CSE 151A - Discussion 08

K-Means Clustering

- Goal : find k cluster centers $\{\vec{\mu}^{(1)} \dots \vec{\mu}^{(k)}\}$ that minimizes the K-Means cost
- $\text{Cost}(\{\vec{\mu}^{(1)} \dots \vec{\mu}^{(k)}\}) = \frac{1}{n} \sum_{i=1}^{n} \min_{j \in \{1 \dots k\}} ||\vec{x}^{(i)} - \vec{\mu}^{(j)}||^2$
  - This is the average squared distance from each data point to its nearest cluster center

Lloyd's Algorithm for K-Means

- Initialize $\{\vec{\mu}^{(1)} \dots \vec{\mu}^{(k)}\}$ in some way
- Repeat until no change in cost :
  - Assign each point $\vec{x}^{(i)}$ to closest center
  - Update $\vec{\mu}^{(i)}$ to be mean of points assigned to it
- Key Facts
  - Converges to local optimum of K-Means cost
  - Cost monotonically decreases as the algorithm progresses
  - Number of iterations unknown
  - Quality of solution depends heavily on initialization

K-Means++ Initialization

- Pick $\vec{\mu}^{(1)}$ uniformly at random from data
- Let $C = \{\vec{\mu}^{(1)}\}$ be the centers chosen so far
- Repeat $k-1$ times :
  - Pick $\vec{x}$ at random, with probability $P(\vec{x}) \propto \min_{\vec{\mu} \in C} ||\vec{x} - \vec{\mu}||^2$
  - Add $\vec{x}$ to C

Gaussian Mixture Models + EM Algorithm

- Mixture of k Gaussians : $\mathbb{P}(\vec{x}) = \sum_{j=1}^{k} \pi_j P_j(\vec{x}^{(i)})$
- Single Gaussian : $P_j = \mathcal{N}(\vec{\mu}^{(j)}, C_j)$
  - Mean : $\vec{\mu}^{(j)} = \frac{1}{\sum_{i=1}^{n} w_{ij}} \sum_{i=1}^{n} w_{ij} \vec{x}^{(i)}$
  - Covariance matrix : $C_j = \frac{1}{\sum_{i=1}^{n} w_{ij}} \sum_{i=1}^{n} w_{ij} \left(\vec{x}^{(i)} - \vec{\mu}^{(j)}\right)\left(\vec{x}^{(i)} - \vec{\mu}^{(j)}\right)^T$
- Mixing weight : $\pi_j = \frac{1}{n} \sum_{i=1}^{n} w_{ij}$

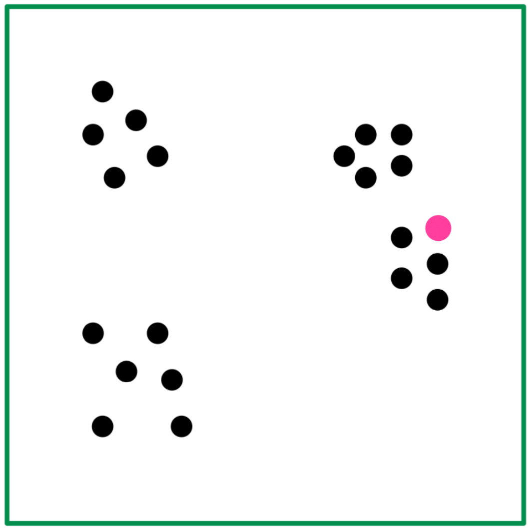
- Responsibility of cluster j for point i : $w_{ij} = \frac{\pi_j P_j(\vec{x}^{(i)})}{\sum_l \pi_l P_l(\vec{x}^{(i)})}$
- Algorithm:
  - Initialize $\pi_1 \cdots \pi_k, \vec{\mu}^{(1)} \cdots \vec{\mu}^{(k)}, C_1 \cdots C_k$
  - Make soft assignment (update responsibilities w_{ij})
  - Update mixing weights (π_j), means ($\vec{\mu}^{(j)}$), covariances (C_j)

Hierarchical Clustering Basics

- Single Linkage : $\mathcal{L}(C, C') = \min_{x,x' \in C,C'} d(x, x') \leftarrow$ smallest distance between any pair of points
- Complete Linkage : $\mathcal{L}(C, C') = \max_{x,x' \in C,C'} d(x, x') \leftarrow$ largest distance between any pair of points
- Average Linkage : $\mathcal{L}(C, C') = \frac{1}{|C||C'|} \sum_{x,x' \in C,C'} d(x, x') \leftarrow$ average distance between all point pairs
- Density Cluster Tree : For a probability density function f, assign clusters $C_f(\lambda)$ to connected components of $\{f \geq \lambda\}$ for any $\lambda > 0$

Problem 1.

Given the data points below, assign $k = 4$ cluster centers using the K-Means++ initialization algorithm, with the initial choice of $\vec{\mu}^{(1)}$ shown in pink. (Note that there are many possible correct solutions.)
Once you have chosen the cluster centers, draw the boundary lines to define the $k = 4$ convex regions.



Problem 2.

For a cluster S consisting of n points, and any arbitrary cluster center $\vec{\mu}$, we define the K-Means cost as :

$$\text{Cost}(\{\vec{\mu}\}) = \frac{1}{n}\sum_{i=1}^{n} ||\vec{x}^{(i)} - \vec{\mu}||^2$$

We can also define the following lemma over any cluster S and any arbitrary cluster center $\vec{\mu}$:

$$\text{Cost}(\{\vec{\mu}\}) = \text{Cost}(\{\bar{x}\}) + ||\vec{\mu} - \bar{x}||^2, \text{ where } \bar{x} = \text{mean}(S)$$

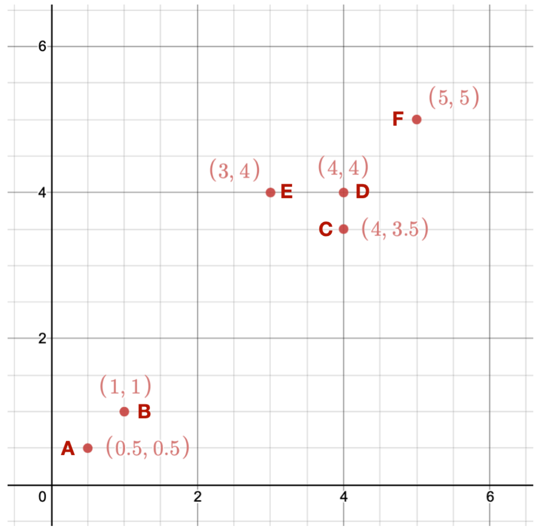
Using the above lemma, prove that for any data point $\vec{x}^{(i)}$ chosen randomly from S, it is true that :

$$\mathbb{E}_{\vec{x}^{(i)}}[\text{Cost}(\vec{x}^{(i)})] = 2 \cdot \text{Cost}(\bar{x})$$

This is to say that the expected value of the cost of clustering set S with any randomly chosen center $\vec{x}^{(i)}$ is twice the cost of a clustering with center $\bar{x} = \text{mean}(S)$.

Problem 3.

Run the single-linkage clustering algorithm on the data points below, drawing the cluster links as you go. Once you are left with a single cluster, draw the corresponding dendrogram.


6
4
2
0
2
4
6
(5, 5)
F
(3, 4)
E
(4, 4)
D
C
(4, 3.5)
(1, 1)
B
A
(0.5, 0.5)