Question 1

• [4 points] Suppose the squared loss is used to do stochastic gradient descent for logistic regression, i.e. $C = \frac{1}{2}\sum_{i=1}^{n}(a_i - y_i)^2$ where $a_i = \frac{1}{1+e^{-wx_i-b}}$. Given the current weight $w = 0.41$ and bias $b = 0.38$, with $x_i = 0.32$, $y_i = 1$, $a_i = 0.63$ (no need to recompute this value), with learning rate $\alpha = 0.68$. What is the updated weight after the iteration? Enter a single number.

• Answer: Calculate

Question 2

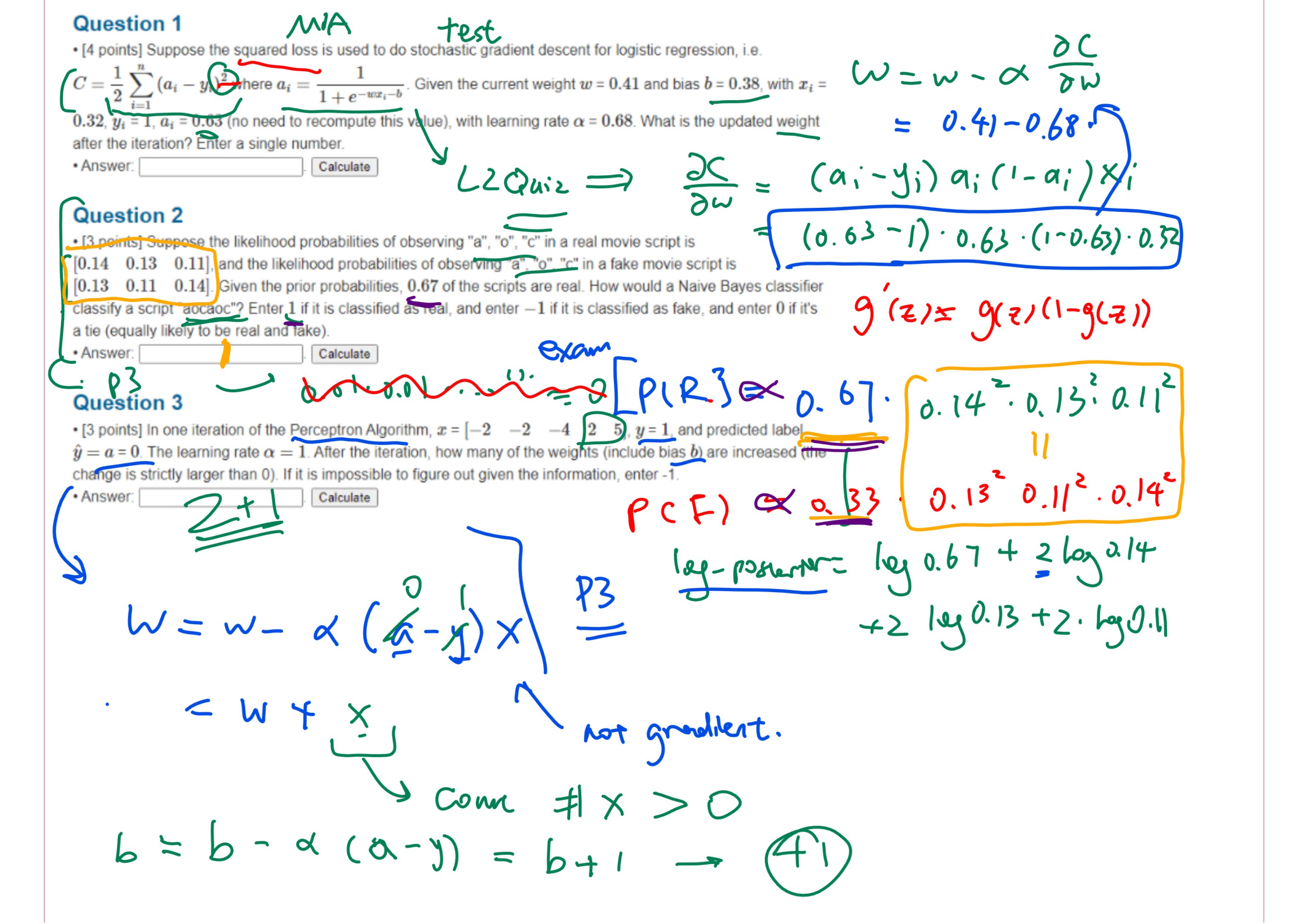
• [3 points] Suppose the likelihood probabilities of observing "a", "o", "c" in a real movie script is $[0.14 \quad 0.13 \quad 0.11]$, and the likelihood probabilities of observing "a", "o", "c" in a fake movie script is $[0.13 \quad 0.11 \quad 0.14]$. Given the prior probabilities, 0.67 of the scripts are real. How would a Naive Bayes classifier classify a script "aocaoc"? Enter 1 if it is classified as real, and enter -1 if it is classified as fake, and enter 0 if it's a tie (equally likely to be real and fake).

