CSEN 240
Homework #6

(Level 0: Team Project) (8 pt)

1. Please write a short summary (about 5 sentences) on your team project progress
    a. What have you observed from the released dataset?
    b. What have you done related to data curation?

(Level 1: high-level understanding)

2. (14 pts) Which of the following statements are true about logistic regression? (Select all that apply.)
    a. Logistic regression predicts binary or categorical outcomes with probability.
    b. Logistic regression use the logistic function (also known as sigmoid function) to output a number between 0 and 1.
    c. Sigmoid function $\sigma(-t) = 1 - \sigma(t)$
    d. Softmax regression can be used for multi-class classification problems.
    e. The output of the softmax function represents class probabilities. The predicted class is the one with the highest probability.
    f. The cross-entropy loss function is typically used in softmax regression, not the mean squared error.
    g. Softmax regression uses gradient descent to minimize the cost function.
3. (8 pts) Which of the following statements are true about k-means clustering? (Select all that apply.)
    a. K-means is a type of unsupervised learning algorithm used for clustering data points into k groups.
    b. K-means require labelled datasets
    c. The algorithm works by iteratively assigning data points to the nearest cluster center and updating the cluster centers based on the new assignments.
    d. K-means is sensitive to the initial choice of cluster centers and may converge to suboptimal solutions.

(Level 2: manual exercise)

4. (15pt) Logistic regression & gradient descent
    a. Given the following dataset of labeled points in two dimensions
        i. Positive samples: {(3,3), (4,4)}
        ii. Negative samples: {(2,1), (1,2)}
    b. Given initial weight $\mathbf{w}^{(0)} = [-4.5, 1, 1]^T$ for logistic regression
    c. Please perform a batch gradient descent to update the weight once
        i. Compute the predicted probability of each sample
            1. Note that $\sigma(-1.5) = 0.18,\ \sigma(1.5) = 0.82,\ \sigma(3.5) = 0.97$
        ii. Compute the gradient (using learning rate α = 1)
            1. Note in real practice, learning rate is often much smaller
        iii. Compute the new weight
5. (15pt) K-means

a. Given the following dataset of unlabeled points in two dimensions
    i. { (1, 2), (2, 1), (3, 3), (4, 3) }
b. Given k=2, and the initial cluster centers of (3, 3) & (4, 3)
c. Please perform the k-means algorithm to put the data points into clusters
    i. Write down the steps
    ii. Write down the final cluster centers
    iii. Write down the cluster assignments

(Level 3: computer-based exercise)

6. (20pt) Logistic regression and gradient descent (Using HW6-6.ipynb)
    a. Given the modified 4-dimensional, 2-class Iris dataset, you will train a logistic regression model using gradient descent.
        i. Read the dataset.
        ii. Train the model with batch gradient descent.
        iii. Plot the cost function as a function of epochs.
        iv. Test the model and observe the accuracy.
7. (20pt) K-means (Using HW6-7.ipynb)
    a. Given the simplified 2-dimensional, 2-class Iris dataset, please implement the K-means unsupervised learning algorithm to cluster them
        i. Read the dataset.
        ii. Train the K-means unsupervised learning algorithm
        iii. Show the cluster assignment of each sample and the two centroids in a plot.