Variational Autoencoders

Let x denote the vector of all observed variables whose joint distribution $p^*(x)$ we would like to model. We will attempt to approximate this underlying distribution with a model $p_\theta(x)$ with parameters θ

$$p_\theta(x) \approx p^*(x)$$

We will include latent variables z in our model. These variables are not observed and are not part of the data, but they participate in the generative process producing the observations x. Now we have a joint distribution $p_\theta(x,z)$ and the marginal over the observed variables is

$$p_\theta(x) = \int p_\theta(x,z)\,dz \qquad (*)$$

Commonly in the factorization

$$p_\theta(x,z) = p_\theta(z)\,p_\theta(x|z),$$

the prior distribution $p_\theta(z)$ and/or $p_\theta(x|z)$ are specified.

The main difficulty here is that the integral in $(*)$ is intractable. This also makes the posterior distribution

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

intractable.

To tackle this issue let's introduce an encoder (inference model) $q_\phi(z|x)$. It has parameters ϕ which we have to optimize so that

$$q_\phi(z|x) \approx p_\theta(z|x)$$

Ex: The distribution q_ϕ can be parametrized using NN's. In that case ϕ include the weights and the biases of the NN. For example,

$$(\mu, \log\sigma) = \mathrm{EncoderNN}_\phi(x)$$

$$q_\phi(z|x) = \mathcal{N}(z; \mu, \mathrm{diag}(\sigma))$$

Variational Autoencoders

Notice that:

$$\log p_\theta(x) = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x)\right] = \mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{p_\theta(x,z)}{p_\theta(z|x)}\right]\right] =$$

$$= \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right]\right]}_{\mathcal{L}_{\theta,\phi}(x)} + \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{q_\phi(z|x)}{p_\theta(z|x)}\right]\right]}_{D_{KL}(q_\phi(z|x)\,\|\,p_\theta(z|x))}$$

The second term above is the Kullback-Liebler (KL) divergence between $q_\phi(z|x)$ and $p_\theta(z|x)$ and is non-negative:

$$D_{KL}(q_\phi(z|x)\,\|\,p_\theta(z|x)) \geq 0$$

The first term $\mathcal{L}_{\theta,\phi}(x)$ is called ELBO (evidence lower bound).

$$\mathcal{L}_{\theta,\phi}(x) = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x,z) - \log q_\phi(z|x)\right] \underset{\substack{\uparrow \\ \text{since } D_{KL}\geq 0.}}{\leq} \log p_\theta(x) \qquad (**)$$

Note that the KL divergence $D_{KL}(q_\phi(z|x)\,\|\,p_\theta(z|x))$ determines two distances:

1) The distance between the approximate and true posteriors.
2) The gap between ELBO $\mathcal{L}_{\theta,\phi}(x)$ and the marginal likelihood $\log p_\theta(x)$

Two for one: The maximization of the ELBO $\mathcal{L}_{\theta,\phi}(x)$ w.r.t. the parameters θ and ϕ, will concurrently optimize two desirable objectives:

1) It will (implicitely) maximize the marginal likelihood $p_\theta(x)$ $\Rightarrow$ the generative model will become better.

2) It will minimize the distance between the approximation $q_\phi(z|x)$ and the "true" posterior $p_\theta(z|x)$. The inference model will become better.

Gradients of the ELBO wrt the generative model parameters are easy to obtain:

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

$$= \mathbb{E}_{q_\phi(z|x)}\left[\nabla_\theta \log p_\theta(x,z)\right] \underset{\uparrow}{\simeq} \nabla_\theta \log p_\theta(x,z)$$

MC estimator; $z \sim p_\theta(z|x) \leftarrow$ random sample

Variational Autoencoders

Gradients wrt the variational parameters ϕ are more difficult

$$\nabla_\phi \mathcal{L}_{\theta,\phi}(x) = \nabla_\phi \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x,z) - \log q_\phi(z|x)]$$

Cannot push the gradient through the expectation.

The reparametrization trick:

Let's express the RV $z \sim q_\phi(z|x)$ as a smooth, bijective transformation of another RV u, given x and ϕ:

$$z = g(u, \phi, x),$$

where the distribution of the RV u is independent of x or ϕ. The expectation can be rewritten in terms of u:

$$\mathbb{E}_{q_\phi(z|x)} [f(z)] = \mathbb{E}_{p(u)} [f(g(u,\phi,x))]$$

Now the gradient operator and the expectation commute:

$$\nabla_\phi \mathbb{E}_{q_\phi(z|x)} [f(z)] = \nabla_\phi \mathbb{E}_{p(u)} [f(g(u,\phi,x))] =$$

$$= \mathbb{E}_{p(u)} [\nabla_\phi f(g(u,x,\phi))] \simeq \nabla_\phi f(g(u,x,\phi))$$

MC estimator; $u \sim p(u) \leftarrow$ random sample

Gradient of ELBO.

