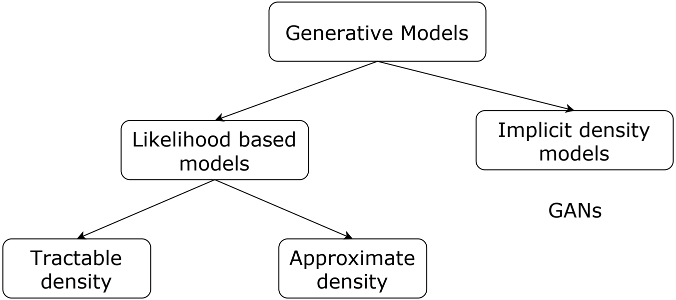
Deep Generative Models

Roman Isachenko

Moscow Institute of Physics and Technology

2019

Generative models zoo



- Autoregressive models
- Flow models

Latent variable models

Bayesian framework

- $\mathbf{x}$ – observed samples;
- $\mathbf{z}$ – unobserved (latent) variables;
- $\boldsymbol{\theta}$ – model parameters.

Discriminative

$$p(\mathbf{z}, \boldsymbol{\theta} | \mathbf{x}) = p(\mathbf{z} | \mathbf{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

Classification/Regression

Generative

$$p(\mathbf{z}, \mathbf{x}, \boldsymbol{\theta}) = p(\mathbf{z}, \mathbf{x} | \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

Generation of new samples $(\mathbf{z}, \mathbf{x})$

Bayesian framework

Bayes theorem

$$p(\boldsymbol{\theta}|\mathbf{X},\mathbf{Z}) = \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{X},\mathbf{Z})} = \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{X},\mathbf{Z})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

Full Bayesian inference

$$p(\mathbf{z}^*,\mathbf{x}^*|\mathbf{X},\mathbf{Z}) = \int p(\mathbf{z},\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{X},\mathbf{Z})d\boldsymbol{\theta}$$

Maximum a posteriori (MAP)

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X},\mathbf{Z}) = \arg\max_{\boldsymbol{\theta}}\big(\log p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\big)$$

Latent variable models

MLE problem

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\mathbf{X}|\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p(\mathbf{x}_i|\boldsymbol{\theta}).$$

Challenge

$p(\mathbf{x}|\boldsymbol{\theta})$ could be intractable.

Extend probabilistic model

Introduce latent variable $\mathbf{z}$ for each sample $\mathbf{x}$

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}); \quad \log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) + \log p(\mathbf{z}).$$

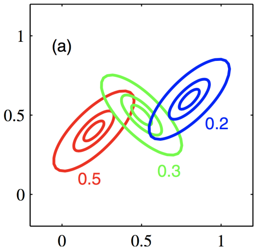
$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z} = \int p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}) d\mathbf{z}.$$

Latent variable models

$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \log \int p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})p(\mathbf{z})d\mathbf{z}$$

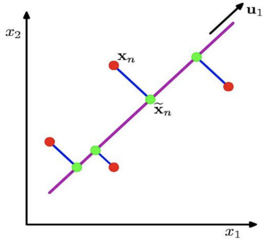
Examples

Mixture of gaussians



- $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{z}, \boldsymbol{\Sigma}_\mathbf{z}^2)$
- $p(\mathbf{z}) = \mathrm{Cat}(\mathbf{z}|\boldsymbol{\pi})$

PCA model



- $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\mathbf{V}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\Sigma}_\mathbf{z}^2)$
- $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|0, \mathbf{I})$

Bishop C. Pattern Recognition and Machine Learning, 2006.

Incomplete likelihood

MLE problem

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta}) =$$

$$= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta}).$$

Since Z is unknown, maximize **incomplete likelihood**.

MILE problem

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \log p(\mathbf{X}|\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \log \int p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) d\mathbf{Z} =$$

$$= \arg\max_{\boldsymbol{\theta}} \log \int p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta}) p(\mathbf{Z}) d\mathbf{Z}.$$

Variational lower bound

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \log \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})} =$$

$$= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})} d\mathbf{Z} = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})q(\mathbf{Z})} d\mathbf{Z} =$$

$$= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} + \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})} d\mathbf{Z} =$$

$$= \mathcal{L}(q,\boldsymbol{\theta}) + KL(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})) \geq \mathcal{L}(q,\boldsymbol{\theta}).$$

Kullback-Leibler divergence

- $KL(q||p) \geq 0$;
- $KL(q||p) = 0 \Leftrightarrow q \equiv p$.