• Answer: Calculate

Question 3

• [3 points] In one iteration of the Perceptron Algorithm, $x = [-2 \quad -2 \quad -4 \quad 2 \quad 5]$, $y = 1$, and predicted label $\hat{y} = a = 0$. The learning rate $\alpha = 1$. After the iteration, how many of the weights (include bias b) are increased (the change is strictly larger than 0). If it is impossible to figure out given the information, enter -1.

• Answer: Calculate



Question 1

• [3 points] Suppose the vocabulary is the alphabet plus space (26 letters + 1 space character), what is the (maximum likelihood) estimated trigram probability $\hat{\mathbb{P}}\{a|x,y\}$ with Laplace smoothing (add-1 smoothing) if the sequence x, y never appeared in the training set. The training set has 500 tokens in total. Enter -1 if more information is required to estimate this probability.

• Answer: 1/27 Calculate

Question 2

• [3 points] Suppose the cumulative distribution function (CDF) of a discrete random variable $X \in \{0, 1, 2, \dots\}$ is given in the following table. What is the probability that 2 is observed.

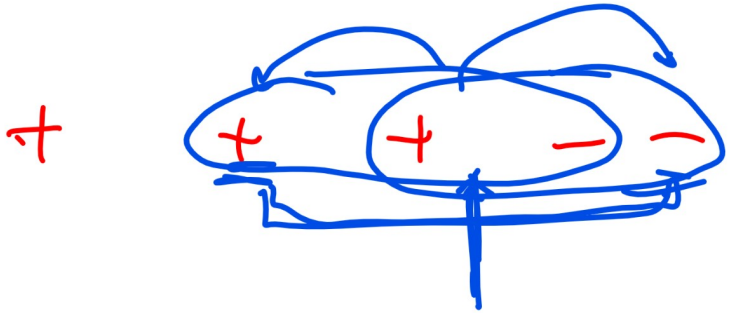
$\mathbb{P}\{X < 0\}$	$\mathbb{P}\{X \leq 0\}$	$\mathbb{P}\{X \leq 1\}$	$\mathbb{P}\{X \leq 2\}$	$\mathbb{P}\{X \leq 3\}$	$\mathbb{P}\{X \leq 4\}$
0	0.15	0.29	0.5	0.76	1

• Answer: 0.21 Calculate

Question 3

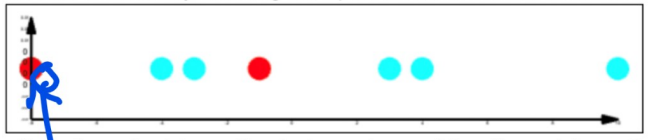
• [3 points] Given an infinite state sequence where the pattern "[2 2 1 2 3 2 3 2]" is repeated infinite number of time. What is the (maximum likelihood) estimated transition probability from state 2 to 2 (without smoothing)?

• Answer: 2/5 Calculate



Question 8

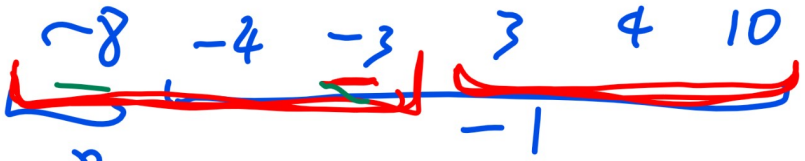
• [3 points] Perform k-means clustering on six points: $x_1 = [3]$, $x_2 = [4]$, $x_3 = [10]$, $x_4 = [-3]$, $x_5 = [-4]$, $x_6 = [-8]$. Initially the cluster centers are at $c_1 = [-8]$, $c_2 = [-1]$. Run k-means for one iteration (assign the points, update center once and reassign the points once). Break ties in distances by putting the point in the cluster with the smaller index (i.e. favor cluster 1). What is the reduction in total distortion? Use Euclidean distance and calculate the total distortion by summing the squares of the individual distances to the center.



