Indian Statistical Institute

Mid Semestral Examination: Machine Learning II

Maximum Marks: 60, Time: 2 Hrs.

September 16, 2024

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

1. Given a dataset $\{x^{(i)}\}_{i=1}^{N}$ with observations assumed to be generated from some unknown distribution $p_{\text{data}}(x)$, define the generative process of the VAE using a latent variable z with a prior distribution $p(z)$. The generative model is parameterized by θ, and the encoder (recognition model) is parameterized by ϕ.

   - Derive the Evidence Lower Bound (ELBO) from the marginal likelihood $p_\theta(x) = \int p_\theta(x|z)p(z)\,dz$. Show how the ELBO can be expressed as:

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

   - Discuss the reparameterization trick used in VAEs to allow for backpropagation through the stochastic latent variable z. Specifically, explain how z is sampled using a differentiable transformation:

   $$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon \quad \text{where} \quad \epsilon \sim \mathcal{N}(0, I)$$

   - Why is the reparameterization trick crucial for training VAEs using gradient-based optimization methods?

   [6 + 5 + 4 = 15]

2. 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]

3. Given the gradient calculated at a point, the Adam optimizer has three distinct steps. First, update the moving averages. Second, apply the bias correction. Third, update the parameters. Consider the moving average of the square of the gradients. It is given by the recursive formula:

   $$S_t = \beta_2 S_{t-1} + (1 - \beta_2) g_t^2.$$

   (i) Write down the expression for st only in terms of the gradients $g_0, g_1, \ldots\ldots, g_t$

   (ii) Given your expression in part (i), what is $E[S_t]$ in terms of $E[g_t^2]$ and β_2? You may assume that g_i's are independent and identically distributed. [5 + 5 = 10]

4. (a) Let $z^{(i)} \in \{0, 1\}^K$ be an indicator vector, such that $z_k^{(i)} = 1$ if $x^{(i)}$ was drawn from Bernoulli(p^k), and 0 otherwise. Let $Z = \{z^{(i)}\}_{i=1,\ldots n}$. Write down the expressions for $P(z^{(i)}|\pi)$ and $P(x^{(i)}|z^{(i)}, p, \pi)$.

   (b) Using the above two quantities, derive the likelihood of the data and the latent variables, $P(Z, X|\pi, p)$.

(c) Let $\eta(z_k{}^{(i)}) = E[z_k{}^i|x^{(i)}, \pi, p]$. Show t at,

$$\eta(z_k{}^{(i)}) = \frac{\pi_i \prod_{d=1}^{D}(p_d{}^{(k)})^{x_d{}^{(i)}}(1-p_d{}^k)^{1-x_d{}^{(i)}}}{\Sigma_j \pi_j \prod_{d=1}^{D}(p_d{}^{(j)})^{x_d{}^{(i)}}(1-p_d{}^j)^{1-x_d{}^{(i)}}}$$

Let $\tilde{p}$, $\tilde{\pi}$ be the new parameters that we'd like to maximize, so p, π are from the previous iteration. Use this to derive the follow ng final expression for the E-step in the expectation-maximization algorithm. SHow all yo r steps.

$$\mathbb{E}[\log P(Z, X|\tilde{p}, \tilde{\pi})|X, p, \pi] = \Sigma_{i=1}{}^N \Sigma_{k=1}{}^K \eta(z_k{}^{(i)})[\log \tilde{\pi}_k + \Sigma_{d=1}{}^D (x_d{}^{(i)} \log \tilde{p_d}{}^{(k)} + (1-x_d{}^{(i)}) \log(1-\tilde{p_d}{}^{(k)}]$$

(d) We need to maximize the above expre sion with respect to $\tilde{\pi}$, $\tilde{p}$. Show that the value of $\tilde{p}$ that maximizes the E-step is

$$\tilde{p}^{(k)} = \frac{\Sigma_{i=1}^N \eta(z_k{}^{(i)}) x^{(i)}}{N_k},$$

where $N_k = \Sigma_{i=1}{}^N \eta(z_k{}^{(i)})$

(e) Show that the value of $\tilde{\pi}$ that maximizes the E-step is

$$\tilde{\pi_k} = \frac{N_k}{\Sigma_{k'} N_{k'}}.$$

The exponential families notation may be useful. Alternatively, you can use Lagrange multipliers. $[4+3+4+4+5=20]$

