In the Name of God

Deep Generative Models (25120)

Problem Set 2

Fall Semester 1404-05

Department of Electrical Engineering

Sharif University of Technology

Instructor: Dr. S. Amini

Due on: According to CW

1 Conditional Variational Autoencoder (CVAE)

In this question, we examine the CVAE, which is the conditional version of VAE. We aim to model the distribution $p_{\text{data}}(x|y)$, and the dataset also consists of pairs $\{(x_i, y_i)\}_{i=1}^{N}$, where x and y can have arbitrary dimensions. For example, x could be an image and y the description of that image in the Persian language.

1. Introduce the distance criterion between the two distributions $p_{\text{data}}(x|y)$ and $p_\theta(x|y)$. Then, provide an approximation of it such that to compute it, samples from $p_{\text{data}}(x, y)$ and the likelihood $p_\theta(x|y)$ are required.

2. Now, assume there exists a latent variable z such that $p_\theta(x|y, z)$ is much simpler than $p_\theta(x|y)$. In the case of having access to it, under this assumption, write the marginal likelihood (or Evidence) $p_\theta(x|y)$ in terms of $p_\theta(x|y, z)$.

3. If we rewrite the marginal likelihood as:

$$p_\theta(x|y) = \mathbb{E}_{z\sim q(z|x,y)}[A]$$

   What is the expression for A inside the mathematical expectation, where $q(z|x, y)$ is an arbitrary conditional distribution over Z.

4. Based on the relation obtained in the previous part, derive the $\text{ELBO}(x, y; \theta, q)$ for $\log p_\theta(x|y)$. (Note: Based on Jensen's inequality, since log is a concave function, we know that $\psi(\mathbb{E}[X]) \geq \mathbb{E}[\psi(X)]$ where $\psi(\cdot)$ is the log function.)

5. The distribution $q(z|x, y)$ that leads to the tightest lower bound (ELBO) to the logarithm of the marginal likelihood. Compute the measure of closeness (i.e., the divergence between $q(z|x, y)$ and the corresponding true posterior) in this case.

6. Write the optimization problem for approximating $q(z|x, y)$ to the best distribution (the distribution that has the tightest lower bound). Express its challenges and provide an alternative solution for this problem (you must justify your alternative solution).

7. Assume you have a CVAE model trained on pairs $\{(x, y)\}$. Now, a new data pair $(x_{\text{new}}, y_{\text{new}})$ is provided. Describe the process of obtaining $\text{ELBO}(x_{\text{new}}, y_{\text{new}}; \theta, q)$. If we want to approximate $p_\theta(x_{\text{new}}|y_{\text{new}})$ using this value, specify which likelihood-approximation method should be used.

8. Assume you have a CVAE model trained on $\{(x_i, y_i)\}$ (image–text description pairs). Now, a new text description y_{new} is provided. Describe the process of generating a sample image corresponding to this text description.

2 Cauchy–Schwarz Divergence

One of the factors that significantly affects the quality of a VAE's results is the choice of the prior distribution $p(z)$ and the posterior distribution $q(z|x)$. In most cases, the posterior is assumed to be Gaussian, but this assumption can be restrictive. If we were able to use more flexible distributions like a Gaussian Mixture Model (GMM) for $p(z)$ or $q(z|x)$, the quality of results could improve when computing the ELBO. However, the closed form of the KL divergence between two GMMs does not exist.

In this problem, we want to explore an alternative divergence measure D_{CS}, defined as:

$$D_{CS}(p\|q) = -\log \frac{\int p(x)q(x)\,dx}{\sqrt{\int p(x)^2\,dx \int q(x)^2\,dx}}$$

Although D_{CS} looks more complicated than D_{KL}, for Gaussian distributions it can actually be computed in closed form, and it satisfies the inequality:

$$D_{CS}(p\|q) \leq \min\{D_{KL}(p\|q),\, D_{KL}(q\|p)\}$$

1. For two univariate normal distributions p and q with means μ_1, μ_2 and variances σ_1^2, σ_2^2, derive $D_{CS}(p\|q)$ and verify that the inequality above holds in this case.

2. The KL divergence between two GMMs does not have a closed form and is difficult to approximate analytically. However, the Cauchy–Schwarz divergence $D_{CS}(p\|q)$, while also lacking a closed form, can be computed numerically in a stable and practical way. Motivated by this, define the following modified objective:

$$L_{CS}(x;\theta,\phi) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \lambda\, D_{CS}(q_\phi(z|x)\|p(z))$$

where λ is a real-valued coefficient.

Instead of providing the final decomposition, describe the approach you would take to rewrite L_{CS} in terms of $\log p(x)$, KL divergences, and D_{CS}. Explain the reasoning for each term and how they relate to the VAE optimization. Then, qualitatively discuss the impact of increasing λ on the model's learning dynamics and latent representations.

3 Posterior Collapse in Variational Autoencoders

1. Define the term *posterior collapse* both intuitively and mathematically in the context of VAEs. Provide at least one formal expression that characterizes this phenomenon. You may use online sources for this section, please cite your sources.