• Note: the red points are the cluster centers and the other points are the training items.

▶ Hint

• Answer: . Calculate

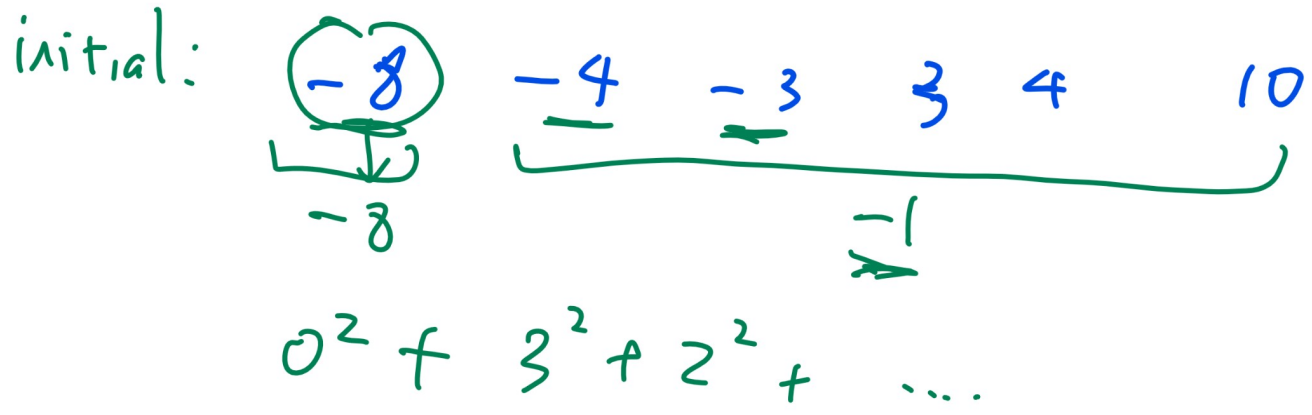


$C = -8$

$C' = -8$

initial distortion $= 0^2 + 3^2 + 2^2 + 2^2 + 3^2 + 9^2$

new distortion $= 0^2 + 4^2 + 5^2 + 1^2 + 2^2 + 8^2$

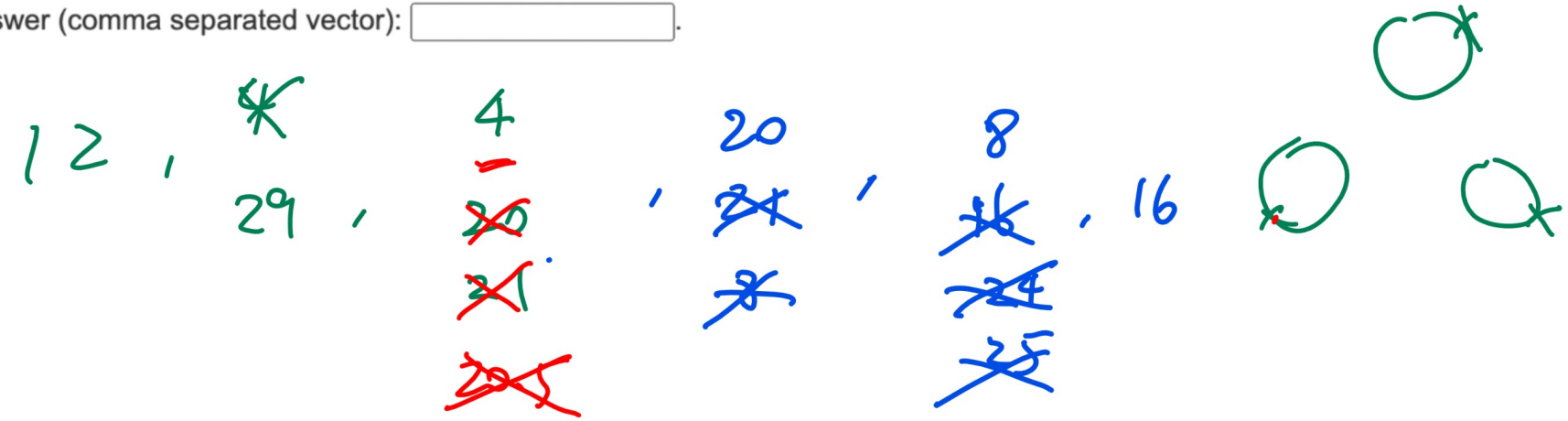


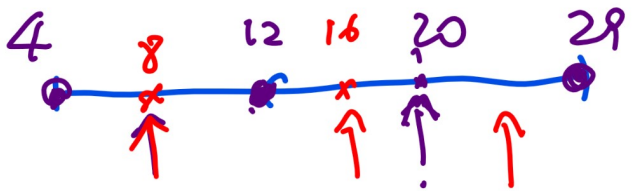
Question 9

• [3 points] Consider the 1D data set: $x_i = i$ for $i = 4$ to 29. To select good initial centers for k-means where $k = 6$, let's set $c_1 = 12$. Then select c_j from the unused points in the data set, so that it is farthest from any already-selected centers $c_1, \ldots, c_{j-1}$ (i.e. $c_j = \text{argmax}_{x_i} \min \{d(c_1, x_i), d(c_2, x_i), \ldots, d(c_{j-1}, x_i)\}$). Enter the initial centers (including c_1) in increasing order (from the smallest to the largest). In case of ties, select the smaller number.

▶ Hint

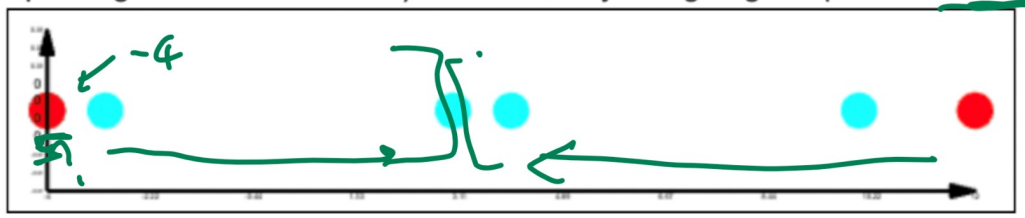
