Deep Probabilistic Generative Models
Exam 2021-2022

21 oct. 2021

Read everything before starting to answer questions. You can answer either in French or English.
Duration: 2 hours.

1 General questions [5 points + 2 bonus points]

1. In this course, what did we call generative models? Why is it different from (for example) classification problems in machine learning? **[1 point]**

2. What does latent mean in *latent random variables*? What does latent random variables mean? Why are they useful? (for example in a Gaussian Mixture Model) **[2 point]**

3. What is the difference between locally normalized models and implicit models in terms of generation? Is there any benefit from this perspective in one of the two families? Explain. **[2 point]**

4. **Bonus:** what is your favourite generative model that we studied in the course? Why? **[2 point]**

2 Locally normalized models [10 point]

Let $Z \in \{1...m\}$ and $X \in \mathbb{R}$ be the latent and observed random variables, respectively. We assume the following generative story:

$$\begin{aligned} &(1) \quad z \sim p(Z; \boldsymbol{\pi}) \\ &(2) \quad x \sim p(X|Z = z; \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \end{aligned}$$

where $\boldsymbol{\pi} \in \triangle^m$, $\boldsymbol{\mu} \in \mathbb{R}^m$ and $\boldsymbol{\sigma} \in \mathbb{R}^m_{++}$. We denote the set of parameters of the generative model $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}\}$. In the following, we will denote $p(Z; \boldsymbol{\theta})$ instead of $p(Z; \boldsymbol{\pi})$ as the corresponding parameters are unambiguous (and similarly for the second distribution). The prior probability distribution and the conditional probability density (PDF) function are defined as follows:

$$\begin{aligned} p(Z = z; \boldsymbol{\pi}) &= \pi_z \\ p(X = x|Z = z; \boldsymbol{\mu}, \boldsymbol{\sigma}^2) &= \mathcal{N}(\mu_z, \sigma_z^2) \quad \text{that is the PDF of a Gaussian} \end{aligned}$$

In other word, this model is a Gaussiam Mixture Model. Let D be the training dataset. We assume a proposal distribution $q(Z|X; \boldsymbol{\phi})$ where $\boldsymbol{\phi} \in \mathbb{R}^{D \times m}$, that is $\forall x \in D, z \in \{1...m\} : q(Z = z|X = x; \boldsymbol{\phi}) = \phi_{x,z}$. Training aims to find the parameters $\boldsymbol{\theta}^*$ that maximizes the log-likelihood of the data:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \frac{1}{|D|} \sum_{\boldsymbol{x} \in D} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\theta})$$

where $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\theta}) = \log \sum_{z \in \{1...m\}} p(z; \theta) p(x|z; \theta)$.

1. Is computing the log-likelihood tractable in this model? Could we use it directly to train the model? **[1 point]**

2. What constraints must be satisfied by $\boldsymbol{\phi}$? **[1 point]**

3. Derive the Evidence Lower Bound $\mathcal{E}(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi})$ and explain each step. **[1 point]**

4. Explain the Expectation-Maximization algorithm. **[2 point]**

5. Derive the closed-form expression of the computation of the Expectation step **using the KKT conditions** (note that you can focus on a single datapoint, why?) For this, it is easier to consider $p(Z, X; \boldsymbol{\theta})$ jointly and never "break" the generative distributions. Could you expect this result? Why? **[3 point]**

6. Explain the difference in the EM training algorithm between GMMs, Sigmoid Belief Networks and Variational Auto-Encoders. **[2 point]**

3 Globally normalized models [10 points]

Let $Z \in \{0,1\}^m$ and $X \in \{0,1\}^n$ be the latent and observed random variables, respectively. We assume a globally normalized generative model, that is the joint distribution probability over random variables is defined as follows:

$$p(\boldsymbol{z}, \boldsymbol{x}; \theta) = \frac{\exp(w(\boldsymbol{z}, \boldsymbol{x}; \theta))}{\exp(c(\boldsymbol{\theta}))} = \exp(w(\boldsymbol{z}, \boldsymbol{x}) - c(\boldsymbol{\theta}))$$

where $c(\theta)$ is the log-partition function and we don't make any assumption on the structure of the function $w : Z \times X \to \mathbb{R}$ for the moment.

