CSEN 240 Machine Learning
Homework #1

1. (5 points) Please hand-written copy the following statement and sign.

"I am committed to being a person of integrity. I pledge, as a member of the Santa Clara University community, to abide by and uphold the standards of academic integrity contained in the Student Conduct Code."

2. [Level 1: high-level understanding] (3pt) AI vs. ML
    a. Give 1 example of AI applications that do not use machine learning. Because the same application may be implemented using machine learning as well, to avoid confusion, please describe the algorithm used in your example.
    b. Give 2 examples of AI applications that use machine learning
3. [Level 1: high-level understanding] (12pt)
    a. Which machine learning algorithm ((a) classification, (b) regression, (c) clustering, and (d) dimension reduction) is the best solution for the following problems: (one each)
        i. Segmenting customers into groups based on their purchasing behaviors.
        ii. Visualizing high-dimensional data in a lower-dimensional space.
        iii. Predicting the stock price of a company based on historical data.
        iv. Identifying fraudulent transactions in a credit card dataset.
    b. Which of the following can describe PLA? (select all that apply)
        i. It is a supervised learning algorithm.
        ii. It is an unsupervised learning algorithm.
        iii. It can be used to solve a classification problem.
        iv. The algorithm is based on the concept of iteratively updating its weights in response to errors in its prediction
    c. Which of the following problems PLA is best suited for? (select one)
        i. Predict housing prices based on location, size, and other factors.
        ii. Classify images, such as identifying the contents of a photograph
        iii. Group customers according to their purchasing behavior
        iv. Extract important features from a dataset, allowing for more accurate analysis
4. [Level 2: manual exercise] (20pt)
    - Given the following dataset of labeled points in two dimensions
        - Positive samples: {(3,3), (4,4)}
        - Negative samples: {(1,2), (2,1)}
    - If the current weight is (-1, 0.3, 0), what should be new weight if we are using PLA to update the weight?
        - Note that $g(\vec{x}) = \text{sign}(w_0 + w_1 x_1 + w_2 x_2)$
    - After the new weight, how many samples does it classify correctly?
        - Note: please just update once. No need to complete the whole epoch or the whole training in this manual exercise.

5. [Level 3: manual exercise] Computer-based exercise] (30 points)
    - Implement a PLA (using HW1-5.ipynb as the starting point)
    - Given the following dataset of labeled points in two dimensions
        - Positive samples: {(3,3), (4,4)}
        - Negative samples: {(1,2), (2,1)}
    - And, its initial weight: (-1, 0.3, 0)
    - Show the weights after each update
6. [Level 3: Computer-based exercise] (30 points)
    - (Using HW1-6.ipynb as the starting point and the PLA routine that you implemented in HW1-5)
    - Assuming (0, 1, 1) is the ground truth of the decision boundary, create 40 unique samples (20 are positive and 20 are negative).
    - First, evenly split the 40 samples into two sets: one is called training samples, and the other is called testing samples.
    - Second, train a PLA boundary using 100% of the training samples, and test the accuracy of the unseen testing samples. (Repeat 100 times for the average accuracy and standard deviation)
    - Third, train a PLA boundary using 50% of the training samples (e.g., 5 positive and 5 negative samples), and test the accuracy of the unseen testing samples. (Repeat 100 times for the average accuracy and standard deviation)
    - What do you observe? Can you explain?