• Answer (comma separated vector): [ ].





Question 10

• [4 points] Suppose K-Means with $K = 2$ is used to cluster the data set $[-4 \quad -3 \quad 3 \quad 4 \quad 10]$ and initial cluster centers are $c_1 = 12$ and $c_2 = x$. What is the smallest value of x if cluster 1 has $n = 1$ points initially (before updating the cluster centers). Break ties by assigning the point to cluster 2.



• Note: the red points are the cluster centers and the other points are the training items.

▶ Hint

• Answer: −4 . Calculate

4 should be closer to x than 12

x [−4 −3 3 4 10]

$c_1 = x$ $c_1 = 12$

$\frac{x + 12}{2} = 4$ $x = -4$

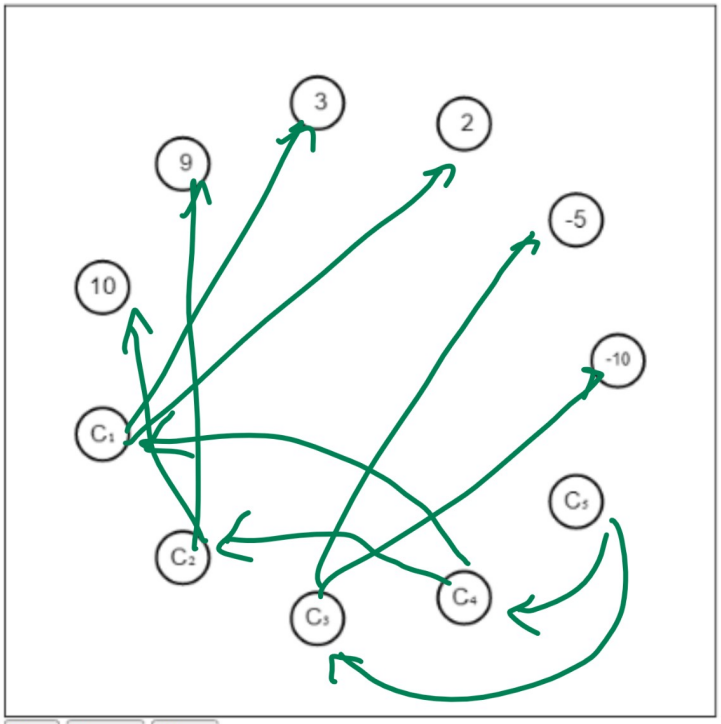
Question 4

• [4 points] Perform hierarchical clustering with single linkage in one-dimensional space on the following points: $[-10]$, $[-5]$, $[2]$, $[3]$, $[9]$, $[10]$. Break ties in distances by first combining the instances with the smallest index (appears earliest in the list). Draw the cluster tree.

