Final Homework

DS-GA 1008 Deep Learning

Fall 2021

This homework covers EBMs and Variational Inference.

For theory questions, you should put all your answers in a PDF file and we will not accept any scanned hand-written answers. It is recommended to use LaTeX.

For the coding question, you need to program in Python. We recommend you to use **PyTorch** and **NumPy** for the experiments and **Matplotlib** to draw the plot. You need to submit your python files and describe the results with plots in the pdf.

If you only submit your python files without relevant description and plots in the pdf, you will get no points for the questions.

The due date of this homework is `23:55 EST` of `12/14`. Submit the following files in a zip file `your_net_id.zip` through NYU Brightspace:

- `theory.pdf`
- `VI.py`

The following behaviors will result in penalty of your final score:

1. 5% penalty for submitting your files without using the correct format. (including naming the zip file, PDF file or python file wrong, or adding extra files in the zip folder, like the testing scripts from coding questions).
2. 20% penalty for late submission within the first 24 hours. We will not accept any late submission after the first 24 hours.
3. 20% penalty for code submission that cannot be executed using the steps we mentioned in coding questions. So please test your code before submit it.

1 Generative Models and EBMs

In this exercise, you will compare Energy-Based-Models with simple energy functions with alternative generative models. To make computations tractable, we will focus on low-dimensional data, of dimension $d = 2$.

1. Generate $n = 1000$ datapoints from a Gaussian mixture using $K = 10$ mixture components with $\pi_k = 1/10$, $\mu_k = R(\cos(2\pi k/K), \sin(2\pi k/K))$, $\Sigma_k = I_2$, $k = 1 \dots K$, with $R = 10$. Write down the probability density function of the model. [5pt]

2. Assuming that $\Sigma_k = I_2$ for each k, use the EM algorithm to estimate the model parameters (π and μ_k). Repeat the experiment using $R = 1$. Interpret the differences. [15pt]

3. Fit a *kernel density estimator* to the data, using the $R = 10$ setting above. Given the data $\{x_1, \dots, x_n\}$, a kernel density estimator is defined as

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \phi_\sigma(x - x_i) \ ,$$

   where ϕ_σ is the pdf of an isotropic Gaussian distribution with covariance σI_2. Tune the parameter σ by using a *validation set*, ie a fresh batch of samples $\{x'_1, \dots, x'_n\}$ drawn from the same GMM model. Using a *test set* of 1000 samples $\{\tilde{x}_1, \dots, \tilde{x}_n\}$, evaluate the log-likelihood of your model

$$\frac{1}{n} \sum_{i=1}^{n} \log \hat{p}(\tilde{x}_i) \ .$$

   [15pt]

Using the same training set, we will now fit an Energy-Based-Model using a simple shallow NN as your energy function.

4. Show that the MLE estimator in the EBM family given by $p_\theta(x) = \exp(E_\theta(x) - A(\theta))$, with $A(\theta) = \log \int \exp(E_\theta(x)) dx$, is the global optimiser of the loss

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} E_\theta(x_i) - A(\theta) \ ,$$

   and

$$\nabla_\theta \mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta E_\theta(x_i) - \mathbb{E}_{x \sim p_\theta} \nabla_\theta E_\theta(x) \ .$$

   [5pt]

5. Since $d = 2$ you can afford to sample using *importance sampling*: consider estimating a quantity of the form $F = \mathbb{E}_{x \sim p_\theta} f(x)$. Consider a base probability measure $q(x)$. Show that

$$F = \frac{\mathbb{E}_{x \sim q}[f(x) \exp(E_\theta(x))/q(x)]}{\mathbb{E}_{x \sim q}[\exp(E_\theta(x))/q(x)]} \ .$$

   [10pt]

6. This suggests the importance sampling estimator

$$\hat{F} = \frac{\hat{N}}{\hat{D}} ,$$

with

$$\hat{N} = \frac{1}{M} \sum_{m=1}^{M} [f(x_m) \exp(E_\theta(x_m))/q(x_m)] , \; \hat{N} = \frac{1}{M} \sum_{m=1}^{M} [\exp(E_\theta(x_m))/q(x_m)]$$

where $\{x_m\}$ are drawn i.i.d. from q. Apply the importance sampling estimator to $\nabla_\theta A(\theta)$ using $M = 5000$ points and q the ground-truth GMM model [Note: you can use the same sample of q for all gradient steps, so you only need to sample once.] [20pt]

7. Using importance sampling to estimate $A(\theta)$, evaluate the test likelihood of your EBM and compare with the kernel-density estimator. [10pt]

2 Variational Inference

In this exercise, we will verify some properties of variational inference. The setup is a mixture model of the form

$$p(x|\theta) = \int_{\mathcal{Z}} p(x|z,\theta) d p_0(\theta) ,$$

where $\theta \in \Theta$ are model parameters, $z \in \mathcal{Z}$ are latent variables defined over a generic domain, and p_0 is a prior distribution over latent variables. For any probability distribution with positive density q over $\mathcal{Z}$, recall the *Variational Lower Bound*

$$\mathcal{L}(q,\theta) := \mathbb{E}_q \log \left(\frac{p(x,z|\theta)}{q(z)} \right) ,$$

which satisfies $\log p(x|\theta) \geqslant \mathcal{L}(q,\theta)$. q is referred as *variational distribution*.

1. Show that $\log p(x|\theta) = \mathcal{L}(q,\theta) + D_{KL}(q||p(z|x,\theta))$. Use this result to argue that maximizing the Variational Lower bound with respect to q is equivalent to minimizing the KL divergence (also w.r.t. q). [10pt]

2. Draw two one-dimensional probability densities p and q such that $D_{KL}(p||q) \gg D_{KL}(q||p)$. Use a multimodal density for p and interpret the conditions on the variational distribution that result in a tighter bound. [10pt]

3. Using the example of the Gaussian Mixture Model from the previous exercise (in which z are the mixture assignments and $x|z = k \sim \mathcal{N}(\mu_k, \Sigma_k)$, use the variational lower bound to estimate $\log p(x)$ using two choices for q: (i) $q = \text{Unif}[1,\ldots,K]$, and (ii) $q = p(z|x,\theta)$, the posterior distribution, using the true parameters of the model. Interpret the results. [20pt]

4. What happens with the variational lower bound as the number of mixture components $K \to \infty$ using each of the previous two choices? Interpret your answer [10pt]