2. Show that if the decoder $p_\theta(x|z)$ is so powerful that it can model the data distribution $p_{\text{data}}(x)$ without depending on z, i.e.,

$$\exists \tilde{p}(x) \text{ such that } p_\theta(x|z) = \tilde{p}(x) \approx p_{\text{data}}(x),$$

then the optimal encoder satisfies $q_\phi(z|x) = p(z)$. Derive this using the ELBO expression.

3. Show that the ELBO can be rewritten as

$$\text{ELBO}(\theta,\phi) = \mathbb{E}_{p_{\text{data}}(x)}\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z)\|p(z)) - I_q(x;z),$$

where $q_\phi(z) = \mathbb{E}_{p_{\text{data}}(x)}[q_\phi(z|x)]$. Explain what posterior collapse corresponds to in this decomposition.

4. List at least three practical factors that can lead to posterior collapse in training VAEs, and explain briefly why each increases the likelihood of collapse.

5. Describe three strategies for preventing posterior collapse, give mathematical intuition.

4 Probabilistic Graph Forecasting with Autoregressive Decoders

We consider a temporal graph sequence denoted by $\{G_t\}_{t=1}^{T}$, where each snapshot

$$G_t = (X_t, A_t)$$

represents the state of the graph at time t. Each component is defined as follows:

- Node feature matrix: $X_t \in \mathbb{R}^{n \times d}$, where each row x_t^v corresponds to the d-dimensional feature vector of node v at time t.
- Adjacency matrix: $A_t \in \{0,1\}^{n \times n}$, representing undirected binary edges between nodes.

To capture temporal dependencies in the evolving graph, we introduce a sequence of latent variables

$$Z_{1:T} = \{Z_1, Z_2, \ldots, Z_T\}, \quad Z_t \in \mathbb{R}^m,$$

where each Z_t encodes a global latent representation of the graph state at time t.

The joint generative process is modeled in an *autoregressive latent-variable* framework, which factorizes as:

$$p_\theta(X_{1:T}, A_{1:T}, Z_{1:T}) = p(Z_1) \prod_{t=1}^{T-1} p_\theta(Z_{t+1} \mid Z_t)\, p_\theta(X_{t+1}, A_{t+1} \mid Z_t).$$

Here, $p_\theta(Z_{t+1} \mid Z_t)$ defines the temporal prior over latent states, and $p_\theta(X_{t+1}, A_{t+1} \mid Z_t)$ specifies the conditional likelihood of the next graph snapshot given the current latent.

The posterior distribution over latent variables is approximated by an amortized inference model:

$$q_\phi(Z_t \mid X_{1:t}, A_{1:t}),$$

implemented using a Graph Neural Network (GNN) that encodes the sequence of observed graph snapshots up to time t.

Finally, the decoder defines separate likelihoods for node features and adjacency matrices as:

$$p_\theta(X_{t+1} \mid Z_t) = \prod_{v=1}^{n} \mathcal{N}\big(x_{t+1}^v \mid \mu_\theta^v(Z_t), \Sigma_\theta^v(Z_t)\big),$$

$$p_\theta(A_{t+1} \mid Z_t) = \prod_{u<v} \text{Bernoulli}\big(A_{t+1}^{uv} \mid \pi_\theta^{uv}(Z_t)\big),$$

where $\mu_\theta^v(Z_t)$ and $\Sigma_\theta^v(Z_t)$ are the mean and covariance parameters for node v, and $\pi_\theta^{uv}(Z_t)$ denotes the predicted edge probability between nodes u and v.

1. Derive a variational lower bound (ELBO) on the log marginal likelihood $\log p_\theta(X_{1:T}, A_{1:T})$ using the encoder q_ϕ. Write the bound as a sum over time steps and identify reconstruction and KL terms.

2. For the Gaussian node feature and Bernoulli edge likelihoods, write the explicit reconstruction log-likelihood $\log p_\theta(X_{t+1}, A_{t+1} \mid Z_t)$.

3. If the decoder is deterministic, analyze its effect on the ELBO, multimodality, and stability.

4. Consider an autoregressive latent-variable model for temporal graphs:

$$p_\theta(X_{1:T}, A_{1:T}, Z_{1:T}) = p(Z_1) \prod_{t=1}^{T-1} p_\theta(Z_{t+1} \mid Z_t)\, p_\theta(X_{t+1}, A_{t+1} \mid Z_t),$$

where Z_t is the latent representation of the graph at time t. Explain how to generate future graphs $(X_{T+1:T+H}, A_{T+1:T+H})$ given an observed history $(X_{1:T}, A_{1:T})$, and discuss under what theoretical conditions the long-term distribution of generated graphs converges to t

5. In practice, the amortized posterior $q_\phi(Z_t \mid X_{1:t}, A_{1:t})$ may deviate from the true posterior $p_\theta(Z_t \mid X_{1:t}, A_{1:t})$, causing accumulated errors during autoregressive generation. Explain qualitatively how this *amortization gap* can lead to distribution drift in predicted graph sequences, and propose one method that could reduce this effect. Briefly relate your explanation to the ELBO optimization and its dependence on accurate inference.