Advanced Deep Representation Learning: Major Exam

Instructor - Prathosh A. P.

8th December 2022

Total Marks: 30, Time: 180 Minutes, Honor code: Needless to mention, any form of cheating will lead to zero marks. **Please write your name, entry number and branch.**

1. Consider that you are supposed to optimize the negative log likelihood of the data sampled from an unknown distribution p_{data} in a VAE setup. Answer the following:
   (a) Construct the objective function in terms of the divergence between the aggregated posterior of the $q_\phi(z) = \int_x q_\phi(z|x)p_{data}dx$ and the latent prior $p_\theta(z)$ and conditional data likelihood under the decoder $p_\theta(x|z)$. Give a method to optimize this objective and argue why should this be better than optimizing the usual VAE objective (2)
   (b) Show that the optimal value of the ELBO is equal to the negative of the data entropy (1)
   (c) Suppose p_{data} is not-Gaussian, then show that it is impossible to reach the optimum value (derived above) for the ELBO with a Gaussianity assumption on the latent prior $p_\theta(z)$ (1)
   (d) Suppose we have to compute the following gradient - $\nabla_\theta E_{p(z;\theta)}[f(z)]$ where $f(z)$ is an arbitrary function. Show that this is equivalent to computing - $E_{p(\epsilon)}[\nabla_\theta f(g(\epsilon,\theta)]$ where $p(\epsilon)$ is an arbitrary distribution and $g()$ is a function taking $p(\epsilon)$ to $p(z)$. Taking $p(z)$ as an exponential distribution with $\lambda = 2$ and $p(\epsilon) \sim U[0,1]$, find $g()$ (2).
2. Consider GAN with JS divergence as a generative model with a unimodal input distribution.
   (a) Show/argue why it is possible/impossible to learn to sample from a distribution that has multiple disjoint supports (1)
   (b) Describe the problem of mode-collapse in a GAN (1)
   (c) Suppose we change the GAN architecture as following - Instead of having one generator and a discriminator, lets have k generators with a single discriminator (D) that outputs a $k+1$ softmax probabilities. Let D_i, $i \in \{1,2,...k+1\}$ represent the output of the i^{th} node of the D network with $D_k(x)$, $\forall i \in \{1,..k\}$ representing the probability that the sample x is coming from the k^{th} generator and $D_{k+1}(x)$ is the probability of x is coming from the true data distribution. Effectively, this leads to the modification of the GAN objective as follows - The joint objective of all generators for a fixed D will be

$$E_{x\sim p_d} log D_{K+1}(x) + \sum_{i=1}^{k} E_{x\sim p_{g_i}} log(1 - D_{k+1}(x))$$

   Similarly, for fixed set of generators, the objective function for the D would be as follows

$$E_{x\sim p_d} log D_{k+1}(x) + \sum_{i=1}^{k} E_{x_i\sim p_{g_i}} log D_i(x_i)$$

   where $\sum_{i=1}^{k+1} D_i(x) = 1$ and $D_i(x) \in [0,1], \forall i$. With these, show the following:
   (d) For fixed Generators, show that the optimal D is given by below (1)

$$D_{k+1}(x) = \frac{p_d(x)}{p_d(x) + \sum_{i=1}^{k} p_{g_i}(x)}$$

$$D_i(x) = \frac{p_{g_i}(x)}{p_d(x) + \sum_{i=1}^{k} p_{g_i}(x)}$$

(e) For the optimal D, the objective to train generators boils down to

$$KL\left(p_d(x)||p_{avg}(x)\right) + k \times KL\left(\frac{1}{k}\sum_{i=1}^{k} p_{g_i}(x)||p_{avg}(x)\right) - (k+1)log(k+1) + klogk$$

where $p_{avg}(x) = \frac{p_d(x)+\sum_{i=1}^{k} p_{g_i}(x)}{k+1}$ (2)

(f) The global minimum is achieved when $p_d = \frac{1}{k}\sum_{i=1}^{k} p_{g_i}$ and find the value of the objective function. Argue that this formulation would avoid mode-collapse (1)

3. Define an EBM and show the following:
   (a) The gradient of the log-likelihood does not require the computation of partition function (1)
   (b) Explain briefly the idea behind contrastive divergence (Need not derive the Loss expression, you may explain in words with one or two equations) (1)
   (c) Define Fisher divergence between the true and the model densities and show that it can be computed without the access to true data density (you may consider a one dimensional data space) (2)
   (d) Let $s(\theta) = \nabla_x logp(x|\theta)$ define the score function for a model, $p(x|\theta)$. The Fisher information matrix (FIM) is defined to be the covariance of the score function:

$$F(\theta) = \mathbb{E}_{x\sim p(x|\theta)}[\nabla_x log\ p(x|\theta)\ \ \nabla_x log\ p(x|\theta)^T]$$

   Also, let $NLL(\theta) = -log\ p(D|\theta)$ denote the negative log likelihood of the dataset D with N iid samples $D = \{x_i, i = 1:N\}$. With these, first show that the expected value of the score function $s(\theta)$ is zero and use this result to show that he FIM is equal to the expected Hessian of the NLL (2)

4. Write down the definitions all the distributions involved in a DDPM. Define forward and reverse diffusion processes (1). Starting from the data log likelihood, derive the loss function involving KL divergences between different distributions (2) Explain how each of the KL is computed. Derive the final loss function as a regression problem over noise and give out a network diagram to implement it (2). Specify the sampling procedure for a DDPM during inference (1)

5. Answer the following questions:
   (a) Formulate Noise Contrastive Estimation (NCE) and show that the objective function boils down to

$$J(f) = 1/2\mathbb{E}\Big\{log[r(f(x) - logp_n(x))] + log[1 - r(f(y) - logp_n(y))]\Big\}$$

   where $r(x) = (1+e^{-x})^{-1}$, $f = logp_m(;\theta)$, the log likelihood of the model, the expectation is with respect to the jointly distribution of $x \sim p_d$ (true data distribution) and $y \sim p_n$ (artificial noise distribution). Subsequently, show that optimal $J(f)$ occurs when $f = log\ p_d$. Hint: Euler Lagrangian Equation: $\frac{\delta L}{\delta f} - \frac{d}{dx}\frac{\delta L}{\delta f'} = 0$ (3)
   (b) Suppose we have a labeled training set from a source distribution $p(x,y)$ which we use to fit a predictive model $p(y|x)$. At test time, we encounter data from the target distribution $q(x,y)$ with $p \neq q$. The data is said to have co-variate shift if $q(x,y) = q(x)p(y|x))$. Suppose our goal is to minimize the risk on the target distribution q, which can be computed using $R(f,q) = \mathbb{E}_{q(x,y)}[l(y,f(x))]$ where l is a loss function and $f(x)$ is the model output. With these, show that the the target risk $R(f,q)$ can be computed in terms of loss computed on the source data weighted by the ratio of the target and the source data marginal distributions, $q(x)$ and $p(x)$, respectively (1)
   (c) Describe the principle of adversarial domain adaptation briefly. Suppose you are not given the source data during adversarial adaptation (it can be used for classifier pre-training but not during adaptation), how would you perform UDA? (2)