HW4 Handwritten Assignment Solution

Lecturor: Pei-Yuan Wu
TAs: Chun-Lin Huang(Problem 1, 2), Yuan-Chia Chang(Problem 3, 4, 5)

December 2023, First Edition

Problem 1 (EM algorithm for mixture of Bernoulli model)(1.5%)

Consider the generative model parameterized by $\theta = (\pi_k, \boldsymbol{\mu}_k)_{k=1}^K$, where $\pi_1, ..., \pi_K \in [0,1]$ satisfies $\sum_{k=1}^K \pi_k = 1$, and that $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K \in [0,1]^D$, so that the probability of generating a D-dimensional binary vector $\mathbf{x} = (x^{(1)}, ..., x^{(D)}) \in \{0,1\}^D$ is

$$p(\mathbf{x}; \theta) = \sum_{k=1}^K \pi_k \prod_{j=1}^D \mu_{kj}^{x^{(j)}} (1-\mu_{kj})^{1-x^{(j)}}$$

In other words, with given $\boldsymbol{\mu}_k$, the elements $x^{(1)}, ..., x^{(D)}$ are independent, where $x^{(j)}$ follows Bernoulli distribution of mean μ_{kj}. Suppose we observe training data of N binary vectors $\mathbf{x}_1, ..., \mathbf{x}_N \in \{0,1\}^D$, derive the E-step and M-step equations of the EM algorithm for optimizing the mixing coefficients π_k and the Bernoulli means μ_{kj} by maximum likelihood.

Solution: Denote $\mathcal{X} = (\mathbf{x}_i)_{i=1}^N$ and latent variables $\mathcal{Z} = (z_1, ..., z_N) \in 1, K^N$, and consider the joint probablity distribution as

$$p(\mathcal{X}, \mathcal{Z}; \theta) = \prod_{i=1}^N p(\mathbf{x}_i, z_i; \theta), \text{ where } p(\mathbf{x}, z = k; \theta) = \pi_k \prod_{j=1}^D \mu_{kj}^{x^{(j)}} (1-\mu_{kj})^{1-x^{(j)}}$$

Start with initial guess $\theta^{(0)}$, the EM algorithm estimates $\theta^{(t+1)} = (\pi_k^{(t+1)}, \boldsymbol{\mu}_k^{(t+1)})_{k=1}^K$ from $\theta^{(t)} = (\pi_k^{(t)}, \boldsymbol{\mu}_k^{(t)})_{k=1}^K$ as follows:

- Expectation Step: Compute

$$Q(\theta|\theta^{(t)}) = \mathbb{E}_{\mathcal{Z}|\mathcal{X};\theta^{(t)}}[\log p(\mathcal{X}, \mathcal{Z}; \theta)] = \sum_{i=1}^N \mathbb{E}_{z_i|\mathbf{x}_i;\theta^{(t)}}[\log p(\mathbf{x}_i, z_i; \theta)]$$

  The posterior probability of latent variables based on current parameters $\theta^{(t)}$:

$$\mathbb{P}[z_i{=}k|\mathbf{x}_i;\theta^{(t)}]{=}\frac{p(\mathbf{x}_i,z_i{=}k;\theta^{(t)})}{\sum_{k'=1}^K p(\mathbf{x}_i,z_i{=}k';\theta^{(t)})}{=}\frac{\pi_k \prod_{j=1}^D \mu_{kj}^{x_i^{(j)}} (1{-}\mu_{kj})^{1-x_i^{(j)}}}{\sum_{k'=1}^K \pi_{k'} \prod_{j=1}^D \mu_{k'j}^{x_i^{(j)}} (1{-}\mu_{k'j})^{1-x_i^{(j)}}}{=}\delta_{ik}^{(t)}$$

  The log-likelihood of parameter θ for jointly generating $\mathbf{x}_i$ and z_i:

$$\log p(\mathbf{x}_i, z_i = k; \theta) = \log \pi_k + \sum_{j=1}^D \left(x_i^{(j)} \log \mu_{kj} + (1 - x_i^{(j)}) \log(1-\mu_{kj}) \right)$$

  Hence

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^N \sum_{k=1}^K \delta_{ik}^{(t)} \left(\log \pi_k + \sum_{j=1}^D \left(x_i^{(j)} \log \mu_{kj} + (1 - x_i^{(j)}) \log(1-\mu_{kj}) \right) \right)$$

- Maximization Step: Choose $\theta^{(t+1)} =_\theta Q(\theta|\theta^{(t)})$. Note that

$$\frac{\partial}{\partial \mu_{kj}} Q(\theta|\theta^{(t)}) = \sum_{i=1}^N \delta_{ik}^{(t)} \left(\frac{x_i^{(j)}}{\mu_{kj}} - \frac{1 - x_i^{(j)}}{1 - \mu_{kj}} \right)$$

By setting partial derivative to zero, the maximal solution for μ_{kj} is

$$\mu_{kj}^{(t+1)} = \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)} x_i^{(j)}}{\sum_{i=1}^{N} \delta_{ik}^{(t)}}$$

As for π_k, due to the constraint $\sum_{k=1}^{K} \pi_k = 1$, we introduce Lagrange multipliers and note

$$\frac{\partial}{\partial \pi_k} \left(Q(\theta|\theta^{(t)}) - \lambda \sum_{k=1}^{K} \pi_k \right) = \frac{1}{\pi_k} \sum_{i=1}^{N} \delta_{ik}^{(t)} - \lambda.$$

By setting the above quantities identially zero for all k, we solve $\lambda = N$ and the maximal solution for π_k is

$$\pi_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \delta_{ik}^{(t)}.$$

Problem 2 (EM algorithm for mixture of exponential model)(1.5%)

Given N samples $x_1, \ldots, x_N \in [0, \infty)$, we would like to cluster them into K clusters. Assume the samples are generated according to Exponential mixture models

$$X \sim \sum_{j=1}^{K} \pi_j Exp(\tau_j)$$

where $\pi_1 + \ldots + \pi_K = 1$, and $Exp(\tau)$ denotes the exponential distribution with probability density function

$$f_\tau(x) = \begin{cases} (1/\tau)e^{-x/\tau} & , \ x \geq 0 \\ 0 & , x < 0 \end{cases}$$

We would like to apply Expectation Maximization algorithm to find the maximum likelihood estimation of parameters $\theta = \{(\pi_k, \tau_k)\}_{k=1}^{K}$.

(a) Please write down the E-step and M-step and show that the parameters are updated from $\theta^{(t)} = \left\{(\pi_k^{(t)}, \tau_k^{(t)})\right\}_{k=1}^{K}$ to $\theta^{(t+1)} = \left\{(\pi_k^{(t+1)}, \tau_k^{(t+1)})\right\}_{k=1}^{K}$ in the following form:

$$\tau_k^{(t+1)} = \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)} x_i}{\sum_{i=1}^{N} \delta_{ik}^{(t)}}, \quad \pi_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \delta_{ik}^{(t)}$$

(b) What is the closed form expression of $\delta_{ik}^{(t)}$?

Solution:

Given data set $\mathcal{X} = \{x_1, \ldots, x_N\}$, the likelihood and log-likelihood functions are given by

$$p(\mathcal{X};\theta) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k f_{\tau_k}(x_i), \quad \log p(\mathcal{X};\theta) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k f_{\tau_k}(x_i) \right)$$

Denote latent variable $z_i \in \{1, \ldots, K\}$ indicating which Exponential distribution x_i is drawn from.

