Final Project

Submit your plots, descriptions, and python scripts including all of your functions.

1. (20 points) **Linear regression**

   Implement gradient descent-based linear regression in Python. Use $\Delta J = 0.00001$ as the stopping criterion.

2. (20 points) **Logistic regression**

   Implement gradient descent-based logistic regression in Python. Use $\Delta J = 0.00001$ as the stopping criterion.

3. (20 points) **PCA and linear regression**

   (1) (30 points) The data in `linear_regression_test_data.csv` contains x, y, and $y-$theoretical. Perform PCA on x and y. Plot y vs x, y-theoretical vs x, and the PC1 axis in the same plot.

   (2) (30 points) Perform linear regression on x and y with x being the independent variable and y being the dependent variable. Plot the regression line in the same plot as you obtained in (1). Compare the PC1 axis and the regression line obtained above. Are they very different or very similar? Could you explain why this happens?

4. (40 points)

   Apply PCA, kmeans clustering, hierarchical clustering, linear regression, LDA, logistic regression, and ANN in sklearn to the heart disease dataset at the UCI Machine Learning Repository and describe what you could find about the data. Use Jupyter notebook to show both your code and plots.

   The data and the description of the data can be found here: Heart Disease Dataset. Use the `processed.clevand.data` that you can download here: Download the Heart Disease Dataset.

   When you apply LDA, logistic regression, and ANN, consider the problem as a binary classification problem to distinguish presence from absence of heart disease. When you apply linear regression, look for variables that are strongly correlated. When you apply clustering, do you see certain patients group together. Describe in detail what you find.