X

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

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

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:

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

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

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

Compute the partial derivatives

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

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

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

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

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

(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:

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

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]

Indian Statistical Institute

End Semester Examination

M. Tech. (CS), Machine Learning II

Maximum Marks: 100 **Duration: 3 Hrs. 15 Mins.**

Note:

- Attempt all questions.
- Assume suitable values wherever necessary and state assumptions clearly.
- All notations follow standard definitions used in deep learning and transformer architectures.
- Use of calculator is permitted.

---

1. Consider a linear autoencoder defined as follows:

$$x \in \mathcal{R}^n, \quad h = W_1 x, \quad \hat{x} = W_2 h,$$

where $W_1 \in \mathcal{R}^{m \times n}$ is the encoder weight matrix and $W_2 \in \mathcal{R}^{n \times m}$ is the decoder weight matrix.

(a) If the hidden layer dimension $m = n$, determine what the product $W_2 W_1$ must be for perfect reconstruction of all inputs x.

(b) If $m < n$, explain mathematically why perfect reconstruction of all inputs is not possible, and describe what $W_2 W_1$ represents in this case.

(c) Briefly interpret the geometric meaning of part (b): what kind of subspace does the autoencoder learn when $m < n$?

[3 + 3 + 2 = 8]

2. Let $p(k)$ be a one-dimensional discrete probability distribution defined on nonnegative integers $k = 0, 1, 2, \ldots$, which we wish to approximate by a Poisson distribution $q(k; \lambda)$ having parameter $\lambda > 0$. The Poisson distribution is given by

$$q(k; \lambda) = \frac{e^{-\lambda} \lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

(a) Show that minimizing the KL divergence $KL(p \| q_\lambda)$ with respect to λ leads to

$$\lambda^* = \sum_{k=0}^{\infty} k\, p(k),$$

that is, the optimal Poisson parameter equals the mean of $p(k)$. Provide all intermediate steps clearly.

(b) Explain intuitively why this result implies that the "best" Poisson approximation to any discrete nonnegative distribution matches its expected value.

(c) Show that $KL(p||q_\lambda)$ is a convex function of λ, and hence verify that the stationary point $\lambda^* = E_p[k]$ corresponds to a unique global minimum.

[6 + 3 + 3 = 12]

3. (a) A simple Recurrent Neural Network (RNN) operates on an input sequence

$$x = [x_1, x_2] = [1, 2],$$

with parameters:

$$W_x = 1.0, \quad W_h = 0.5, \quad b = 0, \quad \text{and activation } f(z) = \tanh(z).$$

Assume the initial hidden state $h_0 = 0$.

i. Compute h_1 and h_2 step-by-step, showing all intermediate values before and after applying tanh.

ii. If the output is computed as $y_t = h_t$, write the final output sequence $[y_1, y_2]$.

(b) Consider an LSTM cell with the same input sequence $x = [1, 2]$ and the following parameters:

$$W_i = W_f = W_o = W_c = 1, \quad U_i = U_f = U_o = U_c = 0, \quad b_i = b_f = b_o = b_c = 0,$$

and the initial states $h_0 = 0$, $c_0 = 0$. Use the activation functions $\sigma(z) = \frac{1}{1+e^{-z}}$ and $\tanh(z)$.

i. Compute the gate activations (i_t, f_t, o_t) and candidate cell state $(\tilde{c}_t)$ for $t = 1, 2$.

ii. Compute the cell states c_1, c_2 and hidden states h_1, h_2.

iii. Compare how the LSTM's memory cell c_t differs from the RNN's hidden state in preserving information across time steps.

[(4+4)+(4+5+3) = 20]

4. A Transformer's Scaled Dot-Product Attention mechanism operates using three matrices: Query (Q), Key (K), and Value (V). You are given the following input matrices for a single attention head:

$$Q = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}, \quad K = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}, \quad V = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

(a) Compute the unnormalized attention scores.

(b) Apply the appropriate scaling factor to the computed scores to obtain the scaled attention scores.

(c) Calculate the normalized attention weights using the Softmax function applied to the scaled scores.

(d) Using the obtained attention weights, compute the final attention output vector as a weighted sum of the value vectors.

[4 + 5 + 4 + 7 = 20]

5. Explainable AI techniques such as LIME and SHAP help interpret the contribution of individual input features to a model's prediction. Consider a simple linear regression model used for predicting a student's exam score based on two features:

$$f(x_1, x_2) = 5 + 3x_1 + 2x_2$$

where x_1 = hours studied and x_2 = number of practice tests taken. We wish to explain the prediction for a specific student with input $(x_1, x_2) = (2, 1)$.