• Note: the node C_1 should be the first cluster formed, C_2 should be the second and so on. All edges to point to the instances (or other clusters) that belong to the cluster.

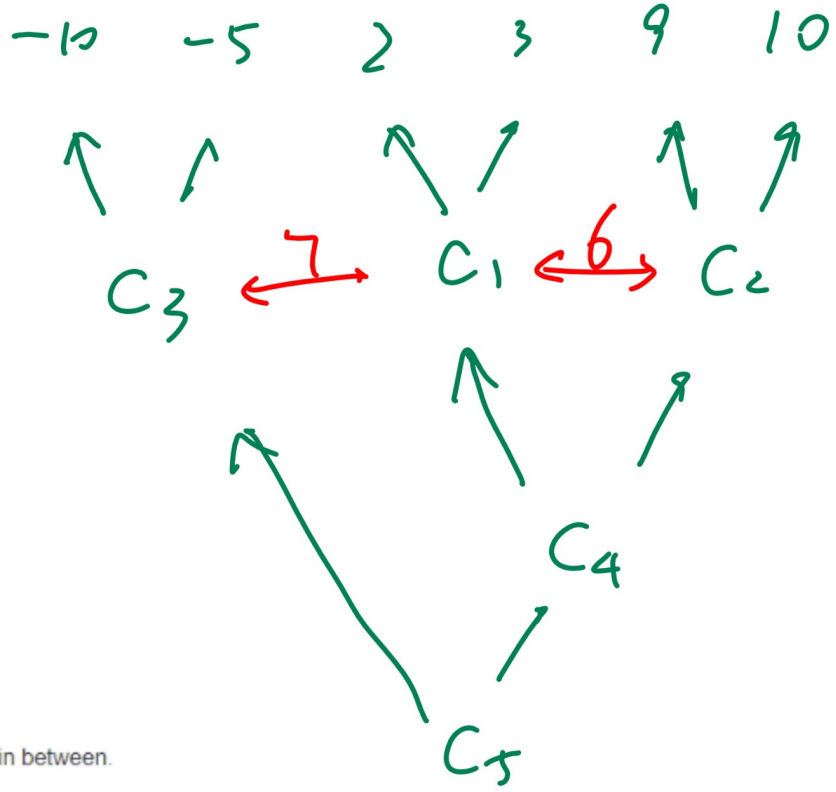
▶ Hint

• Answer:



• Note: to use the eraser, drag it from one node to another to remove the (directed) edge in between.

Single −5 (2,3) ⇒ 7
⇒ 8



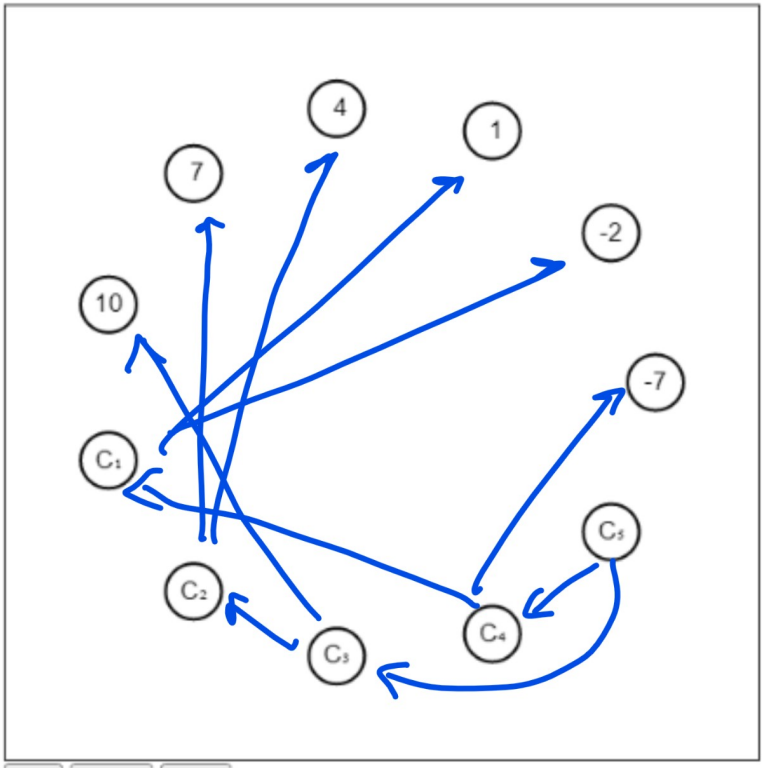
Question 5

• [4 points] Perform hierarchical clustering with complete linkage in one-dimensional space on the following points: $[-7]$, $[-2]$, $[1]$, $[4]$, $[7]$, $[10]$. Break ties in distances by first combining the instances with the smallest index (appears earliest in the list). Draw the cluster tree.

