Week 7 Practice

CSCI 567 Machine Learning

Fall 2025

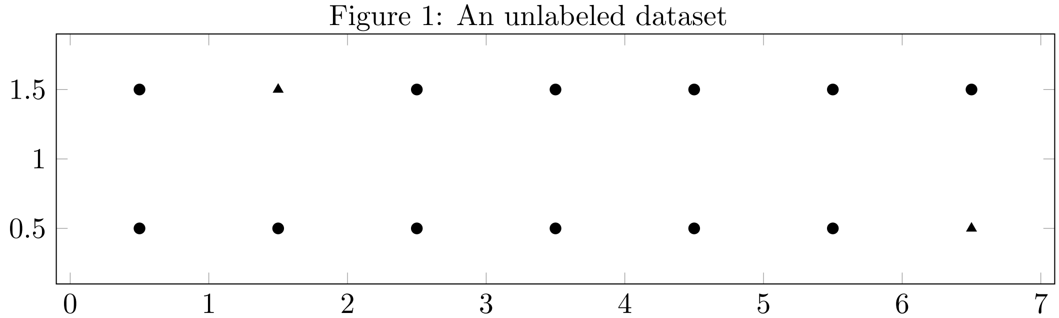
Instructor: Haipeng Luo

1 Clustering

Consider the following dataset. All points are unlabeled and part of the same set. The triangles are used to distinguish two points later.

Suppose that we run the K-means algorithm on this dataset with $K = 2$ and the two points indicated by triangles as the initial centers. When the algorithm converges, there will be two clearly separated clusters. Directly on Figure 1, draw a straight line that separates these two clusters, as well as the centers of these two clusters.

Figure 1: An unlabeled dataset

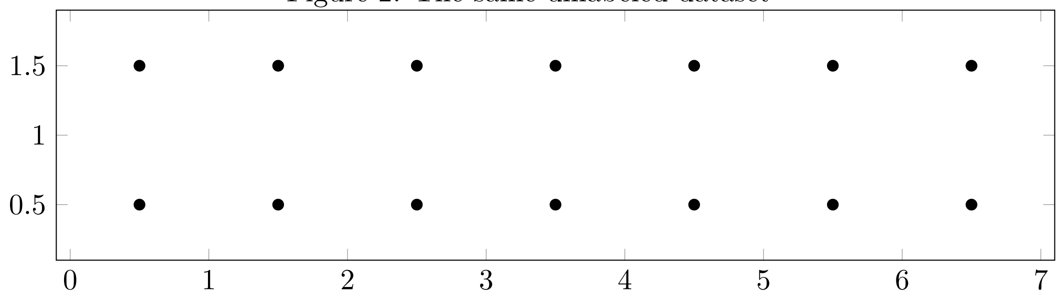


Next, find two other different sets of initialize centers that will converge to the exact same result if we apply K-means. Please follow the instructions below:

- the initialize centers have to be points of the dataset;
- directly on Figure 2, use two triangles to indicate the first set, and two squares to indicate the second;
- these two sets of points can overlap with each other, but of course cannot be the same;

- similarly these sets can overlap with the initialization of Figure 1 but cannot be the same;
- do not pick those that lead to ambiguous results due to different ways of breaking ties.

Figure 2: The same unlabeled dataset



2 MLE, Mixtures, and EM

In this problem you will practice MLE and EM.

(a) Let $X \in \mathbb{R}$ be a random variable uniformly-distributed on some unknown interval $(0, \theta]$, where $\theta > 0$. More specifically, the density function is

$$P(X = x\,; \theta) = \begin{cases} \frac{1}{\theta} & , \text{if } x \in (0, \theta], \\ 0 & , \text{otherwise}, \end{cases} \tag{1}$$

$$= \frac{1}{\theta} \mathbf{1}[0 < x \leq \theta], \tag{2}$$

where $\mathbf{1}[\cdot]$ is an indicator function that outputs 1 when the input condition is true, and 0 otherwise. Suppose that $x_1, x_2, \ldots, x_N$ are i.i.d. samples from this distribution. Write down the likelihood of the observations and then find the maximum likelihood estimator (MLE).

(b) Now suppose that X is distributed according to a **mixture** of two uniform distributions: one on interval $(0, \theta_1]$ and the other on $(0, \theta_2]$, for some unknown $\theta_1, \theta_2 > 0$. More specifically, the density function is

$$\begin{aligned} P(X = x) &= P(X = x, z = 1) + P(X = x, z = 2) \\ &= P(z = 1)P(X = x \mid z = 1) + P(z = 2)P(X = x \mid z = 2) \\ &= \omega_1 U(X = x\,; \theta_1) + \omega_2 U(X = x\,; \theta_2) \end{aligned}$$

where U is the uniform distribution defined as in Eq. (1) or Eq. (2), and ω_1, ω_2 are mixture weights such that

$$\omega_1 \geq 0, \omega_2 \geq 0, \text{ and } \omega_1 + \omega_2 = 1.$$

Suppose that $x_1, x_2, \ldots, x_N$ are i.i.d. samples from this mixture of uniform distributions. MLE does not admit a closed-form for this problem, and we will use the EM algorithm to approximately find the MLE.

- First, the E-Step fixes a set of parameters $\theta_1, \theta_2, \omega_1, \omega_2$ and computes for each n the posterior distribution $\gamma_{nk} = P(z_n = k \mid x_n\,; \theta_1, \theta_2, \omega_1, \omega_2)$ of the latent variable z_n, where $k \in \{1, 2\}$ indicates which mixture component x_n belongs to. Write down the explicit form of this posterior distribution *without using the proportional notation*. Then write down explicitly the expected complete log-likelihood using γ_{nk} (as a function of the four parameters $\theta_1, \theta_2, \omega_1, \omega_2$).

- Next, derive the M-Step by maximizing the expected complete log-likelihood you derived from the last problem over the four parameters. Hint: you will need to use the fact $0 \ln 0 = 0$.