STAT 391
Homework 7
Out Friday May 27, 2022
Due Tuesday June 7, 2022

©Marina Meilă
mmp@cs.washington.edu

Problem 1 – K-means clustering with K-logK Initialization

a. Implement the K-logK Initialization algorithm as a generic function. Inputs: sample $\mathcal{D}$ 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 $\mathcal{D}$ 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 K-logK Initialization (KlKI) with Naive Initialization (NI) (i.e initialization with exactly K points) on the data set `hw7-cluster5-data1000.dat` with $K = 4$ clusters and $T = 100$ iterations. The data file contains $n = 1000$ two dimensional real vectors, one per line.

For KlKI, use the $c = 2$. 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 matter of clarity – any data set with more than 100 elements should be plotted as dots not as circles/crosses/letters...

e. Plot on a graph the cost $\mathcal{L}(\mu_{1:K}) = \sum_k \sum_{i \in C_k} ||x_i - \mu_k||^2$ 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?]

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

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

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

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

$$l(\mu_{1:K}, \sigma_{1:k}, \pi_{1:K}) \;=\; \frac{1}{N} \sum_{i=1}^{N} \log f(x^i) \tag{2}$$

and let γ_{ki} be defined as in the "Lecture 9" slides.

a. Show that the gradient of l w.r.t. μ_k is

$$\frac{\partial l}{\partial \mu_k} \;=\; \frac{1}{N} \sum_i \gamma_{ki}/\sigma_k^2 (x^i - \mu_k) \tag{3}$$

b. Derive the gradient w.r.t π_k. For now, ignore any constraints on π_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 y ou 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

$$\pi_k \;=\; \frac{e^{w_k}}{\sum_{k'=1}^{K} e^{w_{k'}}} \tag{4}$$

Find the expression of $\frac{\partial l}{\partial w_k}$.

Problem 3 – EM for Mixture of Gaussians – Not graded

a. Assume you observe 3 samples, $\mathcal{D}_{1,2,3}$ of sizes n_1, n_2, n_3 respectively, where each $\mathcal{D}_k$ is sampled from an unknown $Normal(\mu_k, \sigma^2)$, with $k = 1, 2, 3$ (three different means, and the same variance σ^2.

Write the formula for the (log-)likelihood of the data $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3$ 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. μ_k (and by analogy with the ML estimation for the $Normal(\mu, \sigma^2)$ distribution), that the ML estimates for the means $\mu_{1,2,k}$ are equal to

$$\mu_k^{ML} \;=\; \frac{1}{n_k} \sum_{i \in \mathcal{D}_k} x_i \tag{5}$$

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

$$(\sigma^2)^{ML} \;=\; \frac{1}{n_1 + n_2 + n_3} \sum_{k=1}^{3} \sum_{i \in \mathcal{D}_k} (x_i - \mu_k^{ML})^2 \tag{6}$$

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

$$f(x) \;=\; \sum_{k=1}^{3} \pi_k f_k(x). \tag{7}$$

Derive the expression of γ_{ki} calculated in the E step of the EM algorithm.

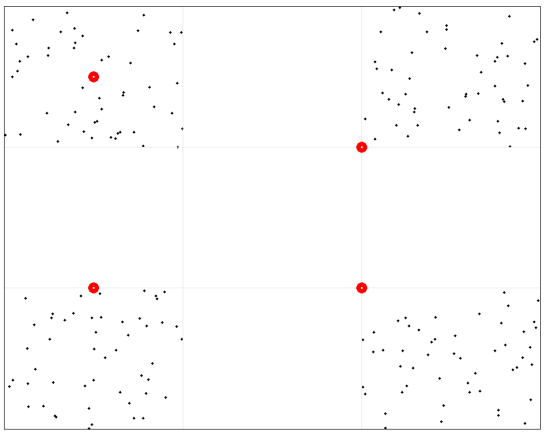
e. Now derive the expression of the M step of the EM algorithm by analogy with the results you obtained in **b, c**. *OK to "transcribe" the results from above with no proof except a one sentence motivation.*

