E9-333: Advanced Deep Representation Learning

Instructor: Prathosh A. P.

7th December 2023

Total Marks: 50

Time: 180 Minutes

1. Suppose you are given N binary images(that only can take one of two discrete values) each in $\mathbb{R}^D$, drawn i.i.d from an unknown distribution. Suppose it is desired to model the underlying density function as a mixture of K Bernoulli distributions.
   (a) Formulate and write down the log likelihood for the dataset and count the number of parameters. Assume each data dimension to be independent of each others (2)
   (b) Derive the expression for iterative estimates for the parameters of the aforementioned model. (5)
   (c) Can the above model be used for data clustering? If so how (1)

2. Define a latent variable probabilistic model.
   (a) Derive a lower bound on the log likelihood of a latent variable model by minimization of a divergence measure between the model latent posterior conditioned on data and variational posterior. (2)
   (b) Reformulate the above lower bound in terms of three terms involving (a) average data reconstruction term, (b) KL between the the aggregated latent posterior - $q(z) = \frac{1}{N}\sum_{n=}^{N} q_\phi(z|x_n)$ and latent prior $p(z)$ and (c) a mutual information term between the sample index n(treated as a random variable) and the latent variable z.Use this result to argue how does the three terms mentioned affect the modeling behavior. Hint: Mutual Information between $n$and z is given by $\mathbb{I}_{q(n,z)}[n, z] = \mathbb{E}_{q(n,z)}\left[log\{\frac{q(n,z)}{q(n)q(z)}\}\right]$, where $q(n, z)$, $q(z)$ and $q(n)$ are respectively the joint distribution between sample index and latent variable, aggregated posterior and marginal of sample index $q(n) = \frac{1}{N}$, N being the number of data samples. (5)
   (c) Derive the optimal prior $p(z)$ to optimize the lower bound and describe a method to realize it in practice. (3)

3. Consider a Generative Adversarial Network that has K generators $G_{1:K}$ instead of one. Each generator G_k maps z to $x = G_k(z)$, thus inducing a single distribution P_{G_k} ; and K generators altogether induce a mixture over K distributions, denoted byP_θ in the data space.The final generated sample is taken to be $G_u(z)$, where index u is drawn from a fixed multinomial distribution $Mult(\pi)$ where $\pi = [\pi_1, \pi_2, \cdots, \pi_k]$. In addition to the usual discriminator D, that distinguishes between the generated and the training samples, There is an additional classifier C that performs multi-class classification to classify samples labeled by the indices of their corresponding generators. With this, the it is sought to optimize the following optimization problem:

$$\underset{C,G_{1:K}}{Min}\ \underset{D}{Max}\ \left\{\mathbb{E}_{x\sim P_{data}}[logD(x)] + \mathbb{E}_{x\sim P_\theta}[log(1 - D(x))] - \beta\left\{\sum_{k=1}^{K}\pi_k\mathbb{E}_{P_{G_k}}[logC_k(x)]\right)\right\}$$

here, $C_k(x)$ is the is the likelihood that x is generated by G_k.

(a) With this setup, Show that the optimal classifier $C^* = C^*_{1:K}$ and the optimal discriminator D^* are given by $C^*_k(x) = \frac{\pi_k P_{G_k}(x)}{\sum_i \pi_i P_{G_i}(x)}$ and $D^*(x) = \frac{P_{data}(x)}{P_{data}(x)+P_\theta(x)}$ when the generators are fixed. (3)

(b) Show that: at the equilibrium point as given in (a), the optimal generators are obtained when the JSD between P_{data} and P_θ is minimized while the JSD among $P_{G_1}, P_{G_2}, ..., P_{G_K}$ is maximized. Hint: JSD $(P_1, P_2, ..., P_K) = \sum_i \pi_i D(P_i||M)$ where $M = \sum_i \pi_i P_i$. (3)

(c) If the true data distribution has the form $P_{data}(x) = \sum_{k=1}^{K} \pi_k q_k(x)$ where $q_k(x)$ are distributions with non-overlapping supports, the show that the optimization problem will ensure that $P_\theta = P_{data} = \sum_{k=1}^{K} \pi_k q_k(x)$. (3)

(d) Argue how and why this GAN formulation is helpful in avoiding mode-collapse. (2)

(e) **Note: For all the above proofs assume that the second derivative is absent in the Euler-Lagrangian equations occurring during function optimization**

4. Diffusion models (use usual notations all through):

   (a) Define (mathematically) the forward and the model distributions in a DDPM. (2)

   (b) Show that sampling from a t^{th} forward step can be accomplished directly from the initial point without having to carry out t diffusion steps. (2)

   (c) In a DDPM, denoising matching term is given by the KL divergence between $q(x_{t-1}|x_t, x_0)$ and $p_\theta(x_{t-1}|x_t)$ which are both assumed to be Gaussian distributions with different means μ_q and μ_θ respectively and same variance Σ_q. Here $\mu_q = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\bar{\alpha}_t}$ and $\Sigma_q = \sigma_q^2 \mathcal{I}$ where $\sigma_q^2 = \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}$. With these, answer the following:

      i. Derive an expression for denoising matching term as a scaled regression problem, regressing on the true data point x_0, via a modeled point x_θ. (2)

      ii. Show that the scaling term in the above regression problem can be expressed in terms of signal to noise ratio (SNR) term, defined as the ratio of the squared of the mean and variance of the forward distribution at all t. Argue what should be the behavior of the this SNR term and suggest any modification in the standard DDPM implementation to emulate the required behavior of the SNR term. (3)

      iii. How is conditional generation achieved in a DDPM? Explain with the appropriate mathematical justification. (2)

5. Mixed bag:

   (a) Define an an EBM and argue (mathematically) why ML estimation is not preferred with them. (2)

   (b) Demonstrate two non-generative modeling use cases for adversarial divergence minimization. (2)

   (c) Suppose in a $K-$class supervised learning problem, instead of a single hard label per data point, the posterior of labels given each data point $(p(y|x_i)$ is known. Can you suggest an alternative to usual Empirical Risk Minimization, by the use of the label posterior, that results in a 'better' model estimation. Justify your answer formally. (4)

   (d) Formulate a question on the course content and answer it. (2)