6 Homework 6 (Clustering Algorithms)

Due date: April 2

1. Consider the following variant of the fuzzy c-means clustering algorithm minimizing the objective function
$$J = \sum_{i=1}^{c} \sum_{k=1}^{n} \ln u_{ik} \left\| x_k - v_i \right\|^2,$$
where $A_i(x_k) = u_{ik}$ is the degree of membership of x_k in cluster A_i, $x_k, k = 1, ..., n$ are datapoints, and $v_i, i = 1, ..., c$ are cluster centers with the constraint, $\sum_{i=1}^{c} A_i(x_k) = 1$, for each $k = 1, .., n$.
a) Write the Lagrangean L for this constraint optiomization problem.
b) Calculate $\frac{\partial L}{\partial u_{ik}}$ and deduce the update step for this version of the algorithm from $\frac{\partial L}{\partial u_{ik}} = 0$.
c) Calculate $\nabla_v L$ and deduce the update step for this version of the algorithm from $\nabla_v L = 0$.

2. Consider the objective function for the EM algorithm
$$J = \sum_{i=1}^{m} \sum_{j=1}^{k} w_{ij} \left[-\frac{1}{2\sigma_j} \left\| x_i - \mu_j \right\|^2 + \ln \varphi_j - \ln w_{ij} \right],$$
where w_{ij} is the membership of x_i in cluster j. and the constraints $\sum_{j=1}^{k} \varphi_j = 1$ and $\sum_{j=1}^{k} w_{ij} = 1$
a) Write the Lagrangean L of the optimization problem.
b) Calculate $\nabla_\varphi L$, and deduce the update step for φ_j from the condition $\nabla_\varphi L = 0$.

3. Write a program that predicts the time interval to the next eruption time for the old faithful geyser
http://www.stat.cmu.edu/~larry/all-of-statistics/=data/faithful.dat
The steps of the algorithm are as follows:
- Use a one dimensional EM model to cluster the eruption times for the given dataset.
- Within each cluster, perform a linear regression to predict the time interval to the next eruption within each cluster.
- Predict the time to the next eruption based on the input eruption time.
Use 250 datapoints for training and print the results of your algorithm for the remaining 22 datapoints.