Problem 4 – K-means in pictures

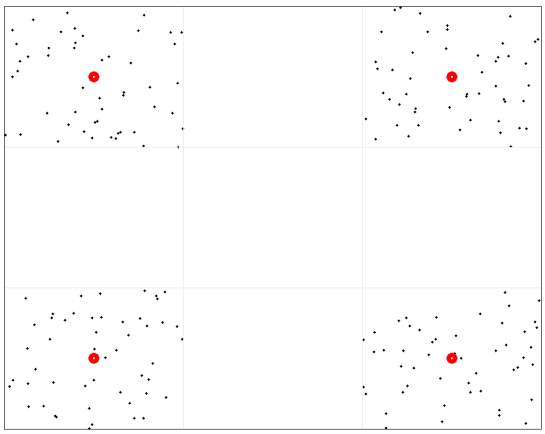
In the picture below, the data are uniformly distributed in the squares shown, and the centers $\mu_{1,2,3,4}$ are represented by colored circles. Draw the location of the centers after 1 iteration of K-means. The exact location of the center is not important; it is sufficient to mark it approximately, 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. Hint for determining the assignments to centers: for each pair of centers $\mu_k, \mu_{k'}$, first find the middle of line segment $[\mu_k, \mu_{k'}]$, then draw a perpendicular on this segment at the middle. This perpendicular separates the points closer to μ_k from the points closer to $\mu_{k'}$.

Example

Initial

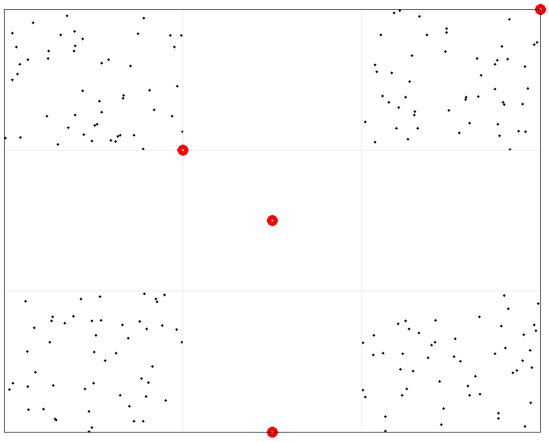


After 1 iteration K-means

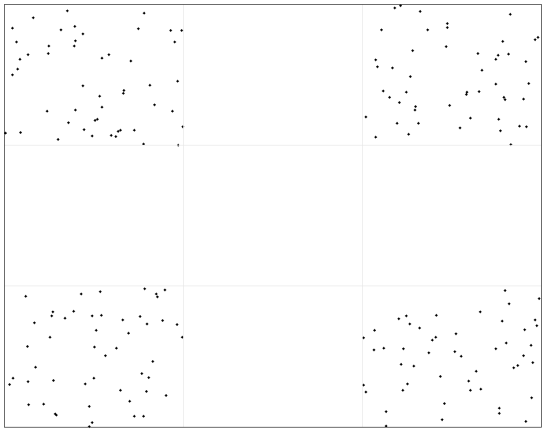


a.

Initial

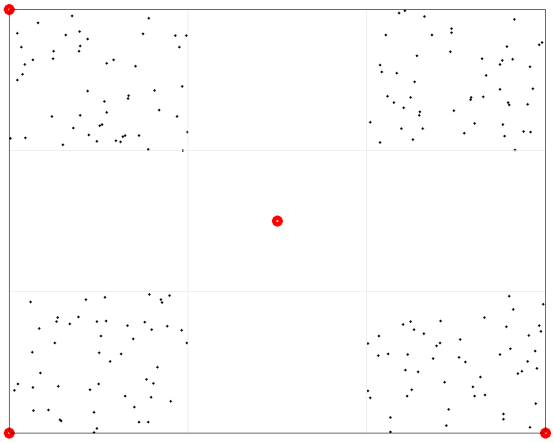


After 1 iteration K-means



b.

Initial



After 1 iteration K-means