Indian Statistical Institute

Mid Semestral Examination: Machine Learning II

Maximum Marks: 60, Time: 2 Hrs. 15 Mins.

September 10, 2025

Answer ALL questions. The maximum score you can obtain is 60.

1. Let $p(k)$ be a one-dimensional discrete distribution that we wish to approximate, with support on nonnegative integers. One way to fit an approximating distribution $q(k)$ is to minimize the Kullback-Leibler (KL) divergence $KL(p||q)$. Show that when $q(k)$ is a Poisson distribution, this KL divergence is minimized by setting λ to the mean of $p(k)$. [10]

2. Let $\{x^{(i)}\}_{i=1}^N$ be data samples from an unknown distribution. A Variational Autoencoder (VAE) models latent variables $z \in \mathbb{R}^k$ with prior

$$p(z) = \mathcal{N}(z; 0, I),$$

and likelihood

$$p_\theta(x|z) = \mathcal{N}(x; f_\theta(z), \sigma^2 I),$$

where f_θ is a neural network (decoder). Since the true posterior $p_\theta(z|x)$ is intractable, we introduce a variational approximation

$$q_\phi(z|x) = \mathcal{N}(z;\, \mu_\phi(x), \mathrm{diag}(\sigma_\phi^2(x))),$$

with parameters ϕ (encoder network).

(a) Show that the marginal log-likelihood of a data point x can be decomposed as

$$\log p_\theta(x) = \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] + \mathrm{KL}\big(q_\phi(z|x) \,\|\, p_\theta(z|x)\big).$$

Conclude that maximizing the expected log-likelihood is equivalent to maximizing the Evidence Lower Bound (ELBO):

$$\mathcal{L}(\theta, \phi; x) := \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathrm{KL}(q_\phi(z|x) \,\|\, p(z)).$$

(b) Prove that the ELBO is indeed a lower bound:

$$\mathcal{L}(\theta, \phi; x) \;\le\; \log p_\theta(x),$$

and equality holds if and only if $q_\phi(z|x) = p_\theta(z|x)$ almost everywhere.

(c) Using the *reparameterization trick* $z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, prove that the gradient of the ELBO w.r.t. θ and ϕ can be written as an expectation with respect to ϵ, thus avoiding direct differentiation through the sampling step.

[5 + 6 + 7 = 18]

3. Consider a very simple RNN with one hidden unit and scalar input defined by

$$h_t = \tanh(w_h h_{t-1} + w_x x_t), \qquad y_t = w_y h_t,$$

where w_h, w_x, w_y are scalar weights, $h_0 = 0$, and $\tanh(z)$ is the hyperbolic tangent activation. Suppose the parameters are

$$w_h = 0.5, \quad w_x = 1.0, \quad w_y = 2.0,$$

and the input sequence is

$$x_1 = 1, \quad x_2 = -1, \quad x_3 = 2.$$

The target at the final time step is $y^* = 1$ and the loss is

$$L = \frac{1}{2}(y_3 - y^*)^2.$$

Compute the partial derivatives

$$\frac{\partial L}{\partial w_h} \qquad \text{and} \qquad \frac{\partial L}{\partial w_x},$$

showing the intermediate backward-pass quantities (e.g. $\delta_t = \partial L / \partial a_t$ with $a_t = w_h h_{t-1} + w_x x_t$) and the final numerical values. [6 + 6 = 12]

4. Consider the forward (noising) process used in denoising diffusion probabilistic models (DDPM). Let $x_0 \in \mathbb{R}^d$ be a data point, and let a fixed sequence $\{\beta_t\}_{t=1}^T$ satisfy $0 < \beta_t < 1$. Define

$$\alpha_t := 1 - \beta_t, \qquad \bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s.$$

The forward diffusion (a Markov chain) is defined by

$$q(x_t \mid x_{t-1}) = \mathcal{N}\big(x_t;\ \sqrt{\alpha_t}\, x_{t-1},\ \beta_t I\big), \qquad t = 1, \ldots, T.$$

(a) Show that the marginal distribution $q(x_t \mid x_0)$ has the closed form

$$q(x_t \mid x_0) = \mathcal{N}\big(x_t;\ \sqrt{\bar{\alpha}_t}\, x_0,\ (1 - \bar{\alpha}_t) I\big).$$

(b) Using the result of (a), prove that x_t can be expressed explicitly as

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I).$$

(c) Show that the conditional posterior distribution is Gaussian:

$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}\big(x_{t-1};\ \mu_q(x_t, x_0),\ \tilde{\beta}_t I\big),$$

where

$$\mu_q(x_t, x_0) = \frac{\beta_t \sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_t}\, x_0 + \frac{\sqrt{\alpha_t}\,(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}\, x_t,$$

and

$$\tilde{\beta}_t = \frac{\beta_t (1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}.$$

[5 + 7 + 8 = 20]

5. Let us consider a Generative Adversarial Network (GAN) consisting of a Generator G and a Discriminator D. Let us also take the original data distribution as p_{data}, generated by G as p_g, and that of the noise as p_z. With these settings, if we replace the traditional Binary Cross Entropy loss in GANs with the Mean Squared Error loss then the modified objective function becomes as follows:

$$min_D = \frac{1}{2}\mathbb{E}_{x\sim p_{data}}\Big(D(x)-b\Big)^2 + \frac{1}{2}\mathbb{E}_{z\sim p_z}\Big(D(G(z))-a\Big)^2,$$
$$min_G = \frac{1}{2}\mathbb{E}_{z\sim p_z}\Big(D(G(z))-c\Big)^2.$$

Let us call this an MSE (Mean Squared Error)-GAN, where a, b, and c, respectively denote the generated data label, the real data label, and the label of the data that the generator wants the discriminator to believe.

- What can be an intuition behind replacing the Binary Cross Entropy loss with a Mean Squared Error?
- There exists a family of divergence measures called f-divergence. The general form of f divergence between a couple of distributions P and Q are defined as:

$$D(P||Q) = \int f\Big(\frac{p(x)}{q(x)}\Big)q(x)dx.$$

Show that the KL divergence is actually a special case of f-divergence for a particular form of the function f, and also derive the form of this generating function.

- The χ^2 divergence is defined as $\chi^2(P||Q) = \int \frac{(p(x)-q(x))^2}{q(x)}dx$. Show that this divergence is a member of the f-divergence family and also derive the corresponding generating function f.
- If we impose the constraints $b-c=1$ and $b-a=2$ then can you show that the objective function of MSE-GAN can also be reduced to a χ^2 divergence between $p_{data}+p_g$ and $2p_g$ in a manner similar to the canonical GAN? $[3+4+6+7=20]$