- Expectation Step (E-step): Compute

$$Q\left(\theta \middle| \theta^{(t)}\right) = \sum_{i=1}^{N} \mathbb{E}_{z_i|\mathbf{x}_i;\theta^{(t)}} \left[\log p(\mathbf{x}_i, z_i; \theta)\right]$$

- Posterior prob. dist. of latent variables z_i based on current parameters $\theta^{(t)}$:

$$\mathbb{P}\left[z_i = k \middle| x_i; \theta^{(t)}\right] = \frac{p\left(x_i, z_i = k; \theta^{(t)}\right)}{\sum_{j=1}^{K} p\left(x_i, z_i = j; \theta^{(t)}\right)} = \frac{\pi_k^{(t)} f_{\tau_k^{(t)}}\left(x_i\right)}{\sum_{j=1}^{K} \pi_j^{(t)} f_{\tau_j^{(t)}}\left(x_i\right)}$$

$$= \frac{\left(\pi_k^{(t)} / \tau_k^{(t)}\right) e^{-x_i/\tau_k^{(t)}}}{\sum_{j=1}^{K} \left(\pi_j^{(t)} / \tau_j^{(t)}\right) e^{-x_i/\tau_j^{(t)}}} = \delta_{ik}^{(t)}$$

- Log-likelihood of parameter θ given data x_i and latent variable z_i:

$$\log p\left(x_i, z_i = k; \theta\right) = \log \pi_k f_{\tau_k}\left(x_i\right) = \log\left(\frac{\pi_k}{\tau_k}\right) - \frac{x_i}{\tau_k}$$

  Hence

$$Q\left(\theta \middle| \theta^{(t)}\right) = \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{ik}^{(t)} \left\{\log\left(\frac{\pi_k}{\tau_k}\right) - \frac{x_i}{\tau_k}\right\}$$

- Maximization Step (M-step): Choose

$$\theta^{(t+1)} = \arg\max_{\theta \in \Theta} Q\left(\theta \middle| \theta^{(t)}\right)$$

- Partial derivative over τ_k:

$$\frac{\partial}{\partial \tau_k} \log Q\left(\theta \middle| \theta^{(t)}\right) = \sum_{i=1}^{N} \delta_{ik}^{(t)} \left(\frac{x_i}{\tau_k^2} - \frac{1}{\tau_k}\right)$$

  Setting derivate equals to zero $\Longrightarrow$ Take $\tau_k^{(t+1)} = \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)} x_i}{\sum_{i=1}^{N} \delta_{ik}^{(t)}}$.

- Partial derivative over π_k:

  By constraint $\sum_{k=1}^{K} \pi_k = 1$, invoke Lagrange multiplier

$$\nabla_{\pi_k}\left(\log Q\left(\theta \middle| \theta^{(t)}\right) - \lambda \sum_{k=1}^{K} \pi_k\right) = \sum_{i=1}^{N} \frac{\delta_{ik}^{(t)}}{\pi_k} - \lambda$$

  Setting derivate equals to zero $\Longrightarrow$ Take $\pi_k^{(t+1)} = \lambda^{-1} \sum_{i=1}^{N} \delta_{ik}^{(t)}$. The constraint $\sum_{k=1}^{K} \pi_k^{(t+1)} = 1$ further leads to $\lambda = N$ and $\pi_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \delta_{ik}^{(t)}$.

Problem 3 (Boosting)(0.5%)

1. Consider training a boosting classifier using decision stumps on the data set illustrated in Figure 1:

Figure 1: AdaBoost Data set

   (a) Which examples will have their weights increased at the end of the first iteration? Circle them.

   (b) How many iterations will it take to achieve zero training error? Justify your answers.

2. Suppose AdaBoost is run on N training examples, and suppose on each round that the weighted training error ϵ_t of the t'th weak hypothesis is at most $1/2 - \gamma$, for some number $0 < \gamma < 1/2$. After how many iterations, T, will the combined hypothesis be consistent with the N training examples, i.e., achieves zero training error? Your answer should only be expressed in terms of N and γ. (Hint: Recall that exponential loss is an upper bound for 0-1 loss. What is the training error when 1 example is misclassified?)

Solution:

1. (a) We execute Adaboost algorithm on the data set. First we note that if the classifier classifies the symbol as circle, $f(x) = 1$. On the other hand, if the classfier classifies the symbol as cross, $f(x) = -1$.We choose our first decision stump as a horizontal line to the right of the plane, which is the red line in the following figure. We consider all points as the cross symbol. $\epsilon_1 = 0.2, \alpha_1 = ln(\sqrt{\frac{0.8}{0.2}}) = 0.693$. For the cross symbol, the weights becomes to 0.5, and the weight of the circle symbol becomes to 2. Hence, the weight of the circle symbol increases at the end of the first iteration.

   (b) We choose the first decision stump as we mentioned above. Next, we choose the blue line in the following figure as the second classifier. The classifier classifies the symbol as the circle if it is on the left to the line, and classifies as the cross if it is on the right to the line. At last, the third decision stump is the green line, which classifies the right hand side symbol as circle and the left hand side symbol as cross. After executing, $\alpha_1 = 0.693, \alpha_2 = 0.549, \alpha_3 = 0.822$. $H(x) = sign(\alpha_1 f_1(x) + \alpha_2 f_2(x) + \alpha_3 f_3(x))$. For the circle symbol, $H(x) = sign(0.678) = 1$. For the left crosses, $H(x) = sign(-0.966) = -1$. For the right crosses, $H(x) = sign(-0.42) = -1$. From the above result, we achieve zero training error by these decision stumps.

2. The training error after T iterations is upper bounded by

$$\prod_{t=1}^{T} \left(2\sqrt{\epsilon_t(1-\epsilon_t)}\right) \leq \prod_{t=1}^{T} \left(2\sqrt{(1/2-\gamma)(1/2+\gamma)}\right) = (1-4\gamma^2)^{T/2}$$

   Zero training error can be achieved when $(1-4\gamma^2)^{T/2} < 1/N$, namely

$$T > \frac{2\log N}{\log(1/(1-4\gamma^2))} \geq \frac{\log N}{2\gamma^2}.$$

   Here the second equality indicates a (looser but simpler) lower bound for T.

Figure 2: First decision stump f_1

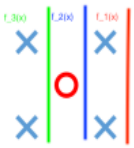

Figure 3: An example of to achieve zero training error by three decision stumps

Problem 4 (Expectation Maximization Interpretation behind Semi-Supervised Learning)(1%)

Given N samples $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^m$ as well as their labels $y_1, \ldots, y_N \in \{0, 1, \ldots, K\}$. Consider the generative model where each sample $\mathbf{x}_i$ is generated independently according to Gaussian mixture model that depends on the label y_i, as represented by random variable

$$X_i \sim \begin{cases} \sum_{j=1}^{K} \pi_j \mathcal{N}\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) & , \ if \ y_i = 0 \\ \mathcal{N}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) & , \ if \ y_i = k \neq 0 \end{cases}$$

where $\pi_1 + \ldots + \pi_K = 1$, and $\mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, with probability density function

$$\mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{1}{\sqrt{(2\pi)^m \left|\boldsymbol{\Sigma}\right|}} \exp\left(-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}\right)\right)$$

We would like to apply Expectation Maximization algorithm to find the maximum likelihood estimation of parameters $\theta = \{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^{K}$.

