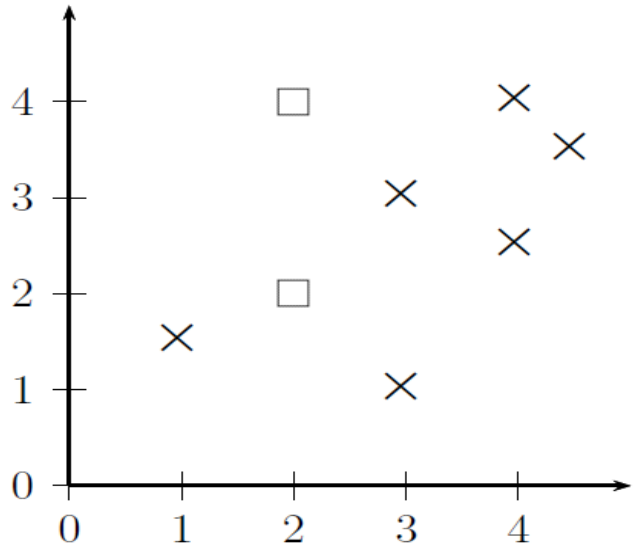
Problem Set – Clustering - Key

1. Explain how k-means differs from EM algorithm?

   K-means yield hard clustering. EM yields soft clustering. K-means algorithm can be considered as a special case of EM algorithm.

2. Staring with two cluster centers indicated by squares, perform k-means clustering on the six data points (denote by X). Stop when converged.



K-means algorithm will converge in two steps. In the first step points at the top will be assigned to the cluster center at the top and points at the bottom will be assigned to the cluster center at the bottom. In second the new cluster centers will mean of points at the top and points at the bottom.

3. Consider cluster 1D data with a mixture of 2 Guassian using the EM algorithm. You are given the ID data points x = [1 10 20]. Suppose the output of the E step is the following matrix

$$R = \begin{bmatrix} 1 & 0 \\ 0.4 & 0.6 \\ 0 & 1 \end{bmatrix}$$

where entry $r_{i,c}$ is the probability of observation x_i belonging to cluster c (the responsibility of cluster c for data point i). You have to compute the M step. You may state the equations for maximum likelihood estimates of these quantities (which you should know) without proof; you just have to apply the equations to this data set. You may leave your answer in fractional form.

   a. Write down the likelihood function you are trying to optimize.
   b. After performing M step for the missing weights π_1, π_2, what are the new values?
   c. After performing M step for the means μ_1, μ_2, what are the new values?

a. Likelihood

$$p(D|\theta) = \prod_{i=1}^{3}\sum_{k=1}^{2} \pi_k \mathcal{N}(x_i|\mu_k, \sigma_k)$$

b. Mixing Wights

$$\pi_1 = \frac{(1+0.4+0)}{3} = \frac{1.4}{3} = 0.46$$

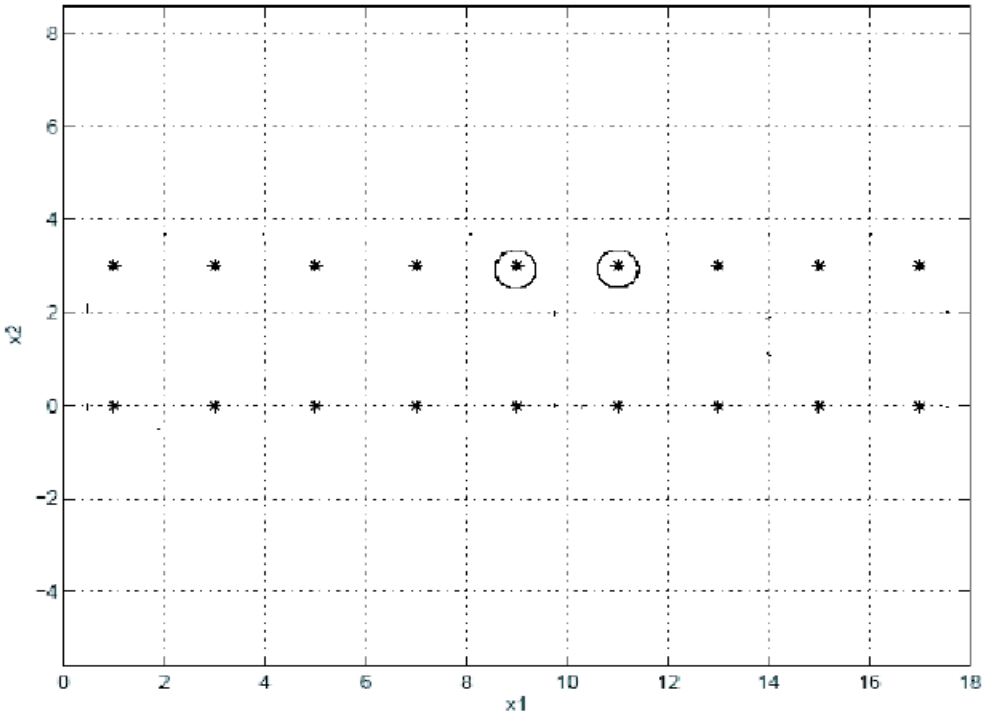
$$\pi_2 = \frac{(0+0.6+1)}{3} = \frac{1.6}{3} = 0.53$$

c. Means

$$\mu_1 = \frac{(1)1 + 0.4(10)}{1.4} = \frac{5}{1.4} = 3.57$$

$$\mu_2 = \frac{0.6(10) + 1(20)}{1.6} = \frac{26}{1.6} = 16.25$$

4. In the following figure some data points are shown which lie on integer grid. Suppose we apply the K-means algorithm to this data, using K =2 and with the centers initialized at the two circled data points. Draw the final clusters obtained after K-means converges.



Solution: K-means algorithm converges in 2 steps. The following figure shows the final means and clusters.