Variational lower bound

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})) \geq \mathcal{L}(q, \boldsymbol{\theta}).$$

ELBO

$$\begin{aligned}\mathcal{L}(q, \boldsymbol{\theta}) &= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} = \\ &= \int q(\mathbf{Z}) \log p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta}) d\mathbf{Z} + \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} \\ &= \mathbb{E}_q \log p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta}) - KL(q(\mathbf{Z})||p(\mathbf{Z}))\end{aligned}$$

Instead of maximizing incomplete likelihood, maximize ELBO

$$\max_{\theta} p(\mathbf{X}|\boldsymbol{\theta}) \quad \rightarrow \quad \max_{q, \theta} \mathcal{L}(q, \boldsymbol{\theta}).$$

EM-algorithm

$$\mathcal{L}(q,\boldsymbol{\theta}) = \int q(\mathbf{Z}) \log p(\mathbf{X}|\mathbf{Z},\boldsymbol{\theta}) d\mathbf{Z} + \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z}.$$

Block-coordinate optimization

- Initialize $\boldsymbol{\theta}^*$;
- E-step

$$q(\mathbf{Z}) = \arg\max_q \mathcal{L}(q,\boldsymbol{\theta}^*) = \arg\min_q KL(q||p) = p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^*);$$

- M-step

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q,\boldsymbol{\theta});$$

- Repeat E-step and M-step until convergence.

Amortized variational inference

E-step

$$q(\mathbf{Z}) = \arg\max_q \mathcal{L}(q, \boldsymbol{\theta}^*) = \arg\min_q KL(q||p) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^*).$$

could be **intractable**.

Idea

Restrict the family of all possible distributions $q(\mathbf{z})$ to the particular parametric class conditioned of sample: $q(\mathbf{z}|\mathbf{x}, \phi)$.

Variational Bayes

- E-step

$$\phi_n = \phi_{n-1} + \eta \nabla_{\phi} \mathcal{L}(\phi, \boldsymbol{\theta}_{n-1})|_{\phi=\phi_{n-1}}$$

- M-step

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\phi_n, \boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{n-1}}$$

ELBO gradient (M-step)

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \mathbb{E}_q \log p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta}) - KL(q(\mathbf{Z}|\mathbf{X}, \phi)||p(\mathbf{Z})) \to \max_{\boldsymbol{\phi}, \boldsymbol{\theta}}.$$

Optimization w.r.t. $\boldsymbol{\theta}$: **mini-batching** (1) + **Monte-Carlo estimation** (2)

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}) &= \sum_{i=1}^{n} \int q(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\phi}) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta}) d\mathbf{z}_i \\
&\overset{(1)}{=} n \int q(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\phi}) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta}) d\mathbf{z}_i, \quad i \sim U[1, n] \\
&\overset{(2)}{=} n \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}_i|\mathbf{z}_i^*, \boldsymbol{\theta}), \quad \mathbf{z}_i^* \sim q(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\phi}).
\end{aligned}$$

ELBO gradient (E-step)

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \mathbb{E}_q \log p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta}) - KL(q(\mathbf{Z}|\mathbf{X}, \phi)||p(\mathbf{Z})) \to \max_{\boldsymbol{\phi}, \boldsymbol{\theta}}.$$

Optimization w.r.t. ϕ: density function depends on the parameters.

Hint 1 (log-derivative trick)

$$\nabla_x p(y|x) = p(y|x) \nabla_x \log p(y|x).$$

Hint 2

$$\begin{aligned}
\nabla_x f(x) &= \nabla_x \int p(y|x) h(y) dy \\
&= \int \left(\nabla_x p(y|x) \right) h(y) dy \Big) dy \\
&\sim h(y_0) \nabla_x \log p(y_0|x) \quad y_0 \sim p(y|x).
\end{aligned}$$

ELBO gradient (E-step)

$$\mathcal{L}(\phi, \boldsymbol{\theta}) = \mathbb{E}_q \log p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta}) - KL(q(\mathbf{Z}|\mathbf{X}, \phi)||p(\mathbf{Z})) \to \max_{\boldsymbol{\phi}, \boldsymbol{\theta}}.$$

Optimization w.r.t. ϕ: density function depends on the parameters.

$$\nabla_\phi \int q(\mathbf{Z}|\mathbf{X}, \phi) \log p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta}) d\mathbf{Z} \sim \log p(\mathbf{x}_i|\mathbf{z}_i^*, \boldsymbol{\theta}) \nabla_\phi \log q(\mathbf{z}_i^*|\mathbf{x}_i, \phi),$$

$$\mathbf{z}_i^* \sim q(\mathbf{z}_i^*|\mathbf{x}_i, \phi).$$

Problem

Unstable solution with huge variance.