1. Please write down the E-step and M-step and show that the parameters are updated from $\theta^{(t)} = \left\{\left(\pi_k^{(t)}, \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)\right\}_{k=1}^{K}$ to $\theta^{(t+1)} = \left\{\left(\pi_k^{(t+1)}, \boldsymbol{\mu}_k^{(t+1)}, \boldsymbol{\Sigma}_k^{(t+1)}\right)\right\}_{k=1}^{K}$ in the following form:

$$\pi_k^{(t+1)} = \frac{\sum_{i:y_i=0} \delta_{ik}^{(t)}}{\sum_{i:y_i=0} 1}$$

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i:y_i=k} \mathbf{x}_i + \sum_{i:y_i=0} \delta_{ik}^{(t)} \mathbf{x}_i}{N_k + \sum_{i:y_i=0} \delta_{ik}^{(t)}}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i:y_i=k} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T + \sum_{i:y_i=0} \delta_{ik}^{(t)} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T}{N_k + \sum_{i:y_i=0} \delta_{ik}^{(t)}}$$

where $N_k = \sum_{i:y_i=k} 1$ is the number of samples in class k. Please show your derivations.

2. What is the closed form expression of $\delta_{ik}^{(t)}$? Please show your derivations.

Solution: Given data set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, the likelihood and log-likelihood functions are given by

$$p(S; \theta) = \left(\prod_{i:y_i=0} \sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)\right) \cdot \prod_{k=1}^{K} \prod_{i:y_i=k} \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right),$$

$$\log p\left(\mathcal{X}; \theta\right) = \sum_{i:y_i=0} \log\left(\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)\right) + \sum_{k=1}^{K} \sum_{i:y_i=k} \log \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$$

Denote latent variable $z_i \in \{1, \ldots, K\}$ indicating which Gaussian distribution x_i is drawn from.

- Expectation Step (E-step): Compute

$$Q\left(\theta \middle| \theta^{(t)}\right) = \sum_{i=1}^{N} \mathbb{E}_{z_i|\mathbf{x}_i; y_i; \theta^{(t)}}\left[\log p\left(\mathbf{x}_i, y_i, z_i; \theta\right)\right]$$

Posterior prob. dist. of latent variables z_i based on current parameters $\theta^{(t)}$:

If $y_i = 0$, then

$$\delta_{ik}^{(t)} = \mathbb{P}\left[z_i = k \middle| \mathbf{x}_i; y_i = 0; \theta^{(t)}\right] = \frac{p\left(\mathbf{x}_i, y_i = 0, z_i = k; \theta^{(t)}\right)}{\sum_{j=1}^{K} p\left(\mathbf{x}_i, y_i = 0, z_i = j; \theta^{(t)}\right)} = \frac{\pi_k^{(t)} \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)}{\sum_{j=1}^{K} \pi_j^{(t)} \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)}\right)}$$

If $y_i = l \neq 0$, then

$$\delta_{ik}^{(t)} = \mathbb{P}\left[z_i = k \middle| \mathbf{x}_i; y_i = l; \theta^{(t)}\right] = \begin{cases} 1 & , \ if \ l = k \\ 0 & , \ if \ l \neq k \end{cases}$$

- Log-likelihood of parameter θ given data (x_i, y_i) and latent variable z_i:

If $y_i = 0$, then

$$\log p\left(\mathbf{x}_i, y_i = 0, z_i = k; \theta\right) = \log\left(\frac{\pi_k}{\sqrt{(2\pi)^m \left|\boldsymbol{\Sigma}_k\right|}}\right) - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)$$

If $y_i = l \neq 0$, then

$$log(\mathbf{x}_i, y_i = k, z_i = l; \theta) = \begin{cases} \log(\frac{1}{\sqrt{(2\pi)^m |\Sigma_k|}}) - \frac{1}{2}(\mathbf{x}_1 - \mu_k)^T \Sigma_k^{-1}(\mathbf{x}_1 - \mu_k) & l = k, \\ log(0) & l \neq k. \end{cases}$$

Hence

$$Q\left(\theta \middle| \theta^{(t)}\right) = \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{ik}^{(t)} \left\{ \log\left(\frac{\pi_k^{\mathbb{1}(y_i=0)}}{\sqrt{(2\pi)^m \left|\boldsymbol{\Sigma}_k\right|}}\right) - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k) \right\}$$

- Maximization Step (M-step): Choose

$$\theta^{(t+1)} = \arg\max_{\theta \in \Theta} Q\left(\theta \middle| \theta^{(t)}\right)$$

Note that $Q\left(\theta \middle| \theta^{(t)}\right)$ takes exactly the same form as in the unsupervised GMM scenario, hence the optimal solution $\theta^{(t+1)}$ is specified by

$$\pi_k^{(t+1)} = \frac{\sum_{i:y_i=0} \delta_{ik}^{(t)}}{\sum_{i:y_i=0} 1}$$

$$\boldsymbol{\mu}^{(t+1)} = \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)} \mathbf{x}_i}{\sum_{i=1}^{N} \delta_{ik}^{(t)}} = \frac{\sum_{i:y_i=k} \mathbf{x}_i + \sum_{i:y_i=0} \delta_{ik}^{(t)} \mathbf{x}_i}{N_k + \sum_{i:y_i=0} \delta_{ik}^{(t)}}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T}{\sum_{i=1}^{N} \delta_{ik}^{(t)}}$$

$$= \frac{\sum_{i:y_i=k} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T + \sum_{i:y_i=0} \delta_{ik}^{(t)} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T}{N_k + \sum_{i:y_i=0} \delta_{ik}^{(t)}}$$

Problem 5 (Label Propagation Algorithm)(1.5%)

In this problem, we will investigate label propagation algorithm by executing on a toy example. Next, we will show that the algorithm will convergence, which can be expressed analytically.

Let's consider the graph that we have seen in HW3 Problem 2.

We have previously known that x_1 node is in Class 1 and x_7 node is in Class 2. Now, we want to separate these 10 nodes into Class 1 and Class 2.

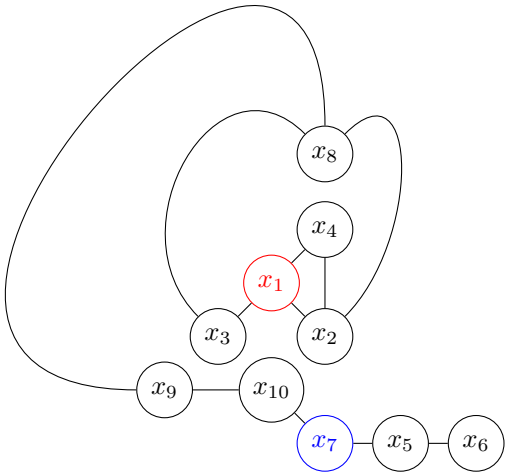

Figure 4: undirected connected graph G with labeled node

Consider the transition matrix $\boldsymbol{T}$,

$$\boldsymbol{T}_{i,j} = \frac{\tilde{\boldsymbol{W}}_{i,j}}{\sum_{k=1}^{10} \tilde{\boldsymbol{W}}_{k,j}}$$

, where $\tilde{\boldsymbol{W}}$ is the adjusted adjacency matrix of the graph G, which is defined as

$$\tilde{\boldsymbol{W}} = \boldsymbol{W} + \delta \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

. $\delta > 0$ is a small number. By adjusting the adjacency matrix, the weight of edge being connected in the original graph G is $1 + \delta$ and the weight of other edges is δ. Through the adjustment, we can prevent that the algorithm runs unsupervised due to the isolated labeled nodes. In the toy example, we set $\delta = 0.01$. $\boldsymbol{T}_{i,j}$ represents the probability that node j will propagate its own state to node i. For example, $\boldsymbol{T}_{2,1} = \boldsymbol{T}_{3,1} = \boldsymbol{T}_{4,1} \approx 0.326 \approx \frac{1}{3}, \boldsymbol{T}_{1,1} = \boldsymbol{T}_{5,1} = \boldsymbol{T}_{6,1} = \boldsymbol{T}_{7,1} = \boldsymbol{T}_{8,1} = \boldsymbol{T}_{9,1} = \boldsymbol{T}_{10,1} \approx 0.003$, which shows that node 1 will transfer its label to three neighbors with probability around 1 over 3. Also, it will transfer its label to other nodes(including itself) with probability slightly greater than zero.

