Theoretical task 14

Recommendations: all solutions should be short, mathematically strict (unless qualitative explanation is needed), precise with respect to the stated question and clearly written.

1. How would the steps of K-Means algorithm change, if in minimization criterion the euclidean distance is replaced by Manhattan distance (L1 distance)? What about its computational complexity?

2. Propose the generalization of K-Means algorithm for any metric space. In that case it is only possible to calculate distance between any pair of objects $\rho(x_i, x_j)$ (feature representation of the objects is forbidden).

3. Suppose that data is generated from a mixture of K multivariate normal distributions. In what cases can a) EM algorithm and b) k-means algorithm exactly recover the original partitioning (from which distribution a particular data point was generated)? What if an algorithm is initialized with the correct means from the original mixture? Explain your answer.

4. Consider a mixture of two one-dimensional Normal distributions:

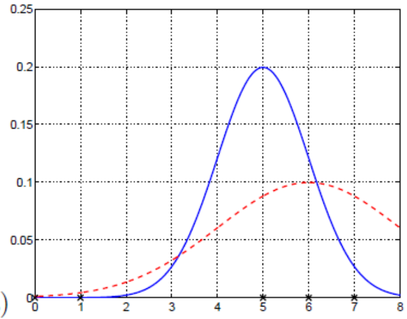
$$p(x) = w_1 N(x, \mu_1, \sigma_1^2) + w_2 N(x, \mu_2, \sigma_2^2)$$

Training dataset consists of 5 points located at 0, 1, 5, 6, 7. Below there are three plots that correspond to:

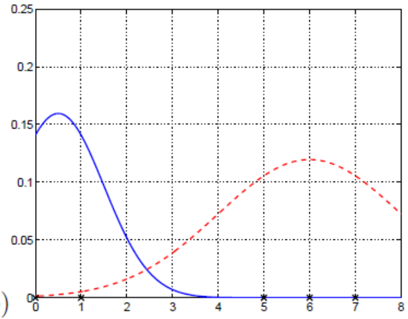
- initial situation before EM,
- situation after one iteration of EM,
- some random situation (one plot is excess).

Which plot corresponds to which situation? Explain your answer.

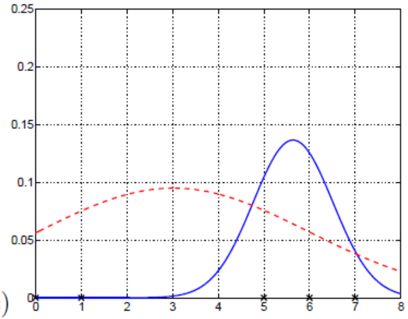
a)



b)



c)



Here blue lines represent $w_1 N(x, \mu_1, \sigma_1^2)$ and red lines represent $w_2 N(x, \mu_2, \sigma_2^2)$.