Generative Models - DRE7053
Lecture 1

NORA Summer School 2024

Rogelio A Mancisidor
Assistant Professor
Department of Data Science and Analytics
BI Norwegian Business School

June 10-14, 2024

Outline

Outline

1. About the course
2. Background and Motivation
   - Joint Probability
   - Directed Graphical Models
   - Learning and Inference
3. Variational Inference
   - Latent Variable Models
   - Bayes' Theorem
   - Variational Inference
   - Evidence Lower Bound
   - Mean-Field Variational Inference

Schedule

Time	Day 1	Day 2	Day 3	Day 4	Day 5
09:00-10:00	Introduction	Reading discussion	Reading discussion	Reading discussion	Reading discussion
10:00-12:00	Bakground & motivation	Variational Inference II	Variational Autoencoder II	Multimodal Variational Autoencoder I	Diffusion Models (if we have time)
12:00-13:00	LUNCH				
13:00-15:00	Variational Inference I	Variational Autoencoder I	Variational Autoencoder III	Multimodal Variational Autoencoder II	Closing
15:00-17:00	Lab session	Lab session	Lab session	Lab session	

Figure: Tentative plan

Practical Information

- 45 min. sessions followed by 15 min. break
- Course GitHub website: link
- I will keep the level of the lab session to *basic*
- If your programming and TensorFlow level is *advanced*, you can skip the lab sessions
- Evaluation details:
  - Individual paper, limited to 8 content pages
  - Ideas for the paper: replicate a previous work with your own data/problem setting or propose a new idea/model, which fits into your own research, clearly stating the research question(s), method(s), and experimental design.
  - Pass/fail
  - Submission by August 23 12:00 in WiseFlow

Lab sessions

- We focus on two aspects:
  1. Object oriented programming
  2. TensorFlow and TensorFlow Probability
- 1) it is easier to think and understand models as classes with methods.
- 2) leverage automatic differentiation libraries and probabilistic programming.

Reading discussions

- We start the day by discussing one (or two) selected paper(s).
- I guide the discussion, but all of you must engage to have a good understanding about the paper(s).
- See the course GitHub website for selected papers.

Outline

The joint probability

- Given a set of n events $\{\cap_{i=1}^{n} E_i\} = \Omega$, the *chain rule of conditional probabilities* is

$$Pr(E_1, E_2, \cdots, E_n) = Pr(E_1)Pr(E_2|E_1)\cdots Pr(E_n|E_{n-1}, \cdots, E_1), \quad (1)$$

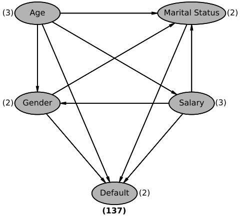
  that is, the joint probability of all possible events in Ω (sample space) can be expressed in terms of conditional probabilities.

A simple example

- Assume that we have the following variables:
  1. age
  2. salary
  3. gender
  4. marital status
  5. a binary class label y
- Continuous variables are discretisized, hence: $age \in \{a_1, a_2, a_3\}$, $salary \in \{s_1, s_2, s_3\}$, $gender \in \{g_1, g_2\}$, $ms \in \{ms_1, ms_2\}$, and $y \in \{0, 1\}$.

Directed Graphical Models - I

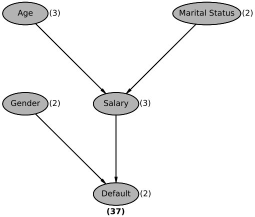
- A directed graphical model specifies probabilistic dependencies



- **Nodes** represent random events (variables)
- **Arrows** indicate probabilistic dependencies

Directed Graphical Models - II

- Based on expert knowledge:
    - *age*, *gender* and *ms* are independent variables
    - *y* only depends on *gender* and *salary*
    - *salary* depends on both *age* and *ms*



What is the joint probability?