1. Please write down the transition matrix $\boldsymbol{T}$.

Solution: $\boldsymbol{T}$ is shown as the figure below

```
[[0.0032 0.3258 0.481  0.481  0.0048 0.0091 0.0048 0.0032 0.0048 0.0048]
 [0.3258 0.0032 0.0048 0.481  0.0048 0.0091 0.0048 0.3258 0.0048 0.0048]
 [0.3258 0.0032 0.0048 0.0048 0.0048 0.0091 0.0048 0.3258 0.0048 0.0048]
 [0.3258 0.3258 0.0048 0.0048 0.0048 0.0091 0.0048 0.0032 0.0048 0.0048]
 [0.0032 0.0032 0.0048 0.0048 0.0048 0.9182 0.481  0.0032 0.0048 0.0048]
 [0.0032 0.0032 0.0048 0.0048 0.481  0.0091 0.0048 0.0032 0.0048 0.0048]
 [0.0032 0.0032 0.0048 0.0048 0.481  0.0091 0.0048 0.0032 0.0048 0.481 ]
 [0.0032 0.3258 0.481  0.0048 0.0048 0.0091 0.0048 0.0032 0.481  0.0048]
 [0.0032 0.0032 0.0048 0.0048 0.0048 0.0091 0.0048 0.3258 0.0048 0.481 ]
 [0.0032 0.0032 0.0048 0.0048 0.0048 0.0091 0.481  0.0032 0.481  0.0048]]
```

Figure 5: $\boldsymbol{T}$

Next, we define the label matrix sequence

$$\boldsymbol{Y}^t \in \mathbb{R}^{10\times 2} \quad t = 0, 1, ...$$

where the i_{th} row of $\boldsymbol{Y}^t$ means the probability distribution of node x_i at time t. In this example, $\boldsymbol{Y}^t_{i,1}$ means the probability that the node x_i lies in Class 1 at time t, and $\boldsymbol{Y}^t_{i,2}$ means the probability that the node x lies in Class 2 at time t. We initialize $\boldsymbol{Y}^0_{1,1} = 1, \boldsymbol{Y}^0_{1,2} = 0$ because x_1 is labeled as Class 1. Also, $\boldsymbol{Y}^0_{7,1} = 0, \boldsymbol{Y}^0_{7,2} = 1$ because x_7 is labeled as Class 2. For other nodes i, we initialize $\boldsymbol{Y}^0_{i,1} = \boldsymbol{Y}^0_{i,2} = 0.5$.

After defining label matrix $\boldsymbol{Y}$ and transition matrix $\boldsymbol{T}$, we introduce the algorithm below:

Algorithm 1 Label Propagation Algorithm in Toy Example

Input label matrix $\boldsymbol{Y}^0$, transition matrix $\boldsymbol{T}$, tolerance level ϵ
Output node $x_i \in$ {Class 1, Class 2} $\quad i = 1, , , 10$

procedure LABEL PROPAGATION($\boldsymbol{Y}^0$, $\boldsymbol{T}$, ϵ=10^{-8})
  t = 0
  repeat
    t = t + 1
    $\boldsymbol{Y}^t = \boldsymbol{T}\boldsymbol{Y}^{t-1}$ ▷ Random walk to its neighbor
    $\boldsymbol{Y}^t_{i,j} = \boldsymbol{Y}^t_{i,j}/(\boldsymbol{Y}^t_{i,1} + \boldsymbol{Y}^t_{i,2}) \quad i = 1, ..., 10, j = 1, 2$ ▷ Normalize the probability distribution
    $\boldsymbol{Y}^t_{1,1} = 1, \boldsymbol{Y}^t_{1,2} = 0, \boldsymbol{Y}^t_{7,1} = 0, \boldsymbol{Y}^t_{7,2} = 1$ ▷ Clamp the labeled data
  until $||\boldsymbol{Y}^t - \boldsymbol{Y}^{t-1}||_F < \epsilon$
  If $\boldsymbol{Y}^t_{i,1} > 0.5$ then node x_i lies in Class 1, otherwise node x_i lies in Class 2, $\quad i = 1, ..., 10$
end procedure

There are three main procedures in the loop. First, every node propagates it own state to its neighbors with the transition probability. Next, we normalize the probability distribution for every node. In the last step, we clamp the probability distribution of the labeled data, which prevents the distribution of labeled data being influenced by unlabeled data and accelerates the convergence speed of the algorithm.

2. Please execute the algorithm. Write down the iteration number t, $\boldsymbol{Y}^t$, which nodes lies in Class 1 and which nodes lies in Class 2. Does the result correspond with the graph?

Solution: $t = 65, \boldsymbol{Y}^t$ is shown below. Node 1, 2, 3, 4, 8 lies in Class 1. Node 5, 6, 7, 9, 10 lies in Class 2. The result corresponds to the graph G.

```
[[1.     0.    ]
 [0.8296 0.1704]
 [0.7938 0.2062]
 [0.8815 0.1185]
 [0.0916 0.9084]
 [0.1221 0.8779]
 [0.     1.    ]
 [0.6413 0.3587]
 [0.3824 0.6176]
 [0.2024 0.7976]]
```

Figure 6: $\boldsymbol{Y^t}$

To show the convergence of label convergence algorithm, we consider more general case as the following statement.

Let $(x_1, y_1), ..., (x_l, y_l)$ be labeled data, where y_i takes value in $\{1, .., C\}$, which indicates that x_i lies in Class y_i. Also, we have u unlabeled data $(x_{l+1}, y_{l+1}), ..., (x_{l+u}, y_{l+u})$, where y_j is an unknown value, which lies in $\{1, ..., C\}$.

We can construct the transition matrix $\boldsymbol{T}$,

$$\boldsymbol{T}_{i,j} = \frac{\tilde{\boldsymbol{W}}_{i,j}}{\sum_{k=1}^{l+u} \tilde{\boldsymbol{W}}_{k,j}}$$

, where $\tilde{\boldsymbol{W}}_{i,j} > 0, \quad i, j = 1, ..., l + u$.

Also, we define the label matrix sequence

$$\boldsymbol{Y}^t \in \mathbb{R}^{(l+u)\times C} \quad t = 0, 1, ...$$

where the i_{th} row of $\boldsymbol{Y}^t$ means the probability distribution of node x_i at time t. $\boldsymbol{Y}^t_{i,j}$ means the probability that the node x_i lies in Class j at time t. For $\boldsymbol{Y}^0$, we clamp the first l rows as following,

$$\boldsymbol{Y}^t_{i,j} = \mathbb{1}\{y_i = j\} \quad i = 1, ..., l, j = 1, ..., C$$

, which indicates that x_i must lies in Class y_i. For the other rows, we initialize

$$\boldsymbol{Y}^t_{i,j} = \frac{1}{C} \quad i = l+1, ..., l+u, j = 1, ..., C$$

We execute by the algorithm below.

