HW3 Handwritten Assignment

Lecturor: Pei-Yuan Wu
TAs: Po-Yang Hsieh, Le-Rong Hsu

October 2024

Problem 1 (Principle Component Analysis) (0.5%)

Given 10 samples in 3D:

$$(1,2,3),(4,8,5),(3,12,9),(1,8,5),(5,14,2),(7,4,1),(9,8,9),(3,8,1),(11,5,6),(10,11,7)$$

1. What are the principal axes? Please write down your derivation or provide your code of computation at HW3 math problems (NTU COOL) through "add another file".
2. Please compute the principal components for each sample and either write down your derivation or provide your code of computation in report.
3. What is the average reconstruction error if reduce dimension to 2D? Here the reconstruction error is defined as the squared loss.

Problem 2 (Reparameterization Trick)(0.5%)

By working on this problem, we hope that students will understand VAEs from the perspective of deep latent variable models, as a supplement to the course materials. [1]

Let's recall the problem setting of the VAE. Let $p_{\mathcal{D}}$ be the data distribution and for each $\mathbf{x} \in \mathcal{D}$, $p_{\mathcal{D}}(\mathbf{x})$ is the probability of $\mathbf{x}$. A deep latent variable model, which is a generative model, represents the joint distribution $p_\theta(\mathbf{x}, \mathbf{z})$ over data and latent variables, and the likelihood of data generated by θ is given by

$$\begin{aligned} p_\theta(\mathbf{x}) &= \sum_{\mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z}) && (\mathbf{z} \text{ discrete}) \\ &= \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} && (\mathbf{z} \text{ continuous}) \end{aligned}$$

For a better understanding of this relation, please refer to the Problem 2 of HW1 as an example. By the property of the joint distribution, one has

$$p_\theta(\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{z}|\mathbf{x})}.$$

Although $p_\theta(\mathbf{x}, \mathbf{z})$ is tractable, the marginal probability $p_\theta(\mathbf{x})$ is intractable because $p_\theta(\mathbf{z}|\mathbf{x})$, the probability of latent variables conditioning on the data distribution, is typically intractable.

To deal with the intractability, VAE is invented by considering an *inference model* $q_\phi(\mathbf{z}|\mathbf{x})$, or encoder, parametrized by a neural network, denoted ϕ, to approximate $p_\theta(\mathbf{z}|\mathbf{x})$, while the generative model $p_\theta(\mathbf{x}, \mathbf{z})$ can be decomposed as

$$\begin{aligned} p_\theta(\mathbf{x}, \mathbf{z}) &= p_\theta(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z}) \\ &= p(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z}) \end{aligned}$$

where $p_\theta(\mathbf{z}) = p(\mathbf{z})$ is called the *prior distribution*, meaning that it is not conditioned on any observation, and $p_\theta(\mathbf{x}|\mathbf{z})$ is called the decoder which reconstructs data from latent variables.

[1] The idea of this problem follows from "An Introduction to Variational Autoencoders" by Diederik P. Kingma and Max Welling.

1. For $\mathbf{x} \in \mathcal{D}$, please decompose $\log p_\theta(\mathbf{x})$ into ELBO and KL divergence. You may refer to the lecture slides but don't just give the final answer.

2. Let $\mathcal{L}_{\phi,\theta}(\mathbf{x})$ denote the ELBO. To perform optimization on ϕ, we have to take gradient of $\mathcal{L}_{\phi,\theta}(\mathbf{x})$ with respect to ϕ:

$$\begin{aligned}\nabla_\phi \mathcal{L}_{\phi,\theta}(\mathbf{x}) &= \nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{z},\mathbf{x}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right] \\ &\neq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\nabla_\phi\left(\log p_\theta(\mathbf{z},\mathbf{x}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right)\right]\end{aligned}$$

   where the last equality is generally not true. What is an alternative way to perform optimization? Please refer to section 2.4 in the reference. Also, show that the gradient of ELBO with respect to ϕ becomes exchangeable and is unbiased.

Problem 3 (Laplacian Eigenmaps)(2%)

Consider an undirected connected graph G, which is shown below. We want to utilize Laplacian Eigenmaps method to reduce these 10 points to 3-dimensional space. Here, undirected graph means that edges in the graph do not have a direction, and connected graph means that there is a path from any node to any other node in the graph.

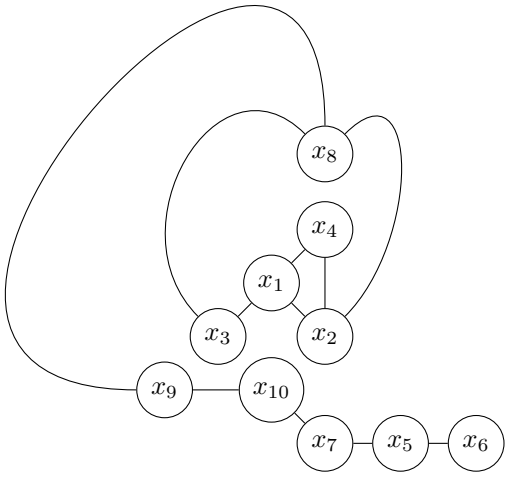

Figure 1: Problem 2 undirected connected graph G

1. (0.3%) Write down the adjacency matrix $\boldsymbol{W}$

2. (0.3%) Write down the diagonal matrix $\boldsymbol{D} = \text{diag}(d_1, ..., d_{10})$, where $d_i = \sum_{j=1}^{10} \frac{\boldsymbol{W_{ij}} + \boldsymbol{W_{ji}}}{2}$ and the Laplacian $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$.