Solution

Reparametrization trick

ELBO gradient (E-step)

Reparametrization trick

$$f(x) = \int p(y|x)h(y)dy$$

$$\nabla_x \int p(y|x)h(y)dy = \nabla_x \int r(\epsilon)h(g(x,\epsilon))d\epsilon$$
$$= \nabla_x h(g(x,\epsilon^*)), \quad \epsilon^* \sim r(\epsilon).$$

Example

$$q(z|x) = \mathcal{N}(z|\mu,\sigma^2), \quad r(\epsilon) = \mathcal{N}(\epsilon|0,1), \quad z = \sigma\epsilon + \mu.$$

ELBO gradient (E-step)

Derivative

$$\nabla_\phi \int q(\mathbf{Z}|\mathbf{X}, \phi) \log p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta}) d\mathbf{Z} \sim$$
$$n\nabla_\phi \int r(\epsilon) \log p(\mathbf{x}_i|g(\mathbf{x}_i, \epsilon, \phi), \boldsymbol{\theta}) d\epsilon \sim$$
$$n\nabla_\phi \log p(\mathbf{x}_i|g(\mathbf{x}_i, \epsilon^*, \phi), \boldsymbol{\theta}), \quad \epsilon^* \sim r(\epsilon).$$

Variational assumption

$$q(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\Sigma}(\mathbf{x})).$$

Variational autoencoder (VAE)

Final algorithm

- pick $i \sim U[1, n]$;
- compute stochastic gradient w.r.t. $\boldsymbol{\theta}$

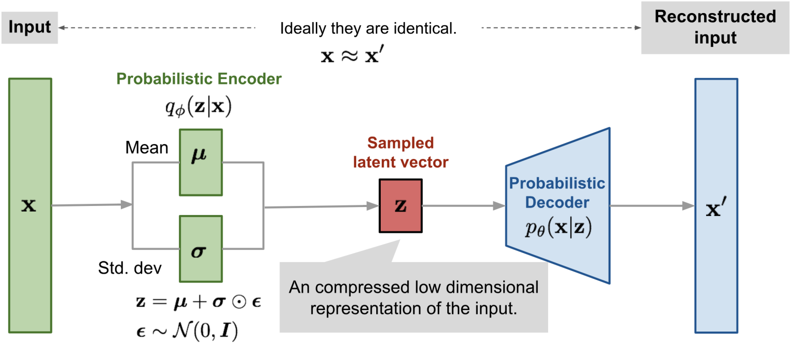
$$n \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}_i | \mathbf{z}_i^*, \boldsymbol{\theta}), \quad \mathbf{z}_i^* \sim q(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\phi});$$

- compute stochastic gradient w.r.t. $\boldsymbol{\phi}$

$$n \nabla_{\boldsymbol{\phi}} \log p(\mathbf{x}_i | g(\mathbf{x}_i, \epsilon^*, \boldsymbol{\phi}), \boldsymbol{\theta}) - \nabla_{\boldsymbol{\phi}} KL(q(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\phi}) || p(\mathbf{z}_i)), \; \epsilon^* \sim r(\epsilon);$$

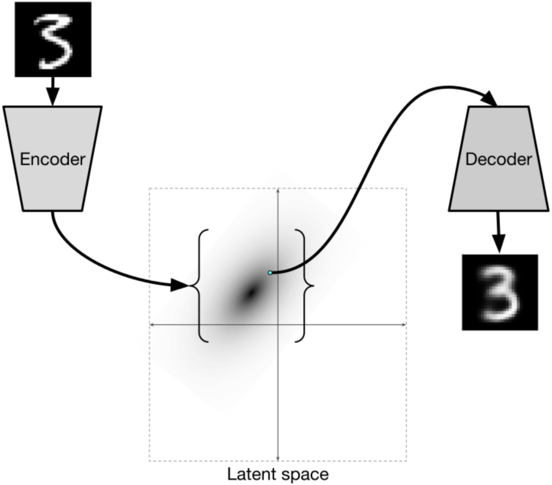
- update $\boldsymbol{\theta}, \boldsymbol{\phi}$ according to the selected optimization method (SGD, Adam, RMSProp).

Variational autoencoder (VAE)



https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html

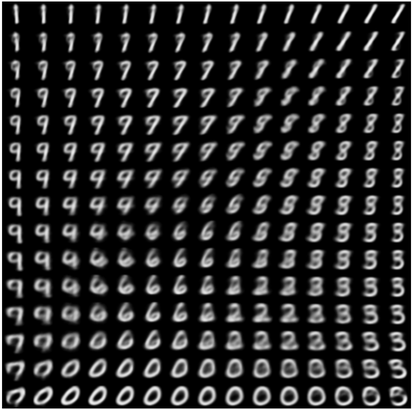
Variational Autoencoder



Isaac Dykeman, http://ijdykeman.github.io/ml/2016/12/21/cvae.html

Variational Autoencoder

Generation objects by sampling the latent space $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$



http://bit.ly/2w73aXB

References

- *Variational Bayesian inference with Stochastic Search*
  https://arxiv.org/abs/1206.6430
- *Stochastic Variational Inference*
  https://arxiv.org/abs/1206.7051
- *Doubly Stochastic Variational Bayes for non-Conjugate Inference*
  http://proceedings.mlr.press/v32/titsias14.pdf
- *Auto-Encoding Variational Bayes*
  https://arxiv.org/abs/1312.6114
- *Markov chain Monte Carlo and variational inference: Bridging the gap*
  https://arxiv.org/pdf/1410.6460.pdf
- *Tutorial on Variational Autoencoders*
  http://arxiv.org/abs/1606.05908