Homework 1 (100 points) – Due 01/20/2025 at 11:59 pm

Logistic Regression Implementation.

In this homework, you will be implementing Logistic Regression on a binary classification task.

<u>Instructions:</u> All your responses must be saved in a jupyter notebook using Google colab and pushed to your git repository. Please DO NOT upload word documents and/or .py files directly to Submitty. Failure to follow these instructions will result in points deduction.

<u>How to Submit:</u> Please upload a .txt file with a link to your file from Git as your submission to Homework 1 gradable on Submitty.

You need to perform the following tasks for this homework:

Task 1 (20 points): Advanced Objective Function and Use Case

1. Derive the objective function for Logistic Regression using Maximum Likelihood Estimation (MLE). Do some research on the MAP technique for Logistic Regression, include your research on how this technique is different from MLE (include citations).
2. Define a machine learning problem you wish to solve using Logistic Regression. Justify why logistic regression is the best choice and compare it briefly to another linear classification model (cite your work if this other technique was not covered in class).
3. Discuss how your dataset corresponds to the variables in your equations, highlighting any assumptions in your derivation from part 1.

Task 2 (20 points): Dataset and Advanced EDA

1. Select a publicly available dataset (excluding commonly used datasets such as Titanic, Housing Prices or Iris). Provide a link to your dataset. Ensure the dataset has at least 10 features to allow for more complex analysis.
2. Perform Exploratory Data Analysis (EDA), addressing potential multicollinearity among features. Use Variance Inflation Factor (VIF) to identify highly correlated variables and demonstrate steps to handle them.
3. Visualize the dataset's feature relationships, ensuring inclusion of at least two advanced visualization techniques (e.g., pair plots with KDE, heatmaps with clustering).

Task 3 (20 points): Logistic Regression Implementation

1. Implement Logistic Regression from scratch, including the vectorized implementation of cost function and gradient descent.

2. Implement and compare the three gradient descent variants (e.g., batch gradient descent, stochastic gradient descent, and mini-batch gradient descent). Explain their convergence properties with respect to your cost function.

(Refer to the research paper discussed in class; you may add additional research too).

Task 4 (40 points): Optimization Techniques and Advanced Comparison

1. Implement or use packages to incorporate any three optimization algorithms (e.g., Momentum, RMSProp, Adam). Compare their performance with the vanilla stochastic gradient descent implementation from Task 3.
2. Define and use multiple evaluation metrics (e.g., precision, recall, F1 score) to analyze and interpret results for each algorithm.
3. Perform a hyperparameter tuning process (manual or automated using grid search/random search) for each optimization algorithm and assess its impact on performance. If you have to do some research for these techniques, please cite your sources.
4. Conclude by discussing the practical trade-offs of the algorithms, including computational complexity, interpretability, and suitability for large-scale datasets.

(For more on evaluation metrics check this link: https://www.kdnuggets.com/2020/05/model-evaluation-metrics-machine-learning.html

Research task (Not Graded): After finishing all the tasks try to think about any novel ways of optimization that you can come up with. Can you improve/update RMSProp and/or Adams? Can you add some minor adjustments to the momentum technique equation? If yes, then you should definitely try experimenting with your new technique. If it gives improved results in a particular scenario, then believe it or not you have invented something of your own and you are ready to publish! Keep thinking.

Note: *Grading will be focused on your understanding of the problem and the solution. Please make sure you explain everything you have implemented in your Jupyter Notebook. You must explain your results e.g. if the algorithm you implemented has a lower accuracy you should comment on some of the reasons behind the results.*

- *Focus on demonstrating a deeper understanding of logistic regression concepts and their applications.*
- *For Full credit, clearly explain every step and decision, providing detailed justifications in your Jupyter Notebook.*
- *Discuss any unexpected outcomes in your results and hypothesize reasons for such behaviors.*