Homework 4 - K-Means Clustering & Logistic Regression

---

1. **K-Means** We want to apply the K-means algorithm to cluster the points below in one dimension (i.e., on the x-axis)

$$2, \quad 5, \quad 10, \quad 22, \quad 26, \quad 32, \quad 42, \quad 54, \quad 66$$

into 3 clusters where the initial cluster centroids are 0, 16 and 30.

   (a) Perform 3 iterations of K-Means. Tabulate your results in a table like the one below but give justification for each entry. Round your results to 2 decimal places. Then draw the points and centroids on a line graph for each iteration indicating the clusters by circling points in each cluster.

   (b) Has the algorithm converged using the criteria that points are no longer moving from one cluster to another? Justify your answer.

	Cluster 1		Cluster 2		Cluster 3	
Iter. #	Points	Mean	Points	Mean	Points	Mean
1						
2						
3						

---

2. **Logistic Regression** The general logistic function is defined by

$$y(x) = \frac{1}{1 + e^{-m(x-b)}}$$

where $x = b$ is the point where $y(b) = 1/2$. The vertical line $x = b$ is called the cutoff.

   (a) Order the functions below from the one with the lowest cutoff value to the highest; that is, if we plotted the curves the one with the lowest cutoff would lie to the left and the highest to the right. Justify your answers.

$$y_1(x) = \frac{1}{1 + e^{-2x-1}} \qquad y_2(x) = \frac{1}{1 + e^{-2x+4}} \qquad y_3(x) = \frac{1}{1 + e^{-3(x+4)}} \qquad y_4(x) = \frac{1}{1 + e^{-3x+4}}$$

   (b) Suppose we use scikit-learn's Logistic Regression model on a set of data which uses one parameter, the weight in kilograms (kg), to determine if a child between 5 and 6 years old is overweight. In this case the denominator of the logistic function is $1 + e^{-t}$ where $t = mx + \tilde{b} = mx - mb$. The algorithm returns m and $\tilde{b}$. Based on a training set of data the algorithm predicts $m = 1.35$, $-mb = -27.7$ where 1 indicates True, the child is overweight and 0 indicates False, the child is not overweight. Based on the result of the classifier, would each of the following children be classified as overweight or not overweight? Explain your reasoning.

Child 1: 19.9 kg Child 2: 20.7kg Child 3: 23.9 kg Child 4: 20.1 kg

   (c) Suppose we use Logistic Regression on a set of data which uses two parameters, the weight in kilograms (kg) and the height in centimeters (cm), to determine if a child between 5 and 6 years old is overweight. Because there are two features in this dataset the denominator of the logistic function is $1 + e^{-t}$ where $t = m1 * x + m2 * y + \tilde{b}$. where x, y are the two features. The algorithm returns $m_1, m_2, \tilde{b}$.
   Assume that the algorithm returns the values $m_1 = 0.082, m_2 = 1.375$ and $\tilde{b} = -37.13$

      i. Write the equation of the line where the height h is on the x-axis and weight on the y-axis; that is write in the form of $w(h) = mh + b$ where m is the slope, h is the height and b is the intercept.

      ii. Based on the result of the classifier, would each of the following children be classified as overweight or not overweight?

Child 1: 98 cm, 19.9 kg Child 2: 112 cm, 20.7 kg