CS236 - Deep Generative Models

Instructor: Stefano Ermon; Aditya Grover; Notes: Adithya Ganesh

November 2, 2018

1 Variational Autoencoder

- Observations: $\mathbf{x} \in \{0, 1\}^d$.
- Latent variables $\mathbf{z} \in \mathbb{R}^k$.
- Goal: learn a latent variable model that satisfies

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

In particular, the VAE is defined by the following generative process:

$$\begin{aligned} p(\mathbf{z}) &= \mathcal{N}(\mathbf{z}|0, I) \\ p(\mathbf{x}|\mathbf{z}) &= \mathrm{Ber}(\mathbf{x}|f_\theta(\mathbf{z})), \end{aligned}$$

where $f_\theta(\mathbf{z})$ is a neural network decoder to obtain the parameters of the d Bernoulli random variables which model the pixels in each image.

For inference, we want good values of the latent variables given observed data (that is, $p(\mathbf{z}|\mathbf{x})$.

Indeed, by Bayes' theorem, we can write

$$\begin{aligned} p(\mathbf{z}|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})} \\ &= \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{\int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\, dz}. \end{aligned}$$

We want to maximize the marginal likelihood $p_\theta(\mathbf{x})$, but the integral over all possible $\mathbf{z}$ is intractable. Therefore, we use a variational approximation to the true posterior.

We write

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mu_\phi(\mathbf{x}), \mathrm{diag}(\sigma^2_\phi(\mathbf{x}))).$$

Variational inference approximates the posterior with a family of distributions $q_\phi(\mathbf{z}|\mathbf{x})$.

To measure how well our variational posterior $q(\mathbf{z}|\mathbf{x})$ approximates the true posterior $p(\mathbf{z}|\mathbf{x})$, we can use the KL-divergence.

The optimal approximate posterior is

$$\begin{aligned} q_\phi(\mathbf{z}|\mathbf{x}) &= \mathrm{argmin}_\phi\, KL(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})) \\ &= \mathrm{argmin}_\phi \left\{ \mathbb{E}_q \left[\log q_\phi(\mathbf{z}|\mathbf{x})\right] - \mathbb{E}_q \left[\log p(\mathbf{x}, \mathbf{z})\right] + \log p(\mathbf{x}) \right\}. \end{aligned}$$

But this is impossible to compute directly, since we end up getting $p(\mathbf{x})$ in the divergence.

We then maximize the lower bound to the marginal log-likelihood:

$$\begin{aligned} \log p_\theta(\mathbf{x}) &\geq \mathrm{ELBO}(\mathbf{x}; \theta, \phi) \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \end{aligned}$$

And this ELBO is tractable, so we can optimize it.

1.1 Reparametrization trick

Instead of sampling

$$z \sim \mathcal{N}(\mu, \Sigma),$$

we can sample

$$\begin{aligned} &z = \mu + L\epsilon; \\ &\epsilon \sim \mathcal{N}(0, I); \Sigma = LL^T \end{aligned}$$

Allows for low variance estimates.

1.2 GMVAE

Same set up as vanilla VAE, except the prior is a mixture of Gaussians. That is,

$$p_\theta(\mathbf{x}) = \sum_{i=1}^{k} \frac{1}{k} \mathcal{N}(\mathbf{z}|\mu_i, \mathrm{diag}(\sigma_i^2))$$

However, the KL term cannot be computed analytically between a Gaussian and a mixture of Gaussians. We can obtain an unbiased estimator, however:

$$
\begin{aligned}
D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) &\approx \log q_\phi(\mathbf{z}^{(1)}|\mathbf{x}) - \log p_\theta(\mathbf{z}^{(1)}) \\
&= \log \mathcal{N}(\mathbf{z}^{(1)}|\mu_\phi(\mathbf{x}), \mathrm{diag}(\sigma_\phi^2(\mathbf{x}))) - \log \sum_{i=1}^{k} \frac{1}{k}\mathcal{N}(\mathbf{z}^{(1)}|\mu_i, \mathrm{diag}(\sigma_i^2)).
\end{aligned}
$$

1.3 IWVAE

The ELBO bound may be loose if $q_\phi(\mathbf{z}|\mathbf{x})$ is a poor approximation to $p_\theta(\mathbf{z}|\mathbf{x})$. For a fixed $\mathbf{x}$, the ELBO is, in expectation, the log of the unnormalized density ratio

$$
\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} = \frac{p_\theta(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})} p_\theta(\mathbf{x}),
$$

where $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$.

1. Prove that IWAE is a valid lower bound of the log-likelihood.

$$
\begin{aligned}
\log p_\theta(\mathbf{x}) &\geq \mathbb{E}_{\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(m)} \sim q_\phi(\mathbf{z}|\mathbf{x})} \left(\log \frac{1}{m} \sum_{i=1}^{m} \frac{p_\theta(\mathbf{x},\mathbf{z}^{(i)})}{q_\phi(\mathbf{z}^{(i)}|\mathbf{x})} \right) \\
&\geq \mathbb{E}_{z^{(1)} \sim q_\phi(\mathbf{z}|\mathbf{x})} \log \frac{p_\theta(\mathbf{x},\mathbf{z}^{(1)})}{q_\phi(\mathbf{z}^{(1)}|\mathbf{x})}
\end{aligned}
$$

   Jensen states that for convex functions, $\mathbb{E}f[X] \geq f\mathbb{E}[X]$. $\log$ is concave. So

1.4 Questions

- Why is the reparametrization trick lower variance? (Asked on Piazza.)