{N_{train} \times D}$, whose rows correspond to the training samples $\mathbf{x_1}, \ldots, \mathbf{x_{N_{train}}} \in \mathcal{R}^{D \times 1}$ and columns to the D features (including the constant 1 for the intercept): $\mathbf{X} = \begin{bmatrix} 1, \mathbf{x_1}^T \\ \vdots \\ 1, \mathbf{x_N}^T \end{bmatrix} \in \mathcal{R}^{N_{train} \times D}$. Then use the ordinary least squares solution that we learned in class: $\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.
Note: You can use libraries for matrix operations and random sampling, but please implement the linear regression algorithm, the 5-fold cross-validation process, and the RSS error computation.

(A-iv) (15 points) Based on your findings from questions (i) and (ii), use linear regression and experiment with different feature combinations. Please report your results.

Note: We would like to have an informative but non-redundant feature space, i.e., the features should be predictive of the outcome of interest but not too correlated to each other.

(A-v) (10 points) Explain the mathematical derivation for implementing and training the linear regression model with ordinary least squares (OLS) solution.

Part B - Logistic Regression

(B-i) (20 points) Data Preprocessing In this question, we will learn about data preprocessing. More specifically, we want to prepare the data so that it will be ready for being fed into a machine learning model. Generally, the data contains missing values and also has categorical (non-numerical) values. We will learn how to prepare such data. The dataset we will use in this assignment is called *Hitters.* Inside the "Homework 1" folder on CANVAS you will find the data file (named "hw1_q2_data.csv") (can also be downloaded from https://github.com/jcrouser/islr-python/blob/master/data/Hitters.csv).

1. Download and read the data. For Python, you may use *pandas* library and use *read csv* function.
2. Print the data. How does the data look like? Add a short description about the data. (You may use *head()* function in *pandas* library)
3. Return the shape of the data. Shape means the dimensions of the data. (In Python, *pandas* dataframe instances have a variable *shape*)
4. Does the data have any missing values? How many are missing? Return the number of missing values. (In *pandas*, check out *isnul()* and *isnul()*.sum())
5. Drop all the rows with any missing data. (In *pandas*, check out *dropna(). dropna()* accepts an argument *inplace*, check out what it does and when it comes in handy.)
6. Extract the features and the label from the data. Our label is *NewLeague* and all the others are considered features.
7. Data preprocessing. We want to do one-hot encoding for categorical features. To do so, we first need to separate numerical columns from nonnumerical columns. (In *pandas*, check out $.select\_dtypes(exclude = ['int64','float64'])$ and $.select\_dtypes($ $include = ['int64','float64'])$. Afterwards, use *get dummies* for transforming to categorical. Then concat both parts (*pd.concat()*).
8. Transform the output into numerical format. If you have selected the label as a *pandas* series, you can use *.replace()* function. In the label, transform *'A'* to 0 and *'N'* to 1.

(B-ii) (20 points) Models for Hitters In this question, we will apply simple classification algorithms, linear regression and logistic regression, to our preprocessed data, completing our first machine learning pipeline. This problem comes after Problem B-i so the input data should be the one you have prepared for that.

1. **Prediction**: Using 80% of the data as a training set and 20% as a testing set, please train a linear regression model and a logistic regression model.
2. Please provide the coefficients for each feature for both models. Are they the same? Are they different? Why? Please describe your observation.
3. Plot the ROC curve for both models. What are the area under the curve measurements?
4. What is the optimal decision threshold to maximize the f1 score? How did you calculate the optimal threshold?

5. **Five-fold Cross-validation**: Repeat (1) using a stratified, five-fold cross-validation.
6. Do the features change in each fold? Please explain.
7. Please provide a mean and 95% confidence interval for the AUROCs for each model.
8. Please provide a mean and 95% confidence interval for the f1 score for each model.

Hint: For plotting ROC curve, we need to use different thresholds and use each threshold to perform binary classification. To do so, we compare the output of the model to each threshold and based on the comparison result, we decide if the prediction should be 1 or 0. In the next step, we compute the false positive rate and true positive rate based on the prediction result. These two rates will give us one single point on the ROC curve (false positive rate on x-axis and true positive rate y-axis). We perform this process for multiple thresholds to get different x and y coordinates. Finally, we plot the ROC curve by connecting these points.

(B-iii) (10 points) Explain the mathematical derivation you used to implement and train your linear and logistic regression models.