Exercises
Learning and Intelligent Systems
SS 2017

Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Gunnar Rätch and Prof. Dr. Thomas Hofmann
Web: `http://las.inf.ethz.ch/teaching/lis-s17/`

Email questions to: Baharan Mirzasoleiman
`baharanm@inf.ethz.ch`

Series 4, May 1th, 2017
(Clustering and K-means)

It is not mandatory to submit solutions and sample solutions will be published in two weeks. If you choose to submit your solution, please send an e-mail from your `ethz.ch` address with subject `Exercise4` containing a PDF (LaTeX or scan) to Baharan Mirzasoleiman `baharanm@inf.ethz.ch` until Tuesday, May 9st 2017.

Problem 1 (K-means convergence):

In the K-means clustering algorithm, you are given a set of n points $x_i \in \mathbb{R}^d$, $i \in \{1, \dots, n\}$ and you want to find the centers of k clusters $\mu = (\mu_1, \dots, \mu_k)$ by minimizing the average distance from the points to the closest cluster center. Formally, you want to minimize the following loss function

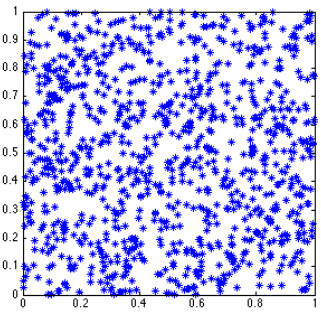
$$L(\mu) = \sum_{i=1}^{n} \min_{j \in \{1,\dots,k\}} \|x_i - \mu_j\|_2^2.$$

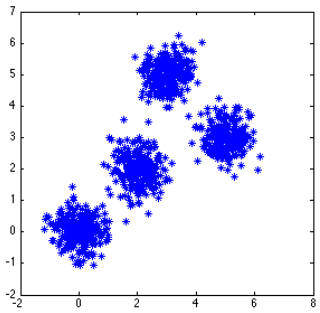
To approximate the solution, we introduce new assignment variables $z_i \in \arg\min_{j \in \{1,\dots,k\}} \|x_i - \mu_j\|_2^2$ for each data point x_i. The K-means algorithm iterates between updating the variables z_i (*assignment step*) and updating the centers $\mu_j = \frac{1}{|\{i:z_i=j\}|} \sum_{i:z_i=j} x_i$ (*refitting step*). The algorithm stops when no change occurs during the *assignment step*.

Show that K-means is guaranteed to converge (to a local optimum). *Hint:* You need to prove that the loss function is guaranteed to decrease monotonically in each iteration until convergence. Prove this separately for the *assignment step* and the *refitting step* .

Problem 2 (K-means initialization):

You are given two example datasets consisting of 1000 two dimensional points each. We want to find 4 clusters in each of them.





We know that K-means is not robust to initialization. Can you provide two different initializations for each of the datasets that would result in qualitatively different clusters? Sketch initializations and resulting clusters.

Problem 3 (K-medians clustering):

In this exercise, you are asked to derive a new clustering algorithm that would use a different loss function given by

$$L(\mu) = \sum_{i=1}^{n} \min_{j \in \{1,\dots,k\}} \|x_i - \mu_j\|_1.$$

(a) Find the update steps for both z_i and for μ_j in this case.

(b) What can you say about the convergence of your algorithm?

(c) In which situation would you prefer to use K-medians clustering instead of K-means clustering?