3. (0.1%) By Neighbor Embedding Slide p.7-p.10 and programming tools (MATLAB, Python..., please hand in your code at HW3 math problems (NTU COOL) through "add another file"), solve the optimization problem
   minimize $\quad \text{Trace}(\boldsymbol{\Psi}^T \boldsymbol{L} \boldsymbol{\Psi})$
   subject to $\quad \boldsymbol{\Psi}^T \boldsymbol{D} \boldsymbol{\Psi} = \boldsymbol{I}_3$

variables $\boldsymbol{\Psi} \in \mathbb{R}^{10\times 3}$
Also, please plot the reduced points $\boldsymbol{z_1}, ..., \boldsymbol{z_{10}}$ in 3-D scatter plot.

4. (0.1%) You may find that the minimal eigenvalue of $\boldsymbol{L}$ is 0, and the corresponding eigenvector is

$$\begin{bmatrix} c \\ c \\ \vdots \\ c \end{bmatrix} \tag{1}$$

where c is a constant. Since all the points fall into a plane, the span of these points is $\mathbb{R}^2$. In order to construct $\boldsymbol{z_1}, ..., \boldsymbol{z_{10}}$ such that $\text{span}\{\boldsymbol{z_1}, ..., \boldsymbol{z_{10}}\} = \mathbb{R}^3$, we need choose the second, third, fourth smallest eigenvalue and the corresponding eigenvectors. Please plot the reduced points by the updated $\boldsymbol{z_1}, ..., \boldsymbol{z_{10}}$ in 3-D scatter plot and verify that whether $\text{Trace}(\boldsymbol{\Psi}^T \boldsymbol{L} \boldsymbol{\Psi}) = 1.098$ and $\boldsymbol{\Psi}^T \boldsymbol{D} \boldsymbol{\Psi} = \boldsymbol{I}_3$.

5. (0.3%) Show that for no matter the graph is, there is an eigenvector of $\boldsymbol{L}$

$$\begin{bmatrix} c \\ c \\ \vdots \\ c \end{bmatrix} \tag{2}$$

where c is a constant, and the corresponding eigenvalue is 0.

6. (0.3%) By Neighbor Embedding Slide p.9, please show that
$\forall \boldsymbol{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \end{bmatrix} \in \mathbb{R}^N, \boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f} = \frac{1}{2} \sum_{1\le i,j\le N} w_{ij}(f_i - f_j)^2.$

7. (0.3%) Show that if $\boldsymbol{f}$ is an eigenvector of $\boldsymbol{L}$ which corresponds to eigenvalue 0, then $\boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f} = 0$.

8. (0.3%) Show that if the graph is connected, the second smallest eigenvalue of $\boldsymbol{L}$ will be nonzero.

Problem 4 (Expectation Maximization Interpretation behind Semi-Supervised Learning)(1.5%)

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

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

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.

Problem 5 (EM for Mixture of Linear Models)(1.5%)

Consider the generative model parameterized by $\theta = ((\pi_k, \mathbf{w}_k, \rho_k))_{k=1}^K$, where $\mathbf{w}_k \in \mathbb{R}^m, \rho_k > 0, \pi_k > 0$ for each k, and that $\sum_{k=1}^K \pi_k = 1$, so that the probabilistic density function (PDF) of generating a real-valued number y given $\mathbf{x} \in \mathbb{R}^m$ is

$$p(y \mid \mathbf{x}; \theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(y; f_k(\mathbf{x}; \theta), \rho_k\right)$$

where $f_k(\mathbf{x}; \theta) = \mathbf{w}_k^\top \mathbf{x}$ and

$$\mathcal{N}(y; \mu, \rho) = \frac{1}{\sqrt{2\pi\rho}} \exp\left(-\frac{(y-\mu)^2}{2\rho}\right)$$

denotes the PDF of Gaussian distribution with mean μ and variance ρ. That is, $p(\cdot \mid \mathbf{x}; \theta)$ describes how the output is generated based on a mixture of linear models $f_k(\cdot; \theta)$ with mixing coefficient π_k and uncertainty ρ_k. Suppose we observe N data inputs $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^m$ and their corresponding outputs $y_1, \ldots, y_N$. The maximum likelihood estimation is given by

$$\theta^{\text{opt}} = \operatorname*{argmin}_\theta \sum_{i=1}^{N} \log p\left(y_i \mid \mathbf{x}_i; \theta\right)$$

Derive the E-step and M-step equations of the EM algorithm for optimizing the linear model coefficients $\mathbf{w}_k$, the mixing coefficients π_k, as well as the variances ρ_k.