Algorithm 2 Generalized Label Propagation Algorithm

Input label matrix $\boldsymbol{Y}^0$ and transition matrix $\boldsymbol{T}$
Output $\boldsymbol{Y}^*$

procedure GENERALIZE LABEL PROPAGATION($\boldsymbol{Y}^0$, $\boldsymbol{T}$, $t = 0$)
  repeat
    t = t + 1
    $\boldsymbol{Y}^t = \boldsymbol{T}\boldsymbol{Y}^{t-1}$ ▷ Random walk to its neighbor
    $\boldsymbol{Y}^t_{i,j} = \boldsymbol{Y}^t_{i,j} / \sum_{k=1}^{C} \boldsymbol{Y}^t_{i,k} \quad i = 1, ...l+u, j = 1, ..., C$ ▷ Normalize the probability distribution
    $\boldsymbol{Y}^t_{i,j} = \mathbb{1}\{y_i = j\} \quad i = 1, ..., l, j = 1, ..., C$ ▷ Clamp the labeled data
  until $\boldsymbol{Y}^t$ converges to $\boldsymbol{Y}^*$
end procedure

Once we get $\boldsymbol{Y}^*$, we can conclude that the unlabeled data (x_i, y_i) belong to Class y_i, where

$$y_i = \arg\max_{\mathrm{J}} \ \boldsymbol{Y}^*_{i,j}$$

Next, we want to calculate $\boldsymbol{Y}^*$ analytically.

3. Please show that the line 4 and 5 in Algorithm 2 can be combined as

$$\boldsymbol{Y}^t = \overline{\boldsymbol{T}}\boldsymbol{Y}^{t-1}$$

, where $\overline{\boldsymbol{T}}_{i,j} = \boldsymbol{T}_{i,j} / \sum_{k=1}^{l+u} \boldsymbol{T}_{i,k}, i, j = 1, ..., l+u$.

Solution:

$$\overline{\boldsymbol{T}}_{i,j}\boldsymbol{Y}^{t-1} = \sum_{r=1}^{l+u} \boldsymbol{T}_{i,r}\boldsymbol{Y}^{t-1}_{r,j} / \sum_{k=1}^{l+u} \boldsymbol{T}_{i,k}$$

After executing line 4 and 5,

$$\boldsymbol{Y}^t_{i,j} = \sum_{r=1}^{l+u} \boldsymbol{T}_{i,r}\boldsymbol{Y}^{t-1}_{r,j} / \sum_{k=1}^{C}\sum_{r=1}^{l+u} \boldsymbol{T}_{i,r}\boldsymbol{Y}^{t-1}_{r,k} = \sum_{r=1}^{l+u} \boldsymbol{T}_{i,r}\boldsymbol{Y}^{t-1}_{r,j} / \sum_{r=1}^{l+u}\sum_{k=1}^{C} \boldsymbol{Y}^{t-1}_{r,k}\boldsymbol{T}_{i,r}$$

$\sum_{k=1}^{C} \boldsymbol{Y}^{t-1}_{r,k} = 1$ for all r because the r_{th} row $\boldsymbol{Y}^{t-1}_{r}$ represents the probability distribution of x_r. Hence,

$$\boldsymbol{Y}^t_{i,j} = \sum_{r=1}^{l+u} \boldsymbol{T}_{i,r}\boldsymbol{Y}^{t-1}_{r,j} / \sum_{r=1}^{l+u} \boldsymbol{T}_{i,r} = \sum_{r=1}^{l+u} \boldsymbol{T}_{i,r}\boldsymbol{Y}^{t-1}_{r,j} / \sum_{k=1}^{l+u} \boldsymbol{T}_{i,k}$$

, which has the same form as $\overline{\boldsymbol{T}}_{i,j}\boldsymbol{Y}^{t-1}$. Hence we get $\boldsymbol{Y}^t = \overline{\boldsymbol{T}}\boldsymbol{Y}^{t-1}$.

We split $\overline{\boldsymbol{T}}$ to $\begin{bmatrix} \overline{\boldsymbol{T}}_{ll} & \overline{\boldsymbol{T}}_{lu} \\ \overline{\boldsymbol{T}}_{ul} & \overline{\boldsymbol{T}}_{uu} \end{bmatrix}$, where $\overline{\boldsymbol{T}}_{ll} \in \mathbb{R}^{l\times l}, \overline{\boldsymbol{T}}_{uu} \in \mathbb{R}^{u\times u}$. Also, we split $\boldsymbol{Y}^t$ to $\begin{bmatrix} \boldsymbol{Y}^t_L \\ \boldsymbol{Y}^t_U \end{bmatrix}$, where $\boldsymbol{Y}^t_L \in \mathbb{R}^l, \boldsymbol{Y}^t_U \in \mathbb{R}^u$

4. Please show that after an iteration,

$$\boldsymbol{Y}^t_U = \overline{\boldsymbol{T}}_{uu}\boldsymbol{Y}^{t-1}_U + \overline{\boldsymbol{T}}_{ul}\boldsymbol{Y}^{t-1}_L, t = 1, 2, ...,$$

$$\boldsymbol{Y}^t_L = \boldsymbol{Y}^{t-1}_L, t = 1, 2, ...$$

Solution:

$$\boldsymbol{Y}^t_{U\,i,j} = \boldsymbol{Y}^t_{l+i,j} = \sum_{r=1}^{l+u} \overline{\boldsymbol{T}}_{l+i,r}\boldsymbol{Y}^{t-1}_{r,j} = \sum_{r=1}^{l} \overline{\boldsymbol{T}}_{l+i,r}\boldsymbol{Y}^{t-1}_{r,j} + \sum_{r=l+1}^{l+u} \overline{\boldsymbol{T}}_{i,r}\boldsymbol{Y}^{t-1}_{r,j}$$

$$= \sum_{r=1}^{l} \overline{\boldsymbol{T}}_{\boldsymbol{ul}\,i,r}\boldsymbol{Y}_{\boldsymbol{L}}{}^{t-1}_{r,j} + \sum_{r=1}^{u} \overline{\boldsymbol{T}}_{\boldsymbol{uu}\,i,r}\boldsymbol{Y}_{\boldsymbol{U}}{}^{t-1}_{r,j} = \sum_{r=1}^{u} \overline{\boldsymbol{T}}_{\boldsymbol{uu}\,i,r}\boldsymbol{Y}_{\boldsymbol{U}}{}^{t-1}_{r,j} + \sum_{r=1}^{l} \overline{\boldsymbol{T}}_{\boldsymbol{ul}\,i,r}\boldsymbol{Y}_{\boldsymbol{L}}{}^{t-1}_{r,j} = \overline{\boldsymbol{T}}_{\boldsymbol{uu}}\boldsymbol{Y}_{\boldsymbol{U}}{}^{t-1} + \overline{\boldsymbol{T}}_{\boldsymbol{ul}}\boldsymbol{Y}_{\boldsymbol{L}}{}^{t-1}$$

. On the other hand, because we clamp $\boldsymbol{Y}^t_L$ at the end of the loop, hence $\boldsymbol{Y}^t_L = \boldsymbol{Y}^{t-1}_L$.

5. From the above result, we let $\boldsymbol{Y}_L = \boldsymbol{Y}^t_L$ for any t. Show that for $t \geq 1$

$$\boldsymbol{Y}^t_U = \overline{\boldsymbol{T}}^n_{uu}\boldsymbol{Y}^0_U + \sum_{i=1}^{n} \overline{\boldsymbol{T}}^{i-1}_{uu}\overline{\boldsymbol{T}}_{ul}\boldsymbol{Y}_L$$