- We can use the directed graphical models to specify joint probability!
- Chain rule of conditional probability:

$$\begin{aligned} p(age, salary, gender, ms, y) = {} & p(age)p(salary|age)p(gender|salary, age) \\ & p(ms|gender, salary, age)p(y|ms, gender, salary, age) \end{aligned}$$

- Using expert knowledge:

$$\begin{aligned} p(age, salary, gender, ms, y) = {} & p(age)p(gender)p(ms) \\ & p(salary|ms, age)p(y|salary, gender) \end{aligned}$$

Unknown parameters and non-redundant parameters

- Number of parameters using the chain rule:
  - unknown probabilities: 138
- Number of parameters based on the factorization using expert knowledge:
  - unknown probabilities: 37
- The number of probabilities has decreased drastically!
- Note that we do not need to learn all unknown probabilities!

Conditional Probability Tables

Advantages of Directed Graphical Models

- Compact representation for complex joint distributions
- Where knowledge is used to specify conditional independence
- Key concept in Deep Generative Models that we (should) leverage!

Outline

Latent Variable Models

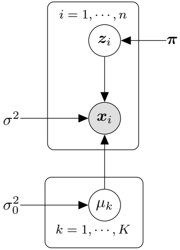
- Suppose that we observe a vector $\boldsymbol{x} = (x_1, x_2, \cdots, x_\ell)^T$, where $\boldsymbol{x} \in \mathbb{R}^\ell$
- It makes sense to assume that $\boldsymbol{x}$ is generated, or governed, by an unseen (latent) variable $\boldsymbol{z} = (z_1, z_2, \cdots, z_d)^T$, $\boldsymbol{z} \in \mathbb{R}^d$
- Latent Variable Models (LVMs) were introduced in a study during World War II.
- A popular LVM is the Bayesian Gaussian Mixture Model (GMM)

The Bayesian GMM - I

- Suppose that we observe
  - $\boldsymbol{X} = \{x_1, x_2, \cdots, x_n\}$ - univariate Gaussian variables
  - $\boldsymbol{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_n\}$ - one-hot-encoded latent variables
  - Hence, for each x_i there is a $\boldsymbol{z}_i$ indicating the k-th component in the GMM to which x_i belongs to
- Let $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_k)^T$ be the vector of expectation parameters
- The variance parameter in the Gaussian likelihood-function is also known
- Assume, for simplicity, μ_k is drawn independently from a common and known Gaussian distribution

The Bayesian GMM - II

- The probabilistic graphical model of the GMM:



- The distributions in the GMM are as follow:

$$\begin{aligned}
\mu_k \sim & \mathcal{N}(0, \sigma_0^2) && k = 1, ..., K, \\
\boldsymbol{z}_i \sim & \text{cat}(\boldsymbol{\pi}) && i = 1, ..., n, \\
x_i | \boldsymbol{z}_i, \boldsymbol{\mu} \sim & \mathcal{N}(\boldsymbol{z}_i^T \boldsymbol{\mu}, \sigma^2) && i = 1, ..., n,
\end{aligned}$$

The Bayes' Theorem

Intractable Model

- (As usual) we need to arrive at the posterior distribution

- How can we evaluate the marginal distribution in the denominator?

- This K-dimensional integral has $\mathcal{O}(K^n)$ complexity!
- Therefore, this (simple) Bayesian GMM model is intractable!

Variational Inference

- Let's look into the most simple case
  - $\boldsymbol{x} = (x_1, x_2, \cdots, x_\ell)^T$ is modulated by a latent variable $\boldsymbol{z} = (z_1, z_2, \cdots, z_d)^T$ through the conditional distribution $p(\boldsymbol{x}|\boldsymbol{z})$
  - we want to arrive at the posterior distribution $p(\boldsymbol{z}|\boldsymbol{x})$, which has a prior distribution $p(\boldsymbol{z})$

$$p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{x})}, \tag{2}$$

- The problem is that the marginal distribution $p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{z})d\boldsymbol{z}$ is intractable, so the posterior!
- VI replaces $p(\boldsymbol{z}|\boldsymbol{x})$ with the *variational distribution* $q(\boldsymbol{z}; \boldsymbol{\lambda}) \in \mathcal{Q}$.
- Where $\boldsymbol{\lambda}$ are variational parameters and $\mathcal{Q}$ is the familiy of variational distributions

Kullback-Leibler Divergence

- It is tempting to minimize the Kullback-Leibler (KL) divergence:

$$KL[q(\boldsymbol{z};\boldsymbol{\lambda})||p(\boldsymbol{z}|\boldsymbol{x})] \tag{3}$$

- wait a second... is $p(\boldsymbol{x})$ not intractable?