• Note: the node C_1 should be the first cluster formed, C_2 should be the second and so on. All edges to point to the instances (or other clusters) that belong to the cluster.

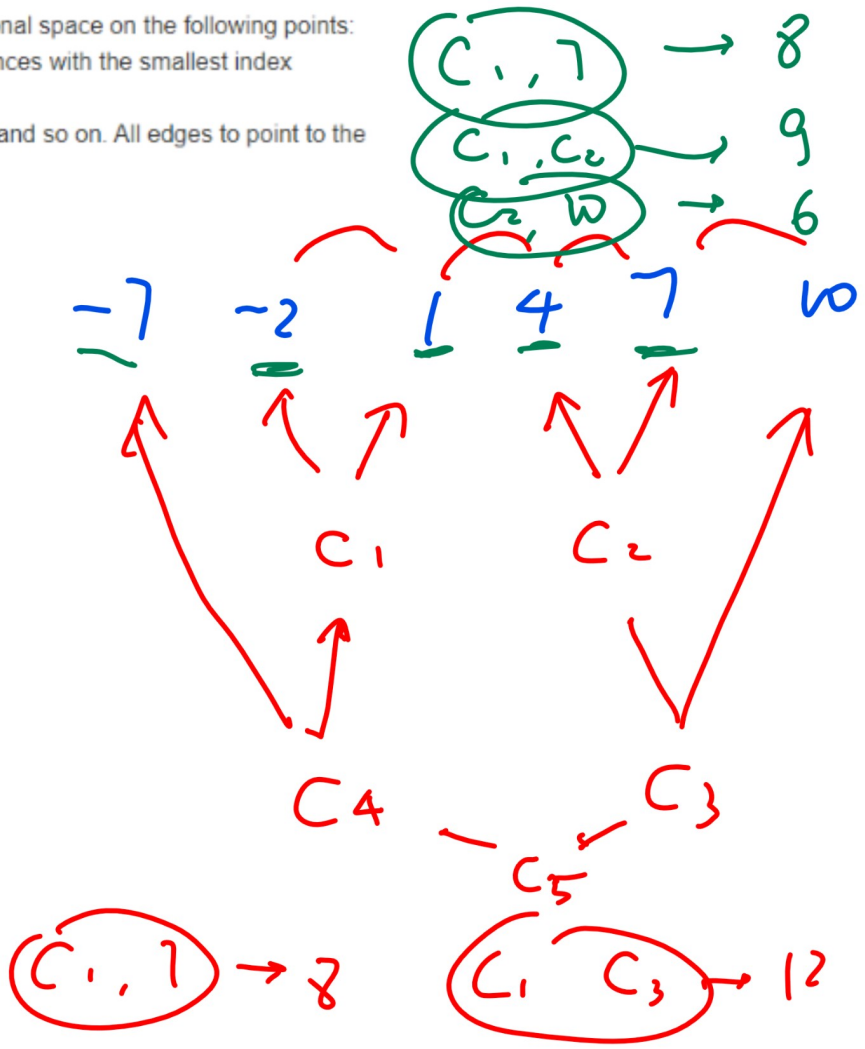
▶ Hint

• Answer:



Pen Eraser Clear

• Note: to use the eraser, drag it from one node to another to remove the (directed) edge in between.



Question 6

• [4 points] You are given the distance table. Consider the next iteration of hierarchical clustering using complete linkage. What will the new values be in the resulting distance table corresponding to the three new clusters? If you merge two columns (rows), put the new distances in the column (row) with the smaller index. For example, if you merge columns 2 and 4, the new column 2 should contain the new distances and column 4 should be removed, i.e. the columns and rows should be in the order (1), (2 and 4), (3).

$$\begin{bmatrix} 0 & 61 & 27 & 11 & 78 \\ 61 & 0 & 88 & 44 & 17 \\ 27 & 88 & 0 & 26 & 73 \\ 11 & 44 & 26 & 0 & 22 \\ 78 & 17 & 73 & 22 & 0 \end{bmatrix}$$

▶ Hint

• Answer (matrix with multiple lines, each line is a comma separated vector)