For this problem, submit your code through the Assignments page link.

Problem 1 – K-means clustering with Power Initialization

a. Implement the Power Initialization algorithm as a generic function. Inputs: sample $D$ of size $n$, consisting of real valued vectors in $d$ dimensions (it is OK to take $d = 2$), number of clusters $K$, a constant $c \geq 1$.

   Set the number of initial centers to $K' = cK \ln K$.

b. Implement the K-means algorithm proper. Inputs: sample $D$ of size $n$, consisting of real valued vectors in $d$ dimensions (it is OK to take $d = 2$), number of clusters $K$, a set of initial centers $\mu_0^1:K$, a maximum number of iterations $T$.

   The algorithm should run no more than $T$ iterations, but it should stop earlier if convergence is reached.

c. Compare Power Initialization (PI) with Naive Initialization (NI) (i.e initialization with exactly $K$ points) on the data set hw6-cluster5-data1000.dat with $K = 4$ (or $K = 3$) clusters and $T = 100$ iterations. The data file contains $n = 1000$ two dimensional real vectors, one per line.

   For PI, use the $c$ constant of your choice. For either method, plot the data as points in the plane, and superimposed on them the trajectories of the $K$ centers for the $T$ iterations. Please make as clear a plot as possible. (Separate or same plot, whatever is more readable.)

d. Also make plots showing the data and the final positions of the centers. Recommended but optional: mark the data points by their cluster assignments (e.g color the points in different colors, or mark the separation lines between clusters; the latter is OK by hand as long as it’s neat enough).

   A small matter of style – any data set with more than 100 elements looks clearer as dots than as circles/crosses/letters...

   [ e. Optional – Extra credit] Plot on a graph the cost $L(\mu_1:K) = \sum_k \sum_{i \in C_k} ||x_i - \mu_k||^2$ versus the iteration $t = 1 : T$ for the two initialization methods. [Not graded, but useful: You are encouraged to experiments by repeating the algorithm from different random initializations. Which algorithm gives a more stable clustering?]

e. Did your algorithms converge? Do you think the clusterings achieved are good clusterings of these data?

   The data set hw6-cluster3-data100-debug.dat with $K = 3$, $d = 2$, $n = 100$ is meant to help you test your code. The optimal cluster labels for this data set are in hw6-cluster3-data100-debug-labels.dat, given as the integers 1,2,3, one per line.
What you need to submit: the code through the web site; the answers and plots from c, d, e, f, on paper or pdf file.

hw6-cluster3-data100-debug.dat  K = 3  hw6-cluster5-data1000.dat  K = 3, 4

Problem 2 – Mixture models (after K. Murphy)
Consider the Gaussian mixture model

\[ f(x) = \sum_{k=1}^{K} \pi_k f_k(x) \quad \text{with} \quad \pi_k \geq 0, \sum_{k=1}^{K} \pi_k = 1 \] (1)

where \( x \in \mathbb{R} \) and \( f_k = \text{Normal}(\mu_k, \sigma_k^2) \). Define the log-likelihood as

\[ l(\mu_1:K, \sigma_1:k, \pi_1:K) = \frac{1}{N} \sum_{i=1}^{N} \log f(x_i) \] (2)

and let \( \gamma_{ki} \) be defined as in the “Lecture 9” slides.

a. Show that the gradient of \( l \) w.r.t. \( \mu_k \) is

\[ \frac{\partial l}{\partial \mu_k} = \frac{1}{N} \sum_i \gamma_{ki} / \sigma_k^2 (x_i - \mu_k) \] (3)

b. Derive the gradient w.r.t. \( \pi_k \). For now, ignore any constraints on \( \pi_k \).

There is a simple, elegant and probabilistically meaningful form of the answer, and this is the one you need to find. Feel free to introduce extra notation if you deem it necessary. (Same holds for question c. below).

c. - Optional, extra credit] One way to enforce the constraint \( \sum_k \pi_k = 1 \) is to use reparametrize \( \pi_1:K \) via the softmax function

\[ \pi_k = \frac{e^{w_k}}{\sum_{k'=1}^{K} e^{w_{k'}}} \] (4)

Find the expression of \( \frac{\partial l}{\partial \mu_k} \).

[Problem 3 – EM for Mixture of Gaussians – Not graded]

a. Assume you observe 3 samples, \( D_{1,2,3} \) of sizes \( n_1, n_2, n_3 \) respectively, where each \( D_k \) is sampled from an unknown \( \text{Normal}(\mu_k, \sigma^2) \), with \( k = 1, 2, 3 \) (three different means, and the same variance \( \sigma^2 \)).
Write the formula for the (log-)likelihood of the data \( D = D_1 \cup D_2 \cup D_3 \) as a function of the parameters \( \mu_{1,3,3}, \sigma^2 \).

b. Prove, by taking the derivative of the log-likelihood above w.r.t. \( \mu_k \) (and by analogy with the ML estimation for the \( \text{Normal}(\mu, \sigma^2) \) distribution), that the ML estimates for the means \( \mu_{1,2,k} \) are equal to

\[
\mu_{k}^{\text{ML}} = \frac{1}{n_k} \sum_{i \in D_k} x_i
\]  

(5)

c. Prove, by taking the derivative of the log-likelihood above w.r.t. \( \sigma^2 \) (and by analogy with the ML estimation for the \( \text{Normal}(\mu, \sigma^2) \) distribution), that the ML estimate for the means \( \sigma^2 \) is equal to

\[
(\sigma^2)^{\text{ML}} = \frac{1}{n_1 + n_2 + n_3} \sum_{k=1}^{3} \sum_{i \in D_k} (x_i - \mu_k^{\text{ML}})^2
\]  

(6)

d. Assume the data \( D = \{x_1, \ldots, x_n\}, x_i \in \mathbb{R} \) come from a mixture of \( K = 3 \) Normal distributions \( f_k = \text{Normal}(\mu_k, \sigma^2), k = 1, 2, 3, \) i.e.

\[
f(x) = \sum_{k=1}^{3} \pi_k f_k(x).
\]  

(7)

Derive the expression of \( \gamma_{ki} \) calculated in the E step of the EM algorithm.

e. Now derive the expression of the M step of the EM algorithm by analogy with the results you obtained in b, c. OK to “transcribe” the results from above with no proof except a one sentence motivation.

Problem 4 – K-means in pictures

In the picture below, the data are uniformly distributed in the squares shown, and the centers \( \mu_{1,2,3,4} \) are represented by colored circles. Draw the location of the centers after 1 iteration of K-means. The exact location of the center is not important; it is sufficient to mark it approximately, i.e. in the square where the center will be, in the right relative position w.r.t other centers in the same square. If a center could be either of two adjacent squares and it can’t be determined in which, place it on the boundary line of the two squares where it could be.
Example

**Initial**

**After 1 step K-means**

a.

**Initial**

**After 1 step K-means**

b.

**Initial**

**After 1 step K-means**