Solution: We show that for any t, $\boldsymbol{Y}^t_U = \overline{\boldsymbol{T}}^n_{uu}\boldsymbol{Y}^0 + \sum_{i=1}^{n} \overline{\boldsymbol{T}}^{i-1}_{uu}\overline{\boldsymbol{T}}_{ul}\boldsymbol{Y}_L$ by mathematical induction.
For $t = 1$, $\boldsymbol{Y}_U = \overline{\boldsymbol{T}}_{uu}\boldsymbol{Y}^0 + \overline{\boldsymbol{T}}_{ul}\boldsymbol{Y}_L$
Assume when $t = k$, $\boldsymbol{Y}^k_U = \overline{\boldsymbol{T}}^k_{uu}\boldsymbol{Y}^0 + \sum_{i=1}^{k} \overline{\boldsymbol{T}}^{i-1}_{uu}\overline{\boldsymbol{T}}_{ul}\boldsymbol{Y}_L$,
Then for $t = k+1$, $\boldsymbol{Y}^{k+1}_U = \overline{\boldsymbol{T}}_{\boldsymbol{uu}}\boldsymbol{Y}_{\boldsymbol{U}}{}^k + \overline{\boldsymbol{T}}_{\boldsymbol{ul}}\boldsymbol{Y}_{\boldsymbol{L}}{}^k = \overline{\boldsymbol{T}}_{\boldsymbol{uu}}\overline{\boldsymbol{T}}^k_{uu}\boldsymbol{Y}^0 + \overline{\boldsymbol{T}}_{\boldsymbol{uu}}\sum_{i=1}^{k} \overline{\boldsymbol{T}}^{i-1}_{uu}\overline{\boldsymbol{T}}_{ul}\boldsymbol{Y}_L + \overline{\boldsymbol{T}}_{\boldsymbol{ul}}\boldsymbol{Y}_{\boldsymbol{L}} = \overline{\boldsymbol{T}}^{k+1}_{uu}\boldsymbol{Y}^0 + \sum_{i=1}^{k+1} \overline{\boldsymbol{T}}^{i-1}_{uu}\overline{\boldsymbol{T}}_{ul}\boldsymbol{Y}_L$. Hence by mathematical induction, we show that for $N \geq 1$, $\boldsymbol{Y}^t_U = \overline{\boldsymbol{T}}^n_{uu}\boldsymbol{Y}^0 + \sum_{i=1}^{n} \overline{\boldsymbol{T}}^{i-1}_{uu}\overline{\boldsymbol{T}}_{ul}\boldsymbol{Y}_L$.

6. Please show that $\sum_{j=1}^{u} \overline{\boldsymbol{T}}_{\boldsymbol{uu}i,j} = \gamma_i$, where $0 < \gamma_i < 1$, for $i = 1, ..., u$. Use the fact to derive $\sum_{j=1}^{u} \overline{\boldsymbol{T}}^{\boldsymbol{n}}_{\boldsymbol{uu}i,j} \leq \gamma^n$, for $i = 1, ..., u$, where $\gamma = \max_{i=1,...,u} \gamma_i$. Last, derive $\lim_{n\to\infty} \overline{\boldsymbol{T}}^{\boldsymbol{n}}_{\boldsymbol{uu}} = \boldsymbol{O}$.

Solution: Because $\boldsymbol{W}_{i,j} > 0, \overline{\boldsymbol{T}}_{i,j} > 0$, $\sum_{j=1}^{u} \overline{\boldsymbol{T}}_{\boldsymbol{uu}i,j} > 0$.
Also, $\sum_{j=1}^{u} \overline{\boldsymbol{T}}_{\boldsymbol{uu}i,j} = \sum_{j=1}^{l+u} \overline{\boldsymbol{T}}_{l+i,j} - \sum_{j=1}^{l} \overline{\boldsymbol{T}}_{l+i,j}$, where $\sum_{j=1}^{l+u} \overline{\boldsymbol{T}}_{l+i,j} = \sum_{j=1}^{l+u} \boldsymbol{T}_{l+i,j} / \sum_{j=1}^{l+u} \boldsymbol{T}_{l+i,j} = 1$, $\sum_{j=1}^{l} \overline{\boldsymbol{T}}_{l+i,j} > 0$. Hence, $\sum_{j=1}^{l} \overline{\boldsymbol{T}}_{\boldsymbol{uu}i,j} < 1$.
Combined the above results, we get $0 < \gamma_i < 1$. Also $0 < \gamma < 1$.
Next, we will show that $\sum_{j=1}^{u} \overline{\boldsymbol{T}}^{\boldsymbol{n}}_{\boldsymbol{uu}i,j} \leq \gamma^n$ by mathematical induction.
When $n = 1$, $\sum_{j=1}^{u} \overline{\boldsymbol{T}}_{\boldsymbol{uu}i,j} \leq \gamma$ holds.
Assume for $n = k$, $\sum_{j=1}^{u} \overline{\boldsymbol{T}}^{\boldsymbol{k}}_{\boldsymbol{uu}i,j} \leq \gamma^k$
Then for $n = k+1$, $\sum_{j=1}^{u} \overline{\boldsymbol{T}}^{\boldsymbol{k+1}}_{\boldsymbol{uu}\ i,j} = \sum_{j=1}^{u} \sum_{p=1}^{u} \overline{\boldsymbol{T}}^{\boldsymbol{k}}_{\boldsymbol{uu}i,p} \overline{\boldsymbol{T}}_{\boldsymbol{uu}p,j} = \sum_{p=1}^{u} \sum_{j=1}^{u} \overline{\boldsymbol{T}}_{\boldsymbol{uu}p,j} \overline{\boldsymbol{T}}^{\boldsymbol{k}}_{\boldsymbol{uu}i,p} \leq \sum_{p=1}^{u} \gamma \overline{\boldsymbol{T}}^{\ k}_{\boldsymbol{uu}i,p} = \gamma \sum_{p=1}^{u} \overline{\boldsymbol{T}}^{\ k}_{\boldsymbol{uu}i,p} \leq \gamma\gamma^k = \gamma^{k+1}$. Hence, by mathematical induction, we have $\sum_{j=1}^{u} \overline{\boldsymbol{T}}^{\boldsymbol{n}}_{\boldsymbol{uu}i,j} \leq \gamma^n$.
Last, $\lim_{n\to\infty} \sum_{j=1}^{u} \overline{\boldsymbol{T}}^{\boldsymbol{n}}_{\boldsymbol{uu}i,j} = 0$, where every element in $\lim_{n\to\infty} \overline{\boldsymbol{T}}^{\boldsymbol{n}}_{\boldsymbol{uu}} \geq 0$ since $\overline{\boldsymbol{T}}_{\boldsymbol{uu}ij} \geq 0$ for every $1 \leq i, j \leq u$. Hence, all the element in $\lim_{n\to\infty} \overline{\boldsymbol{T}}^{\boldsymbol{n}}_{\boldsymbol{uu}} = 0 \Rightarrow \lim_{n\to\infty} \overline{\boldsymbol{T}}^{\boldsymbol{n}}_{\boldsymbol{uu}} = \boldsymbol{O}$.

7. Define $\boldsymbol{S}_n = \boldsymbol{I} + \overline{\boldsymbol{T}}_{uu} + \overline{\boldsymbol{T}}^2_{uu} + ... + \overline{\boldsymbol{T}}^{n-1}_{uu}$, $\boldsymbol{S}_n(\boldsymbol{I} - \overline{\boldsymbol{T}}_{uu}) = \boldsymbol{I} - \overline{\boldsymbol{T}}^n_{uu}$. Use the fact to show that $\lim_{n\to\infty} \boldsymbol{S}_n = (\boldsymbol{I} - \overline{\boldsymbol{T}}_{uu})^{-1}$. Combined all the result above, please show that $\boldsymbol{Y}^*_U = (\boldsymbol{I} - \overline{\boldsymbol{T}}_{uu}))^{-1} \overline{\boldsymbol{T}}_{ul} \boldsymbol{Y}_L$. Hence, $\boldsymbol{Y}^* = \begin{bmatrix} \boldsymbol{Y}_L \\ \boldsymbol{Y}^*_U \end{bmatrix}$, which can be obtained analytically regardless of the initial value $\boldsymbol{Y}^0$.

