Due February 23, 2022 at 11:59PM.

Problem 1

K-means convergence

[10 points]

In k-means, we minimize the loss function:

$$L(\boldsymbol{\mu}, \boldsymbol{\alpha}) = \sum_{k=1}^{K} \sum_{i=1}^{n} ||x_i - \mu_k||_2^2 \mathbb{1}\{\alpha_i = k\},$$

where,

- x_i is a data point with p dimensions,
- n is the total number of data points,
- μ_k is the center of cluster k and is of dimension p,
- $\mathbb{1}\{\cdot\}$ is an indicator function. That is, it is equal to one if the test is true, zero otherwise,
- α_i is the label of the i-th data point.

(a) Let $\alpha_i^{(t)}$ be the assignment in iteration (t). Show that

$$L(\boldsymbol{\mu}, \boldsymbol{\alpha}^{(t+1)}) \leq L(\boldsymbol{\mu}, \boldsymbol{\alpha}^{(t)}).$$

(b) After the assignment, k-means will do a refitting of $\boldsymbol{\mu}$ conditional on the latest assignments. Show that the update

$$\mu_k = \frac{1}{\sum_i^n \mathbb{1}\{\alpha_i = k\}} \sum_{i=1}^{n} x_i \mathbb{1}\{\alpha_i = k\}$$

is the best you can do given this loss function. Hint: one way is to use those weird derivative things.

Problem 2

[10 points]

Soft k-means updates

Please refer to the notation in Chapter 8, section *Soft k-means Clustering*. Consider the following data in two dimensions:

data ID	x_{i1}	x_{i2}
1	0.1	0.2
2	0.2	0.1
3	0.3	0
4	1	1.2
5	0.8	1
6	9	0.1

and the following centers:

cluster ID	μ_{i1}	μ_{i2}
1	0.1	0.9
2	0.5	0
3	0.9	0.5

(a) Using the **partition function** and $\beta = 0.5$ compute the E-step and show the Hidden-Matrix.

(b) Using the assignments you made in (a), compute the M-step.

(c) Using the Newtonian inverse-square law of gravitation, compute the E-step.

(d) Using the assignments you made in (c), compute the M-step.

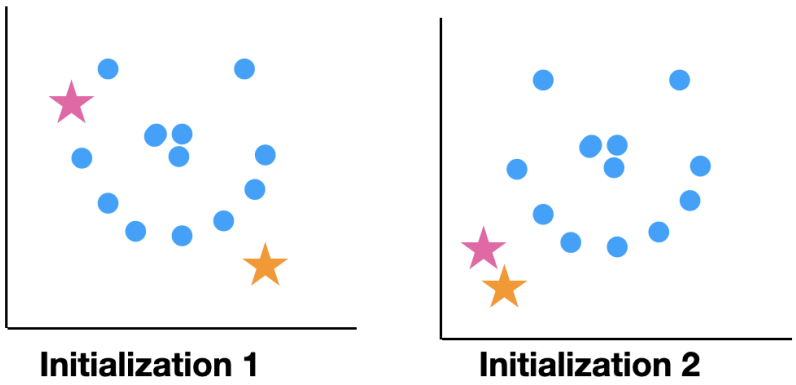
(e) Any observations comparing the two different distance functions?

Problem 3

[10 points]

Decision boundaries in standard k-means

Consider the data in the figure below:



The blue dots are the data points, and the stars are the cluster centers.

(a) How are the data points clusters in Initialization 1 and Initialization 2?

(b) Is there a conceptual difference?

(c) If you were to run the Lloyd algorithm with both initializations, how would it behave?

(d) Do you see a picture in the data?

(e) Draw some clustered data in two dimensions that is trivially easy to cluster by eye, but impossible to cluster correctly using k-means.