CSE 151A - Homework 08

Due: Wednesday, May 27, 2020

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. Unless otherwise noted by the problem's instructions, show your work or provide some justification for your answer. Homeworks are due via Gradescope on Wednesday at 11:59 p.m.

Essential Problem 1.

Consider the following set of data points in $\mathbb{R}^2$.

x_1	x_2
0.2	0.5
0.4	0.1
0.3	0.1
1.0	1.2
1.3	1.1
1.4	1.0

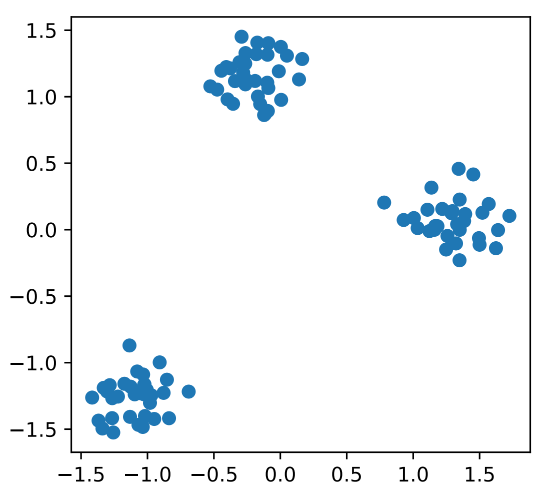
Run k-means with $k = 2$ and initial cluster centers $\vec{\mu}^{(1)} = (1.1, 1.1)^T$ and $\vec{\mu}^{(2)} = (1.3, 0.8)^T$ until convergence. At every step, give the current cluster center positions, and show your work.

Essential Problem 2.

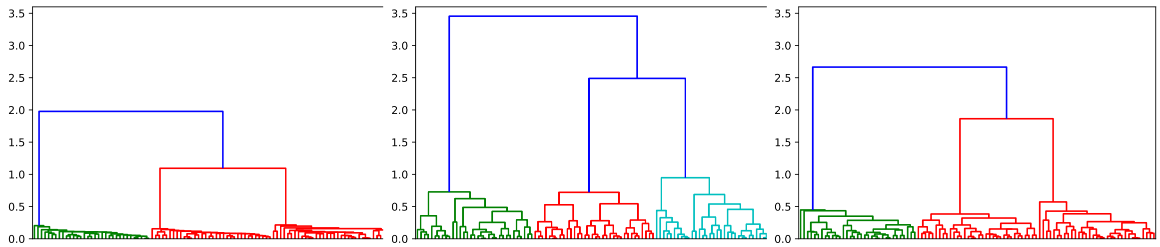
Using the same data as in the problem above, run the EM algorithm to fit a mixture of two Gaussians (both with full covariance matrices). Intialize the means of the Gaussians to be $\vec{\mu}^{(1)} = (1.1, 1.1)^T$ and $\vec{\mu}^{(2)} = (1.3, 0.8)^T$, set the initial covariance matrices to be the identity matrix, and set the initial mixing coefficients to both be 0.5. Run 3 iterations of the algorithm, reporting the cluster means, covariances, and mixing coefficients at each step, as well as the responsibilities of each point.

Essential Problem 3.

Consider the following set of data points:



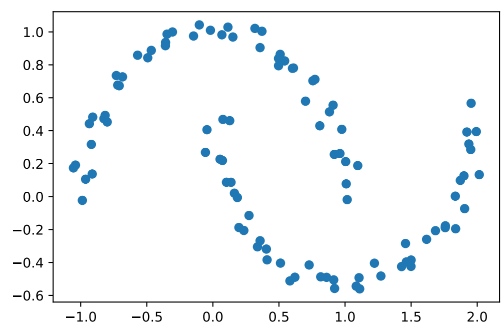
Single-linkage, average-linkage, and complete-linkage clustering were all performed on this data. Which of the dendrograms below is the result of single-linkage, which is the result of average-linkage, and which is the result of complete-linkage? Explain your reasoning.



(a) (b) (c)

Essential Problem 4.

Consider the data below:



Out of all of the clustering methods discussed this week, which method would you use to cluster this data if you believe that there are two clusters? Explain your reasons.

Plus Problem 1. (8 plus points)

In lecture, it was claimed that Lloyd's algorithm will converge to a local minimum of the k-means objective function. In this problem, we'll prove that claim.

a) Let $\vec{x}^{(1)}, \ldots, \vec{x}^{(n)}$ be a set of n vectors in d-dimensional space. Define:

$$L(\vec{v}) = \sum_{i=1}^{n} \|\vec{x}^{(i)} - \vec{v}\|^2.$$

You can think of L as a convex function measuring how well the vector $\vec{v}$ represents the set of data points, with smaller numbers being better. Prove that L is minimized by the mean, $\vec{\mu} = \frac{1}{n}\sum_{i=1}^{n} \vec{x}^{(i)}$.

b) Recall that the k-means objective function is defined to be:

$$\text{Cost}(\vec{\mu}^{(1)}, \ldots, \vec{\mu}^{(k)}) = \frac{1}{n}\sum_{i=1}^{n} \min_{j \in \{1,\ldots,k\}} \|\vec{x}^{(i)} - \vec{\mu}^{(j)}\|^2$$

Prove that the value of the k-means objective decreases (or stays the same) between iterations of Lloyd's algorithm.

Hint: Consider an arbitrary iteration of Lloyd's algorithm. Let $\vec{\mu}^{(1)}, \ldots, \vec{\mu}^{(k)}$ be the centers before the iteration, and let $\vec{\nu}^{(1)}, \ldots, \vec{\nu}^{(k)}$ be the centers after the iteration. Show that the cost of the new centers is smaller than the cost of the old centers. Remember that the nearest cluster center to a data point can change in an iteration, so make sure to take this into account.

After you have proven (a) and (b), it only remains to be shown that the algorithm terminates. It can be shown that k-means takes only a finite number of iterations, and must converge eventually. Here's the idea: apart from before the initial iteration, a cluster center must be the mean of a subset of data. There are exponentially-many subsets of the data set, but an exponential number is finite. Therefore there are finitely-many cluster centers. Once the algorithm has reached a certain configuration of cluster centers, this configuration will never be seen again (as it can be shown that the objective function strictly decreases unless the centers stay the same). Therefore the algorithm terminates eventually (perhaps after an exponential number of iterations).

Plus Problem 2. (6 plus points)

Consider the following set of points in 1-dimensional space:

$$\{0, 1, 3, 6, 10, 15\}$$

Simulate the robust single linkage algorithm discussed in class with $k = 2$ and $\alpha = 0.5$ and show the clusters that the algorithm forms at every step from time $r = 1$ to $r = 9$. If you use code, please provide it in your submission.