1. What is the formula of $c(\boldsymbol{\theta})$? What does this function achieve in the definition of the joint probability distribution? Is it easy to compute in the general case? **[2 point]**

2. Derive and explain the Monte-Carlo estimation of the gradient for training this model. **[2 point]**

3. In the course, we studied two Markov Chain Monte Carlo methods to sample from a generative model: Gibbs sampling and Metropolis–Hastings. What assumption we need to make on the generative model in order to be able to use them? **[2 point]**

In a Boltzmann Machine, the function w is defined as follows:

$$w(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{a}^\top \boldsymbol{z} + \boldsymbol{b}^\top \boldsymbol{x} + \boldsymbol{z}^\top \boldsymbol{C} \boldsymbol{x} + \sum_{i \in \{1 \dots m\}} \sum_{j \in \{1 \dots m\}} D_{i,j} z_i z_j + \sum_{i \in \{1 \dots n\}} \sum_{j \in \{1 \dots n\}} E_{i,j} x_i x_j$$

where parameters are $\boldsymbol{\theta} = \{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{E}\}$ with $\boldsymbol{a} \in \mathbb{R}^m, \boldsymbol{b} \in \mathbb{R}^n, \boldsymbol{C} \in \mathbb{R}^{m \times n}, \boldsymbol{D} \in \mathbb{R}^{m \times m}$ and $\boldsymbol{E} \in \mathbb{R}^{n \times n}$.

4. What is the difference between Boltzmann Machines and Restricted Boltzmann Machines? **[1 point]**

5. In Restricted Boltzmann Machines, we saw that one of the term in the objective was easy to compute thanks to a reformulation. Is it still the case? Why? **[1 point]**

6. What does that change for the Gibbs sampling algorithm that we used to train Restricted Boltzmann Machines? **[2 point]**

4 Implicit models [5 points]

Let $Z \in \mathbb{R}^n$ and $X \in \mathbb{R}^n$ be the latent and observed random variables, respectively. The generative story of an implicit model is defined as follows:

$$\boldsymbol{z} \sim p(Z)$$
$$\boldsymbol{x} = g(\boldsymbol{z}; \boldsymbol{\theta})$$

During the course, we focused on normalizing flows, so the following questions concern normalizing flows.

- Why do we need the change of variable theorem? **[1 point]**
- What assumption do we need to make on function g? **[1 point]**
- When we define the function g, what do we need to be careful about if we want training to be fast? Why? **[1 point]**
- In the course, we saw a trick several times: always keep a part of the input vector of g fixed in its output (or in the various normalizing flow layers that g contains). Explain two benefits of this. **[2 point]**

Cheat sheet

Consider the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{v}} \quad & f(\boldsymbol{v}) \\ \text{s.t.} \quad & g_i(\boldsymbol{v}) = 0 \quad \forall i \in \{1...m\} \\ & h_i(\boldsymbol{v}) \leq 0 \quad \forall i \in \{1...n\} \end{aligned}$$

where $\boldsymbol{v} \in \mathbb{R}^k$ are the variables, f the objective function and g_i and h_i the equality and inequality constraints, respectively. The Lagrangian of the problem is:

$$L(\boldsymbol{v}, \lambda, \mu) = f(\boldsymbol{v}) + \sum_{i \in \{1...m\}} \theta_i g_i(\boldsymbol{v}) + \sum_{i \in \{1...n\}} \mu_i h_i(\boldsymbol{v})$$

where $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^n_+$ are dual variables associated with equalities and inequalities, respectively. The KKT optimality conditions are defined as follows:

$$\begin{aligned} \text{(stationarity)} \quad & \forall i \in 1...k : \frac{\partial}{\partial v_i} L(\boldsymbol{v}, \lambda, \mu) = 0 \\ \text{(primal feasibility)} \quad & \forall i \in 1...m : g_i(\boldsymbol{v}) = 0 \\ & \forall i \in 1...n : h_i(\boldsymbol{v}) \leq 0 \\ \text{(dual feasibility)} \quad & \forall i \in 1...n : \mu_i \geq 0 \\ \text{(complementary slackness)} \quad & \sum_{i \in 1...n} \mu_i h_i(\boldsymbol{v}) = 0 \end{aligned}$$