Solution: Because $\lim_{n\to\infty} \overline{\boldsymbol{T}}^{\boldsymbol{n}}_{\boldsymbol{uu}} = \boldsymbol{O}$, $\lim_{n\to\infty} \boldsymbol{S}_n(\boldsymbol{I} - \overline{\boldsymbol{T}}_{uu}) = I$. Hence, $\lim_{n\to\infty} \boldsymbol{S}_n = (\boldsymbol{I} - \overline{\boldsymbol{T}}_{uu})^{-1}$.
$\boldsymbol{Y}^*_U = \lim_{n\to\infty} \boldsymbol{Y}^n_U = \lim_{n\to\infty} \overline{\boldsymbol{T}}^n_{uu} \boldsymbol{Y}^0 + \sum_{i=1}^{n} \overline{\boldsymbol{T}}^{i-1}_{uu} \overline{\boldsymbol{T}}_{ul} \boldsymbol{Y}_L = \boldsymbol{0} + \lim_{n\to\infty} \boldsymbol{S}_n \overline{\boldsymbol{T}}_{ul} \boldsymbol{Y}_L = (\boldsymbol{I} - \overline{\boldsymbol{T}}_{uu}))^{-1} \overline{\boldsymbol{T}}_{ul} \boldsymbol{Y}_L$.

8. Please calculate the analytical solution $\boldsymbol{Y}^*$ on the toy example above. Does the solution correspond to the iteration solution $\boldsymbol{Y}^t$?

Solution: $\boldsymbol{Y}^*$ is shown below, which is identical to the iteration solution $\boldsymbol{Y}^t$ we have seen previously.

```
[[1.     0.    ]
 [0.8296 0.1704]
 [0.7938 0.2062]
 [0.8815 0.1185]
 [0.0916 0.9084]
 [0.1221 0.8779]
 [0.     1.    ]
 [0.6413 0.3587]
 [0.3824 0.6176]
 [0.2024 0.7976]]
```

Figure 7: $\boldsymbol{Y}^*$

Version Description

1. First Edition: Finish Problem Answer 1 to 5

HW4 Problem 4 Discussion

Yuan-Chia, Chang

December 2023

Original Problem

Given N samples $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^m$ as well as their labels $y_1, \ldots, y_N \in \{0, 1, \ldots, K\}$. Consider the generative model where each sample $\mathbf{x}_i$ is generated independently according to Gaussian mixture model that depends on the label y_i, as represented by random variable

$$X_i \sim \begin{cases} \sum_{j=1}^{K} \pi_j \mathcal{N}\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) & , \ if \ y_i = 0 \\ \mathcal{N}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) & , \ if \ y_i = k \neq 0 \end{cases}$$

where $\pi_1 + \ldots + \pi_K = 1$, and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, with probability density function

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^m |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

We would like to apply Expectation Maximization algorithm to find the maximum likelihood estimation of parameters $\theta = \{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^{K}$.

1. Please write down the E-step and M-step and show that the parameters are updated from $\theta^{(t)} = \left\{\left(\pi_k^{(t)}, \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)\right\}_{k=1}^{K}$ to $\theta^{(t+1)} = \left\{\left(\pi_k^{(t+1)}, \boldsymbol{\mu}_k^{(t+1)}, \boldsymbol{\Sigma}_k^{(t+1)}\right)\right\}_{k=1}^{K}$ in the following form:

$$\pi_k^{(t+1)} = \frac{\sum_{i:y_i=0} \delta_{ik}^{(t)}}{\sum_{i:y_i=0} 1}$$

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i:y_i=k} \mathbf{x}_i + \sum_{i:y_i=0} \delta_{ik}^{(t)} \mathbf{x}_i}{N_k + \sum_{i:y_i=0} \delta_{ik}^{(t)}}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i:y_i=k} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T + \sum_{i:y_i=0} \delta_{ik}^{(t)} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T}{N_k + \sum_{i:y_i=0} \delta_{ik}^{(t)}}$$

where $N_k = \sum_{i:y_i=k} 1$ is the number of samples in class k. Please show your derivations.

2. What is the closed form expression of $\delta_{ik}^{(t)}$? Please show your derivations.

Solution: Given data set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, the likelihood and log-likelihood functions are given by

$$p(S; \theta) = \left(\prod_{i:y_i=0} \sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)\right) \cdot \prod_{k=1}^{K} \prod_{i:y_i=k} \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right),$$

$$\log p(\mathcal{X}; \theta) = \sum_{i:y_i=0} \log\left(\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)\right) + \sum_{k=1}^{K} \sum_{i:y_i=k} \log \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$$

Denote latent variable $z_i \in \{1, \ldots, K\}$ indicating which Gaussian distribution x_i is drawn from.

- Expectation Step (E-step): Compute

$$Q\left(\theta \middle| \theta^{(t)}\right) = \sum_{i=1}^{N} \mathbb{E}_{z_i|\mathbf{x}_i;y_i;\theta^{(t)}}\left[\log p\left(\mathbf{x}_i, y_i, z_i; \theta\right)\right]$$

Posterior prob. dist. of latent variables z_i based on current parameters $\theta^{(t)}$:

If $y_i = 0$, then

$$\delta_{ik}^{(t)} = \mathbb{P}\left[z_i = k \middle| \mathbf{x}_i; y_i = 0; \theta^{(t)}\right] = \frac{p\left(\mathbf{x}_i, y_i = 0, z_i = k; \theta^{(t)}\right)}{\sum_{j=1}^{K} p\left(\mathbf{x}_i, y_i = 0, z_i = j; \theta^{(t)}\right)} = \frac{\pi_k^{(t)} \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)}{\sum_{j=1}^{K} \pi_j^{(t)} \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)}\right)}$$

If $y_i = l \neq 0$, then

$$\delta_{ik}^{(t)} = \mathbb{P}\left[z_i = k \middle| \mathbf{x}_i; y_i = l; \theta^{(t)}\right] = \begin{cases} 1 & , \ if \ l = k \\ 0 & , \ if \ l \neq k \end{cases}$$

- Log-likelihood of parameter θ given data (x_i, y_i) and latent variable z_i:

If $y_i = 0$, then

$$\log p\left(\mathbf{x}_i, y_i = 0, z_i = k; \theta\right) = \log\left(\frac{\pi_k}{\sqrt{(2\pi)^m \left|\boldsymbol{\Sigma}_k\right|}}\right) - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)$$

If $y_i = l \neq 0$, then

$$log(\mathbf{x}_i, y_i = k, z_i = l; \theta) = \begin{cases} \log(\frac{1}{\sqrt{(2\pi)^m |\Sigma_k|}}) - \frac{1}{2}(\mathbf{x}_1 - \mu_k)^T \Sigma_k^{-1}(\mathbf{x}_1 - \mu_k) & l = k, \\ log(0) & l \neq k. \end{cases}$$

Hence

$$Q\left(\theta \middle| \theta^{(t)}\right) = \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{ik}^{(t)} \left\{ \log\left(\frac{\pi_k^{\mathbb{1}(y_i=0)}}{\sqrt{(2\pi)^m \left|\boldsymbol{\Sigma}_k\right|}}\right) - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k) \right\}$$

- Maximization Step (M-step): Choose

$$\theta^{(t+1)} = \arg\max_{\theta \in \Theta} Q\left(\theta \middle| \theta^{(t)}\right)$$

Note that $Q\left(\theta \middle| \theta^{(t)}\right)$ takes exactly the same form as in the unsupervised GMM scenario, hence the optimal solution $\theta^{(t+1)}$ is specified by

$$\pi_k^{(t+1)} = \frac{\sum_{i:y_i=0} \delta_{ik}^{(t)}}{\sum_{i:y_i=0} 1}$$

$$\boldsymbol{\mu}^{(t+1)} = \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)} \mathbf{x}_i}{\sum_{i=1}^{N} \delta_{ik}^{(t)}} = \frac{\sum_{i:y_i=k} \mathbf{x}_i + \sum_{i:y_i=0} \delta_{ik}^{(t)} \mathbf{x}_i}{N_k + \sum_{i:y_i=0} \delta_{ik}^{(t)}}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T}{\sum_{i=1}^{N} \delta_{ik}^{(t)}}$$

$$= \frac{\sum_{i:y_i=k} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T + \sum_{i:y_i=0} \delta_{ik}^{(t)} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T}{N_k + \sum_{i:y_i=0} \delta_{ik}^{(t)}}$$

Motivation of the revision

The updated prior of each gaussian distribution

$$\pi_k^{(t+1)} = \frac{\sum_{i:y_i=0} \delta_{ik}^{(t)}}{\sum_{i:y_i=0} 1}$$

contradicts to the result shown in the "semi-supervised learning"slide p.10. The reason is the difference between the semi-supervised learning model.