Evidence Lower Bound - Version 1

- Let's instead do the following:

- Note that ELBO $= \log p(\boldsymbol{x}) - KL[p(\boldsymbol{z}|\boldsymbol{x})||q(\boldsymbol{z}|\boldsymbol{x})]$, i.e., maximizing the ELBO is equivalent to minimizing the KL divergence.

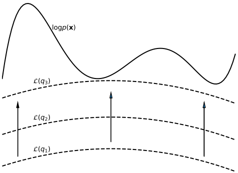


Figure: Learn $\boldsymbol{\lambda}$ to minimize the variational gap

Mean-Field (MF) Approximation - I

- MF assumes that $\mathcal{Q}$ is in the class of fully factorized distributions

$$q(\boldsymbol{Z};\boldsymbol{\lambda}) = \prod_{i=i}^{N} q(\boldsymbol{z}_i;\boldsymbol{\lambda}_i)$$

- Intuitively, MF optimizes each $\boldsymbol{z}_i$ one at a time while holding the other fixed
- Let's write the ELBO for the entire data set as

$$\begin{aligned} ELBO =& \mathbb{E}_q\Big[\log \frac{p(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})}\Big] \\ =& \log p(\boldsymbol{X}) + \mathbb{E}_q[\log p(\boldsymbol{Z}|\boldsymbol{X})] - \mathbb{E}_q[q(\boldsymbol{Z})] \\ \propto& \mathbb{E}_q[\log p(\boldsymbol{Z}|\boldsymbol{X})] - \mathbb{E}_q[q(\boldsymbol{Z})] \end{aligned}$$

Mean-Field Approximation - II

- Due to the *mean-field* assumption and chain rule of conditional probabilites

$$ELBO = \sum_i \mathbb{E}_q[\log p(\boldsymbol{z}_i|\boldsymbol{z}_{-i}, \boldsymbol{X})] - \mathbb{E}_q[q(\boldsymbol{Z})]$$

  where $\boldsymbol{z}_{-i}$ means all variables in $\boldsymbol{Z}$ but $\boldsymbol{z}_i$.

- The ELBO as a function of the j-th variational distribution is

$$\begin{aligned} ELBO(q_j) =& \mathbb{E}_{q_j}[\mathbb{E}_{q_{-j}}[\log p(\boldsymbol{z}_i|\boldsymbol{z}_{-i}, \boldsymbol{X})]] - \mathbb{E}_{q_j}[q_j(\boldsymbol{Z})] \\ =& \mathbb{E}_{q_j}[\log(\exp\{\mathbb{E}_{q_{-j}}[\log p(\boldsymbol{z}_i|\boldsymbol{z}_{-i}, \boldsymbol{X})]\})] - \mathbb{E}_{q_j}[q_j(\boldsymbol{Z})] \\ =& -KL[q_j || \exp\{\mathbb{E}_{q_{-j}}[\log p(\boldsymbol{z}_i|\boldsymbol{z}_{-i}, \boldsymbol{X})]\})] \end{aligned}$$

Mean-Field Approximation - III

- Recall that $KL[p||q]$ is minimized when $p = q$, hence
- Hence, the MF update is

$$q_j^* \propto \exp\{\mathbb{E}_{q_{-j}}[\log p(\boldsymbol{z}_i|\boldsymbol{z}_{-i}, \boldsymbol{X})]\}$$

- Note that
  - the updating step computes an expectation (hence *mean* in MF)
  - we need to be able to evaluate the expectation (an overoptimistic assumption)
  - optimization does not scale to large datasets