

Deep Generative Models

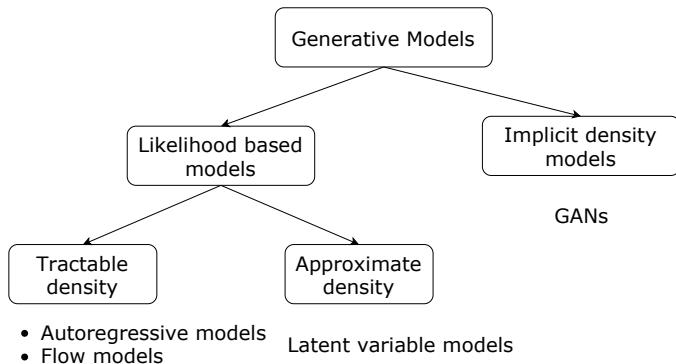
Lecture 2

Roman Isachenko

Moscow Institute of Physics and Technology

2020

Generative models zoo



Bayesian framework

- ▶ \mathbf{x} – samples;
- ▶ \mathbf{y} – target variables;
- ▶ θ – model parameters.

Discriminative

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

- ▶ Find conditional probability of \mathbf{y} given \mathbf{x} .
- ▶ Samples \mathbf{x} are given.
- ▶ Used for classification, regression.

Generative

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

- ▶ Find joint probability of (\mathbf{x}, \mathbf{y}) .
- ▶ Samples \mathbf{x} should be modelled.
- ▶ Generation of new samples (\mathbf{x}, \mathbf{y}) .

Generative models

Given samples $\{\mathbf{x}_i\}_{i=1}^n \in X$ from unknown distribution $p(\mathbf{x})$.

Goal

learn a distribution $p(\mathbf{x})$ for

- ▶ evaluating $p(\mathbf{x})$ for new samples;
- ▶ sampling from $p(\mathbf{x})$.

Challenge

Data is complex and high-dimensional (curse of dimensionality).

Solution

Fix probabilistic model $p(\mathbf{x}|\boldsymbol{\theta})$ – the set of parameterized distributions .

Instead of searching true $p(\mathbf{x})$ over all probability distributions, learn function approximation $p(\mathbf{x}|\boldsymbol{\theta}) \approx p(\mathbf{x})$.

Latent variable models

Suppose that our probabilistic model $p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$ instead of $p(\mathbf{x}|\boldsymbol{\theta})$.

- ▶ Here \mathbf{z} are latent variables.
- ▶ We observe only samples \mathbf{x} .
- ▶ Latent variables \mathbf{z} are unknown.
- ▶ Parameters $\boldsymbol{\theta}$ are not random.

MLE problem for LVM

$$\begin{aligned}\boldsymbol{\theta}^* &= \arg \max_{\boldsymbol{\theta}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \arg \max_{\boldsymbol{\theta}} \prod_{i=1}^n p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta}) = \\ &= \arg \max_{\boldsymbol{\theta}} \sum_{i=1}^n \log p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta}).\end{aligned}$$

What if $\boldsymbol{\theta}$ are random variables with distribution $p(\boldsymbol{\theta})$?

Bayesian framework

What if θ are random variables with distribution $p(\theta)$?

Before we get any data, we do not know anything about θ except the **prior** distribution $p(\theta)$.

When we get data, we could change the **prior** distribution to the **posterior**.

Bayes theorem

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

Full Bayesian inference

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

Bayesian framework

Full Bayesian inference

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

Maximum a posteriori (MAP)

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

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

Latent variable models

MLE problem

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

Challenge

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

Extend probabilistic model

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

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

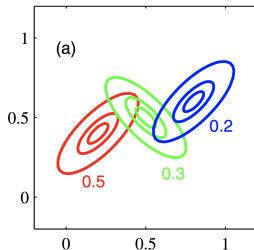
$$p(\mathbf{x}|\theta) = \int p(\mathbf{x}, \mathbf{z}|\theta) d\mathbf{z} = \int p(\mathbf{x}|\mathbf{z}, \theta)p(\mathbf{z}) d\mathbf{z}.$$

Latent variable models

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

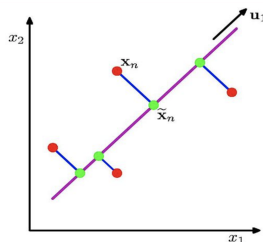
Examples

Mixture of gaussians



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

PCA model

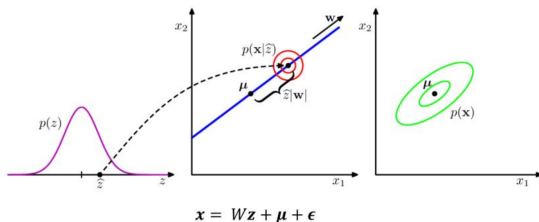


- ▶ $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\Sigma}_{\mathbf{z}}^2)$
- ▶ $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$

Latent variable models

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

PCA goal: Project original data \mathbf{X} onto low latent space while maximizing the variance of the projected data.



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

Incomplete likelihood

MLE problem

$$\begin{aligned}\theta^* &= \arg \max_{\theta} p(\mathbf{X}, \mathbf{Z} | \theta) = \arg \max_{\theta} \prod_{i=1}^n p(\mathbf{x}_i, \mathbf{z}_i | \theta) = \\ &= \arg \max_{\theta} \sum_{i=1}^n \log p(\mathbf{x}_i, \mathbf{z}_i | \theta).\end{aligned}$$

Since \mathbf{Z} is unknown, maximize **incomplete likelihood**.

MILE problem

$$\begin{aligned}\theta^* &= \arg \max_{\theta} \log p(\mathbf{X} | \theta) = \arg \max_{\theta} \log \int p(\mathbf{X}, \mathbf{Z} | \theta) d\mathbf{Z} = \\ &= \arg \max_{\theta} \log \int p(\mathbf{X} | \mathbf{Z}, \theta) p(\mathbf{Z}) d\mathbf{Z}.\end{aligned}$$

Variational lower bound

$$\begin{aligned}\log p(\mathbf{X}|\theta) &= \log \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{p(\mathbf{Z}|\mathbf{X}, \theta)} = \\&= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{p(\mathbf{Z}|\mathbf{X}, \theta)} d\mathbf{Z} = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\theta)q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)q(\mathbf{Z})} d\mathbf{Z} = \\&= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} d\mathbf{Z} + \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)} d\mathbf{Z} = \\&= \mathcal{L}(q, \theta) + KL(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X}, \theta)) \geq \mathcal{L}(q, \theta).\end{aligned}$$

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

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

EM-algorithm

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

Block-coordinate optimization

- ▶ Initialize θ^* ;
- ▶ E-step

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

- ▶ M-step

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

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

Amortized variational inference

E-step

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

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

$$\phi_k = \phi_{k-1} + \eta \nabla_{\phi} \mathcal{L}(\phi, \theta_{k-1})|_{\phi=\phi_{k-1}}$$

► M-step

$$\theta_k = \theta_{k-1} + \eta \nabla_{\theta} \mathcal{L}(\phi_k, \theta)|_{\theta=\theta_{k-1}}$$

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>