Labs
Machine Learning Course
Fall 2025

Westlake University
Department of Artificial Intelligence, SOE
Tao Lin
https://github.com/LINs-lab/course_machine_learning

Problem Set Lab 11(graded), Dec 11, 2025
(Beyond Supervised Learning)

Goals. The goals of this exercise are to:

- Apply one iteration of K-means clustering and update cluster centers in Euclidean space.
- Carry out the E-step and M-step for a two-component Gaussian mixture model.
- Implement one round of label propagation on a graph for semi-supervised learning.
- Derive the form of the optimal discriminator in the GAN minimax objective.
- Obtain the Evidence Lower Bound (ELBO) for a variational autoencoder and understand its gap to the log-likelihood.

Submission instructions:

- Please submit a PDF file to canvas.

Problem 1 (K-means Clustering (20)):
Given the six two-dimensional points

$$\mathcal{X} = \{(0,0),\ (0,2),\ (2,0),\ (2,2),\ (8,8),\ (9,9)\},$$

we wish to cluster them into $K = 2$ clusters using Euclidean distance. The initial cluster centers are

$$\mu_1^{(0)} = (0,0), \quad \mu_2^{(0)} = (9,9).$$

1. Perform one iteration of K-means: assign each point to its nearest center, and list the assignments.
2. Recompute the cluster centers $\mu_1^{(1)}$ and $\mu_2^{(1)}$ based on these assignments.

Problem 2 (EM Algorithm for a One-Dimensional Gaussian Mixture Model (20)):
Let the observed data be

$$\mathcal{D} = \{\, x_1 = 0,\ x_2 = 1,\ x_3 = 5\},$$

and initialize the two-component GMM by

$$\pi_1 = \pi_2 = 0.5, \quad \mu_1 = 0,\ \mu_2 = 5, \quad \sigma_1^2 = \sigma_2^2 = 1.$$

1. Compute the posterior responsibilities

$$\gamma_{ik} = \frac{\pi_k\, \mathcal{N}(x_i \mid \mu_k, 1)}{\sum_{j=1}^{2} \pi_j\, \mathcal{N}(x_i \mid \mu_j, 1)}, \quad i = 1,2,3,\ k = 1,2,$$

   and give their values (you may leave them in exponential or density form).

2. Write down the M-step update formulas

$$\pi_k^{\mathrm{new}} = \frac{1}{N}\sum_{i=1}^{N} \gamma_{ik}, \quad \mu_k^{\mathrm{new}} = \frac{\sum_{i=1}^{N} \gamma_{ik}\, x_i}{\sum_{i=1}^{N} \gamma_{ik}}, \quad (\sigma_k^2)^{\mathrm{new}} = \frac{\sum_{i=1}^{N} \gamma_{ik}(x_i - \mu_k^{\mathrm{new}})^2}{\sum_{i=1}^{N} \gamma_{ik}}.$$

Problem 3 (Label Propagation in Semi-Supervised Learning (20)):
Consider the undirected graph $G = (V, E)$ with

$$V = \{1, 2, 3, 4\}, \quad w_{12} = w_{21} = 1,\ w_{23} = w_{32} = 1,\ w_{34} = w_{43} = 1, \quad w_{ij} = 0 \text{ otherwise.}$$

Nodes 1 and 4 are labeled $y_1 = +1$ and $y_4 = -1$; nodes $2, 3$ are unlabeled. We keep $f_1 = +1$ and $f_4 = -1$ fixed, and initialize

$$f_2^{(0)} = 0, \quad f_3^{(0)} = 0.$$

1. Construct the row-normalized transition matrix P with
$$P_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}.$$

2. Perform one label-propagation iteration:
$$f_i^{(t+1)} = \sum_j P_{ij}\, f_j^{(t)}, \quad i = 2, 3,$$
keeping f_1, f_4 fixed. Write the update equations for f_2 and f_3, and compute $f_2^{(1)}$ and $f_3^{(1)}$.

Problem 4 (Generative Adversarial Network (20)):
The minimax objective is

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))].$$

Fix the generator G and show that the optimal discriminator is

$$D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)}.$$

Problem 5 (Variational Autoencoder ELBO (20)):
Let

$$p_\theta(x, z) = p_\theta(x \mid z)\, p(z)$$

be a latent-variable model, and let $q_\phi(z \mid x)$ be an approximate posterior. Starting from $\log p_\theta(x) = \log \int p_\theta(x, z)\, dz$, derive the Evidence Lower Bound (ELBO)

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}\big[\log p_\theta(x \mid z)\big] - \mathrm{KL}\big(q_\phi(z \mid x) \,\|\, p(z)\big),$$

and show that $\mathcal{L}(\theta, \phi; x) \le \log p_\theta(x)$.