Winter 2022 **CSC 566:** Advanced Data Mining Alexander Dekhtyar

Assignment 3: Logistic Regression and Linear Discriminant Analysis

Due date: Tuesday, February 1, end of day.

This is a **very short** assignment. The more complex followup will come next Tuesday.

Assignment

This is a fairly straightforward **individual assignment**.

I recommend that you complete this assignment in Jupyter and submit a Jupyter notebook. You can use (a) the current CS Jupyter Server (link on the web page, use your Cal Poly login credentials), (b) Google Colabs, or (c) your own local install of Jupyter (e.g., as part of Anaconda) to complete the task.

Dataset. For this assignment you will be working with one specific dataset. The dataset is a Kaggle Water Quality Dataset available at:

https://www.kaggle.com/adityakadiwal/water-potability

The dataset has nine (9) independent variables measuring a variety of chemical characteristics of water. The dependent variable is called Potability and it is a binary variable (0 = water is not potable, 1 = water is potable).

Note, the dataset contains null values. For the purposes of this assignment you can impute the values as the column means prior to classification.

Task. Implement **from scratch** two classification methods:

- **Linear Discriminant Analysis** using gradient descent. (Note: you can also implement a closed form solution, but it is not required).
- **Logistic Regression** using gradient descent.

In addition, compare your implementations with the results of running:

- `sklearn.discriminant_analysis.LinearDiscriminantAnalysis()` (the Linear Disciminant Analysis implementation in `scikitlearn`)
- `sklearn.linear_model.LogisticRegression()` (the "standard" logistic regression implementation in `scikitlearn`).

Ideally, your Logistic Regression and LDA implementations should match or come close to `sklearn` implementations (in terms of outcomes, not necessarily in terms of performance).

Evaluation. For each method, produce the final evaluation using and 80-20 test-train split. Select 80% of your data at random (but in a replicable way - i.e., all tests should be performed on the same training set) for training data, and 20% for validation/test data.

Reporting. Both LDA and Logistic Regression output linear models. For all four of your methods extract the optimal models produced, and compare the LDA models to each other, and the Logistic Regression models to each other.

In addition, report the confusion matrices for each method, and where applicable, compare them to each other.

Write a short report that includes the information about your implementation, the results of your comparative study, and final analysis. Your report shall answer two questions:

- Were you successful in replicating/approximating the `SciKitLearn` implementations?
- Which of the two linear classification methods: Linear Discriminant Analysis or Logistic Regression appears to be more accurate when predicting water potability?

Deliverables. Submit all your code (you can submit Jupyter notebooks directly) and your report. Use the following `handin` command.

```
handin dekhtyar 566-a03 <files>
```

GOOD LUCK!