The ELBO can be rewritten as:

$$\mathcal{L}_{\theta,\phi}(x) = \mathbb{E}_{p(u)} [\log p_\theta(x,z) - \log q_\phi(z|x)], \quad z = g(u,\phi,x)$$

We can form a simple MC estimator $\tilde{\mathcal{L}}_{\theta,\phi}(x)$ of the individual-datapoint ELBO where we use a single noise sample $u \sim p(u)$

$$\tilde{\mathcal{L}}_{\theta,\phi}(x) = \log p_\theta(x,z) - \log q_\theta(z|x), \quad z = g(u,\phi,x).$$

We can than optimize ELBO using minibatch SGD. Notice also that the gradient is an unbiased estimator of the exact single-datapoint ELBO gradient; when averaged over the noise u, the estimated gradient matches the 'true' gradient:

$$\mathbb{E}_{p(u)} [\nabla_{\theta,\phi} \tilde{\mathcal{L}}_{\theta,\phi}(x;u)] = \nabla_{\theta,\phi} \mathcal{L}_{\theta,\phi}(x).$$

Computation of $\log q_\phi(z|x)$

We know the density $p(u)$ of the 'noise' distribution. And we have

$$\log q_\phi(z|x) = \log p(u) - \log \left| \det\left(\frac{\partial z}{\partial u}\right) \right|$$

Jacobian matrix

ex: For example, we can choose

$$u \sim N(0, I)$$

$$(\mu, \log \sigma) = \mathrm{EncoderNN}_\phi(x)$$

$$z = \mu + \sigma \odot u \qquad (= g(u, \phi, x)).$$

The Jacobian of this transformation from u to z is:

$$\frac{\partial z}{\partial u} = \mathrm{diag}(\sigma)$$

and then

$$\log \left| \det\left(\frac{\partial z}{\partial u}\right) \right| = \sum_i \log \sigma_i$$

and the posterior density is

$$\log q_\phi(z|x) = \log p(u) - \log \left| \det\left(\frac{\partial z}{\partial u}\right) \right| =$$

$$= \sum_i \log N(u_i; 0, 1) - \log \sigma_i = -\sum_i \left(\frac{1}{2}(u_i^2 + \log(2\pi)) + \log \sigma_i \right).$$

ex: Full covariance Gaussian posterior.

$$u \sim N(0, I)$$

$$z = \mu + Lu,$$

where L is a lower (or upper) triangular matrix with nonzero entries on the diagonal. Now

$$\log q_\phi(z|x) = \log p(u) - \sum_i \log |L_{ii}|$$

Note also that the covariance of z is $\Sigma = LL^T$. As usual the parameters are computed with a neural network:

$$(\mu, L) \leftarrow \mathrm{EncoderNN}_\phi(x)$$

Variational Autoencoders

Computation of $\log p_\theta(x,z)$.

Notice that

$$\log p_\theta(x,z) = \log p_\theta(x|z) + \log p_\theta(z)$$

Since $p_\theta(z) \sim N(0, I)$ we have

$$\log p_\theta(z) = -\sum_i \frac{1}{2}\left(z_i^2 + \log(2\pi)\right)$$

For the decoder we we assume a factorized Bernoulli distribution

$$p = \text{DecoderNN}_\theta(z)$$

$$\log p(x|z) = \sum_{j=1}^{D} \log p(x_j|z) = \sum_{j=1}^{D} \log\left(\text{Bernoulli}(x_j; p_j)\right)$$

$$= \sum_{j=1}^{D} x_j \log p_j + (1-x_j)\log(1-p_j)$$

Algorithm: Computation of an unbiased estimate of single-datapoint ELBO for VAE with full-covariance Gaussian inference model and a factorized Bernoulli generative model:

x - data point; u - random sample from $p(u) \sim N(0, I)$;

θ - generative model params; ϕ - inference model params

$q_\phi(z|x)$ - inference model; $p_\theta(x,z)$ - generative model

$\tilde{\mathcal{L}}$ - unbiased estimate of the single-datapoint ELBO $\mathcal{L}_{\theta,\phi}(x)$

$(\mu, L) \leftarrow \text{EncoderNN}_\phi(x)$

$u \sim N(0, I)$

$z \leftarrow Lu + \mu$

$\mathcal{L}^1 = \tilde{\mathcal{L}}[\log q_\phi(z|x)] \leftarrow -\sum_i \left(\frac{1}{2}(u_i^2 + \log(2\pi)) + \log \sigma_i\right)$

$\mathcal{L}^2 = \tilde{\mathcal{L}}[\log p_\theta(z)] \leftarrow -\sum_i \left(\frac{1}{2}(z_i^2 + \log(2\pi))\right)$

$p \leftarrow \text{DecoderNN}_\theta(z)$

$\mathcal{L}^3 = \tilde{\mathcal{L}}[\log p_\theta(x|z)] \leftarrow \sum_i \left(x_i \log p_i + (1-x_i)\log(1-p_i)\right)$

$\tilde{\mathcal{L}} = -\mathcal{L}^1 + \mathcal{L}^2 + \mathcal{L}^3$