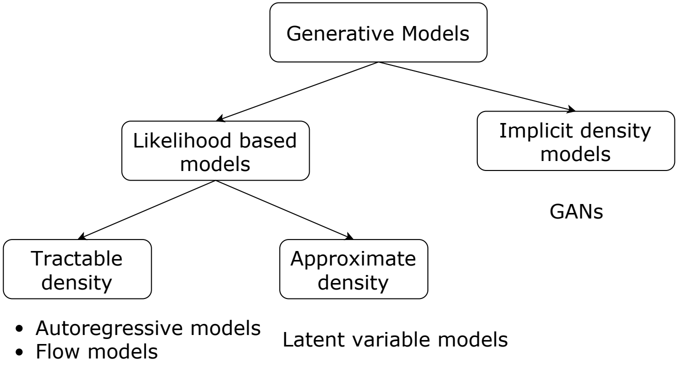
Deep Generative Models
Lecture 2

Roman Isachenko

Moscow Institute of Physics and Technology

2020

Generative models zoo



Bayesian framework

- $\mathbf{x}$ – samples;
- $\mathbf{y}$ – target variables;
- $\boldsymbol{\theta}$ – model parameters.

Discriminative

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

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

Generative

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

- 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

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

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

Bayesian framework

What if $\boldsymbol{\theta}$ are random variables with distribution $p(\boldsymbol{\theta})$?
Before we get any data, we do not know anything about $\boldsymbol{\theta}$ except the **prior** distribution $p(\boldsymbol{\theta})$.
When we get data, we could change the **prior** distribution to the **posterior**.

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

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

Bayesian framework

Full Bayesian inference

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

Maximum a posteriori (MAP)

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

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

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 $\mathbf{x}$

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

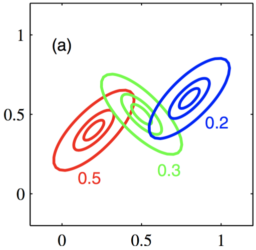
$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z} = \int p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}) d\mathbf{z}.$$

Latent variable models

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

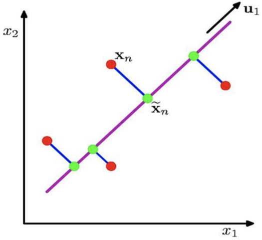
Examples

Mixture of gaussians



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

PCA model



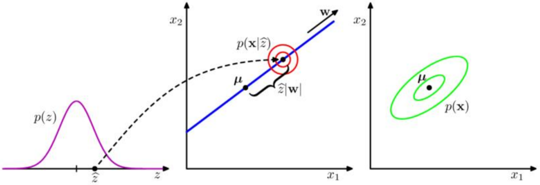
- $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\Sigma}_{\mathbf{z}}^2)$
- $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|0, \mathbf{I})$

Bishop C. Pattern Recognition and Machine Learning, 2006.

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.



$$\boldsymbol{x} = W\boldsymbol{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

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

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

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

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

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

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

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