(a) Using the SHAP framework, assume the following coalition function values (model outputs when only a subset of features is present):

$$f(\emptyset) = 5,$$
$$f(\{x_1\}) = 11,$$
$$f(\{x_2\}) = 9,$$
$$f(\{x_1, x_2\}) = 15.$$

Compute the Shapley values ϕ_{x_1} and ϕ_{x_2} for this prediction and verify that:

$$f(x_1, x_2) = f(\emptyset) + \phi_{x_1} + \phi_{x_2}.$$

(b) Interpret the computed Shapley values: Which feature contributes more to the model's output for this particular prediction, and by how much relative to the baseline?

[5+ 3 = 8]

6. Let $s \in R^n$ be a vector of attention scores with components s_j, and define the *softmax* function

$$a = \text{softmax}(s), \qquad a_i = \frac{e^{s_i}}{\sum_{t=1}^{n} e^{s_t}}, \qquad i = 1, \ldots, n.$$

The vector a represents the attention weights corresponding to s.

(a) Prove that the softmax function is invariant to adding the same constant to all entries of s; that is, for any scalar $c \in R$,

$$\text{softmax}(s + c\mathbf{1}) = \text{softmax}(s),$$

where $\mathbf{1}$ denotes the all-ones vector.

(b) Derive the *Jacobian matrix* of the softmax function, defined as

$$J_s = \frac{\partial a}{\partial s} \in R^{n \times n}, \quad \text{where} \quad (J_s)_{ij} = \frac{\partial a_i}{\partial s_j}.$$

Show that $J_s = \text{diag}(a) - aa^\top$ and hence that for all i, j,

$$\frac{\partial a_i}{\partial s_j} = a_i(\delta_{ij} - a_j),$$

where δ_{ij} is the Kronecker delta.

(Hint: Use the quotient rule starting from $a_i = e^{s_i} / \sum_t e^{s_t}$.)

[4+6=10]

7. (a) Consider a graph G with node indices $\{0, 1, 2, 3, 4\}$ and edge indices given by:

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 2 & 2 & 3 & 3 & 3 & 4 \\ 1 & 2 & 3 & 0 & 0 & 3 & 0 & 2 & 4 & 3 \end{bmatrix}$$

Draw the computation graph for node ID 2 for a 2-hop message passing operation. Assume the node feature matrix of G is:

$$X = \begin{bmatrix} 0 & -1 \\ 1 & -2 \\ 0 & 2 \\ 1 & 0 \\ -1 & 3 \end{bmatrix}$$

Estimate the aggregated features for node 2 using a bottom-to-top approach with the mean aggregator (no self-loop addition required). Further, compute A^2X, where A is the normalized adjacency matrix, to aggregate two-hop node features. Show that the updated node features for node 2 are equal under both aggregation approaches.

(b) Two nodes p and q in a connected graph respectively have degrees m and n. Assume all nodes possess identical features. If neighborhood aggregation is performed, design two different aggregators that yield distinct neighborhood representations for nodes p and q under the following cases:

(a) $m = n$

(b) $m \neq n$

(c) Let L be a graph Laplacian with two eigenvectors

$$u = \begin{bmatrix} -1 & 3a & b \end{bmatrix}, \quad v = \begin{bmatrix} 2b & c & -a \end{bmatrix}$$

corresponding to eigenvalues greater than zero, where $a, b, c \in R$. Find the values of a, b, and c that satisfy orthogonality and normalization conditions of Laplacian eigenvectors.

(d) Discuss three advantages of employing **Knowledge Distillation** from a larger (teacher) network to a smaller (student) network, in comparison to designing an end-to-end trainable compact model.

[8 + 4 + 5 + 3 = 20]

8. (a) A binary classifier predicts whether a candidate is *selected* ($\hat{Y} = 1$) or *not selected* ($\hat{Y} = 0$) in a job interview for the position of Data Scientist. The following table shows predictions for 12 candidates, equally divided between males and females.

Candidate ID	Gender	Predicted Outcome ($\hat{Y}$)
1	Female	0
2	Male	1
3	Male	1
4	Female	1
5	Male	1
6	Female	0
7	Female	0
8	Male	0
9	Female	1
10	Male	1
11	Female	0
12	Male	1

i. Compute the *Demographic Parity Difference (DPD)* between male and female candidates.

ii. Interpret your result: which gender group receives more favorable predictions?

(b) A probabilistic language model assigns conditional probabilities to each token in a sequence $(w_1, w_2, \ldots, w_N)$. The perplexity (PP) of the model is defined as:

$$\mathrm{PP} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i \mid w_1, \ldots, w_{i-1})\right).$$

i. Derive this expression starting from the cross-entropy formulation

$$H(p,q) = -\frac{1}{N}\sum_{i=1}^{N}\log q(w_i).$$

ii. Compute the perplexity for a 4-token sequence where the model assigns probabilities $P(w_1) = 0.4$, $P(w_2|w_1) = 0.3$, $P(w_3|w_1, w_2) = 0.2$, and $P(w_4|w_1, w_2, w_3) = 0.1$.

iii. Interpret what this perplexity value implies about the model's confidence and prediction quality.

[(4+2)+(2+2+2) = 12]