Revised Problem

Given N samples $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^m$ as well as their labels $y_1, \ldots, y_N \in \{0, 1, \ldots, K\}$. For the data with $y_i = l \neq 0$, it belongs to the specific normal distribution. For other data with $y_i = 0$, consider the generative model where each sample $\mathbf{x}_i$ is generated independently according to Gaussian mixture model.

For $y_i = l \neq 0$,

$$p(\mathbf{x}_i, y_i = l|\theta) = p(\mathbf{x}_i|y_i = l; \theta)p(y_i = l|\theta) = \pi_i \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$$

For $y_i = 0$,

$$p(\mathbf{x_i}, y_i = 0|\theta) = \sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j).$$

where $\pi_1 + \ldots + \pi_K = 1$, and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, with probability density function

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^m |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

We would like to apply Expectation Maximization algorithm to find the maximum likelihood estimation of parameters $\theta = \{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^K$.

1. Please write down the E-step and M-step and show that the parameters are updated from $\theta^{(t)} = \left\{\left(\pi_k^{(t)}, \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)\right\}_{k=1}^K$ to $\theta^{(t+1)} = \left\{\left(\pi_k^{(t+1)}, \boldsymbol{\mu}_k^{(t+1)}, \boldsymbol{\Sigma}_k^{(t+1)}\right)\right\}_{k=1}^K$ in the following form:

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)}}{N}$$

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i:y_i=k} \mathbf{x}_i + \sum_{i:y_i=0} \delta_{ik}^{(t)} \mathbf{x}_i}{N_k + \sum_{i:y_i=0} \delta_{ik}^{(t)}}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i:y_i=k} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T + \sum_{i:y_i=0} \delta_{ik}^{(t)} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T}{N_k + \sum_{i:y_i=0} \delta_{ik}^{(t)}}$$

where $N_k = \sum_{i:y_i=k} 1$ is the number of samples in class k. Please show your derivations.

2. What is the closed form expression of $\delta_{ik}^{(t)}$? Please show your derivations.

Solution: Given data set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, the likelihood and log-likelihood functions are given by

$$p(S; \theta) = \left(\prod_{i:y_i=0} \sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\right) \cdot \prod_{k=1}^{K} \prod_{i:y_i=k} \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

$$\log p(\mathcal{X}; \theta) = \sum_{i:y_i=0} \log\left(\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\right) + \sum_{k=1}^{K} \sum_{i:y_i=k} \pi_k \log \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Denote latent variable $z_i \in \{1, \ldots, K\}$ indicating which Gaussian distribution x_i is drawn from.

- Expectation Step (E-step): Compute

$$Q\left(\theta \middle| \theta^{(t)}\right) = \sum_{i=1}^{N} \mathbb{E}_{z_i|\mathbf{x}_i;y_i;\theta^{(t)}}\left[\log p\left(\mathbf{x}_i, y_i, z_i; \theta\right)\right]$$

Posterior prob. dist. of latent variables z_i based on current parameters $\theta^{(t)}$:

If $y_i = 0$, then

$$\delta_{ik}^{(t)} = \mathbb{P}\left[z_i = k \middle| \mathbf{x}_i; y_i = 0; \theta^{(t)}\right] = \frac{p\left(\mathbf{x}_i, y_i = 0, z_i = k; \theta^{(t)}\right)}{\sum_{j=1}^{K} p\left(\mathbf{x}_i, y_i = 0, z_i = j; \theta^{(t)}\right)} = \frac{\pi_k^{(t)} \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)}{\sum_{j=1}^{K} \pi_j^{(t)} \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)}\right)}$$

If $y_i = l \neq 0$, then

$$\delta_{ik}^{(t)} = \mathbb{P}\left[z_i = k \middle| \mathbf{x}_i; y_i = l; \theta^{(t)}\right] = \begin{cases} 1 & , \ if \ l = k \\ 0 & , \ if \ l \neq k \end{cases}$$

- Log-likelihood of parameter θ given data (x_i, y_i) and latent variable z_i:

If $y_i = 0$, then

$$\log p\left(\mathbf{x}_i, y_i = 0, z_i = k; \theta\right) = \log\left(\frac{\pi_k}{\sqrt{(2\pi)^m \left|\boldsymbol{\Sigma}_k\right|}}\right) - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)$$

If $y_i = l \neq 0$, then

$$log(\mathbf{x}_i, y_i = k, z_i = l; \theta) = \begin{cases} \log(\frac{\pi_k}{\sqrt{(2\pi)^m |\Sigma_k|}}) - \frac{1}{2}(\mathbf{x}_1 - \mu_k)^T \Sigma_k^{-1}(\mathbf{x}_1 - \mu_k), & l = k. \\ log(0) & l \neq k. \end{cases}$$

Hence

$$Q\left(\theta \middle| \theta^{(t)}\right) = \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{ik}^{(t)} \left\{\log\left(\frac{\pi_k}{\sqrt{(2\pi)^m \left|\boldsymbol{\Sigma}_k\right|}}\right) - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\right\}$$

- Maximization Step (M-step): Choose

$$\theta^{(t+1)} = \arg\max_{\theta \in \Theta} Q\left(\theta \middle| \theta^{(t)}\right)$$

Note that $Q\left(\theta \middle| \theta^{(t)}\right)$ takes exactly the same form as in the unsupervised GMM scenario, hence the optimal solution $\theta^{(t+1)}$ is specified by

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)}}{N}$$

$$\boldsymbol{\mu}^{(t+1)} = \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)} \mathbf{x}_i}{\sum_{i=1}^{N} \delta_{ik}^{(t)}} = \frac{\sum_{i:y_i=k} \mathbf{x}_i + \sum_{i:y_i=0} \delta_{ik}^{(t)} \mathbf{x}_i}{N_k + \sum_{i:y_i=0} \delta_{ik}^{(t)}}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i=1}^{N} \delta_{ik}^{(t)} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T}{\sum_{i=1}^{N} \delta_{ik}^{(t)}}$$

$$= \frac{\sum_{i:y_i=k} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T + \sum_{i:y_i=0} \delta_{ik}^{(t)} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^T}{N_k + \sum_{i:y_i=0} \delta_{ik}^{(t)}}$$

Conclusion

We utilize the chain rule to consider the joint probability $p(\mathbf{x}_i, y_i|\theta) = p(\mathbf{x}_i|y_i, \theta)p(y_i|\theta)$ for labeled data in the revised problem. The only difference between the original model and the revised model is that we find the connection between the prior π_i and the labeled data, which shows that the updated prior $\pi_i^{(t+1)}$ relates to labeled data. Also, the interpretation of the revised prior is the average posterior probability to the specific class, which follows our intuition.

Reference

https://pages.cs.wisc.edu/ jerryzhu/pub/sslchicago09.pdf