Assignment

Homework 03

Andrew Park

Applied Machine Learning Homework Assignment

CORNELL TECH | HOME OF THE JACOBS TECHNION-CORNELL INSTITUTE

October 3, 2023

1. General Summary

The main task of the assignment was to understand the EM algorithm and apply it to a dataset called 'Old Faithful Geyser Data,' while comparing the clustered results with K-means. Overall, the assignment went well, considering that the homework required a fundamental understanding of the parameters and what happens between the E-steps and M-steps in coding exercises (b) and (c).

To be specific about the submitted assignment, I plotted the trajectory of the iterating μ, the mean of the Gaussian component in the mixture model, to observe how the optimization takes place, and I think the results seem reasonable. When comparing the results of the EM algorithm and K-means, data points for the central area and a point at approximately (2.4, 70) appeared different. As I learned from the lecture, this seems to be a limitation of K-means, considering that instinctive analysis suggests that the clustering by the EM algorithm is much more natural.

Regarding the written exercises, it was a good opportunity to briefly learn and recall about the advantages and disadvantages of K-means and GMMs, as well as the overall concept of unsupervised learning models. Also, with the second question, I understood that there are some moments when the weighted Euclidean distance can be the same as the unweighted Euclidean distance.

WRITTEN EXERCISES

Problem 1

True/false questions. Please provide responses to the following statements, indicating whether they are true or false, and briefly provide explanations for your assessment.

Solution.

a. Training and testing split is beneficial in the context of the K-means algorithm.

   False - Training and testing splits aren't typically used in K-means clustering since it's an unsupervised learning technique that doesn't require labeled data.

b. There exist scenarios where the K-means algorithm will not converge and thus the algorithm will not terminate.

   False - K-means will always converge but it might settle at a local minimum depending on the initial cluster centers, not necessarily the global optimum.

c. There exist scenarios where you would obtain different cluster assignments by running the same K-means algorithm multiple times, potentially with different cluster initialization.

   True - Running K-means multiple times, especially with different initial cluster center placements, can lead to varied cluster assignments because the algorithm might converge to different local optima.

d. Training and testing split is beneficial in the context of GMMs for clustering.

   False - GMMs are also unsupervised, so the concept of a training and testing split isn't typically applied.

e. GMM for clustering will yield consistent cluster partitions across repeated runs.

   False - The results of GMM can vary between runs with different initial parameters as they can converge to different local optima. Multiple runs with varied initializations are often done to obtain more reliable clustering results.

Problem 2

Weights for clustering. In clustering algorithms like K-means, we need to compute distances in the feature space. Sometimes people use weights to value some feature more than others. Show that weighted Euclidean distance for p dimensional data points x_i and $x_{i'}$

$$d_e^{(w)}(x_i, x_{i'}) = \frac{\sum_{l=1}^{p} w_1 (x_{il} - x_{i'l})^2}{\sum_{l=1}^{p} w_l}$$

safisfies

$$d_e^{(w)}(x_i, x_{i'}) = d_e(z_i, z_{i'}) = \sum_{l=1}^{p} (z_{il} - z_{i'l})^2,$$

where

$$z_{il} = x_{il} \cdot \left(\frac{w_l}{\sum_{l=1}^{p} w_l}\right)^{1/2} .$$

Thus weighted Euclidean distance based on x is equivalent to unweighted Euclidean distance based on a proper transformed data z.

Solution.

a) Starting from the definition of z_{il}, we have

$(z_{il} - z_{i'l})^2$

$= (x_{il} \cdot (\frac{w_l}{\sum_{l=1}^{p} w_l})^{1/2} - x_{i'l} \cdot (\frac{w_l}{\sum_{l=1}^{p} w_l})^{1/2})^2$

$= (\frac{{w_l}^{1/2}(x_{il} - x_{i'l})}{(\sum_{l=1}^{p} w_l)^{1/2}})^2$

$= \frac{w_l (x_{il} - x_{i'l})^2}{\sum_{l=1}^{p} w_l}$

b) According to $d_e(z_i, z_{i'})$,

$d_e(z_i, z_{i'})$

$= \sum_{l=1}^{p} (z_{il} - z_{i'l})^2$

$= \sum_{l=1}^{p} \frac{w_l (x_{il} - x_{i'l})^2}{\sum_{l=1}^{p} w_l}$ which is the definition of $d_e^{(w)}(x_i, x_{i'})$

$\therefore d_e^{(w)}(x_i, x_{i'}) = d_e(z_i, z_{i'})$

So, under given a proper transformed data z, the weighted Euclidean distance based on x is equivalent to unweighted Euclidean distance.