



1. In single cell RNA-Sequencing techniques, what technique is used to find from which cells the reads are coming from? [3]
2. Suppose we are using k -means to cluster gene expression data from 3 different conditions. The initial centroids are chosen randomly, leading to the following cluster centers: $C1 = [1, 1]$, $C2 = [10, 10]$, $C3 = [20, 20]$. Assume the actual data points are tightly grouped around $[5,5]$, $[15,15]$, and $[25,25]$. Will k -means converge to the correct clusters in this case? Explain how centroid initialization affects convergence and final clustering outcome. What is a probable solution to this problem? [6]

3. Soft k -means algorithm uses fuzzy or soft membership and weighted update of the centroids according to the following two formula.

membership	$w_{ik} = \frac{e^{-\beta \ x_i - \mu_k\ ^2}}{\sum_{j=1}^K e^{-\beta \ x_i - \mu_j\ ^2}}$
centroid update	$\mu_k = \frac{\sum_{i=1}^N w_{ik} x_i}{\sum_{i=1}^N w_{ik}}$

Given two cluster centers, $\mu_1 = (2, 3)$, $\mu_2 = (6, 7)$ and two data points $x_1 = (4, 5)$, $x_2 = (2, 3)$, perform the following considering $\beta = 0.5$:

- (a) Find the membership of each of these data points to the clusters. [3]
- (b) Update the cluster centers. [3]