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? $[3+4+6+7=20]$

Part II: Answer each of the following questions

21. Provide answers with brief explanations for each of the following questions.

    (a) In the multi-arm-bandit problem, which value of ε among $0, 0.1, 0.01$ and 1 would produce best result with sufficiently large number of actions, and why?

    (b) Recall the definition of λ-return:

$$G_t^{\lambda} == (1-\lambda) \sum_{n=1}^{T-t-1} \lambda^{n-1} G_{t:t+n} + \lambda^{T-t-1} G_t.$$

    Which value of λ corresponds to TD-0 and which value of λ corresponds to MC?

    (c) Suppose the environment in an episodic task provides reward only at the terminal state. Which method among tabular MC and tabular SARSA will be able to receive updates for more state-action pairs after the first completed episode?

    (d) Why does Q-learning become challenging in case of continuous action space?

    $(2.5 \times 4 = 10$ marks$)$

22. Highlight the main differences between tabular methods and function approximation methods on the following aspects: (a) representation of states, (b) scalability, (c) generalization ability, (d) convergence and stability, and (e) sample efficiency (number of samples required for successful training).

    $(2 \times 5 = 10$ marks$)$

23. Assume that a globally trained movie recommender system (defined by a set of global parameters) needs to be personalized for each user using reinforcement learning. Every week, the system recommends 5 movies to the user. Depending on whether the user views them or not, the system receives a reward and updates the parameters accordingly. Identify an algorithm of your choice and outline the agent that learns the parameters for a single user using a pseudo-code. Include the update formulae as much as possible and write comments with the pseudo-code as explanation.

    (10 marks)

Indian Statistical Institute

M. Tech. (CS) II

Machine Learning 2

End Semestral Examination, Date: November 25, 2024

Timing: 3 hrs. 30 mins | Maximum marks: 100

You may attempt ALL the questions. The maximum score achievable is 100. Use of a calculator is permitted.

1. Given an LSTM cell with standard equations for Forget Gate: f_t, Input Gate: i_t, Candidate Cell State: $\tilde{C}_t$, Cell State Update: C_t, Output Gate: o_t, and the Hidden State Update: h_t.

   Assume:

   - Input $x_t = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$
   - Previous hidden state $h_{t-1} = \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix}$
   - Previous cell state $C_{t-1} = \begin{bmatrix} 0.2 \\ -0.2 \end{bmatrix}$

   The weight matrices and biases are given as:

$$W_f = \begin{bmatrix} 0.4 & -0.2 & 0.3 & 0.1 \\ 0.1 & 0.3 & -0.4 & 0.2 \end{bmatrix}, \quad b_f = \begin{bmatrix} 0.1 \\ 0.2 \end{bmatrix}$$

$$W_i = \begin{bmatrix} 0.3 & 0.2 & 0.1 & -0.3 \\ -0.3 & 0.4 & 0.5 & 0.2 \end{bmatrix}, \quad b_i = \begin{bmatrix} -0.1 \\ 0.1 \end{bmatrix}$$

$$W_C = \begin{bmatrix} 0.5 & -0.4 & 0.2 & 0.3 \\ 0.2 & 0.3 & 0.1 & -0.1 \end{bmatrix}, \quad b_C = \begin{bmatrix} 0 \\ 0.1 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.2 & 0.3 & 0.1 & -0.2 \\ -0.2 & 0.4 & -0.3 & 0.3 \end{bmatrix}, \quad b_o = \begin{bmatrix} 0.1 \\ -0.1 \end{bmatrix}$$

   Calculate the following values for the LSTM cell at time step t and show all your work:

   - Updated cell state C_t
   - Output gate o_t
   - Updated hidden state h_t

   Note: Use the *sigmoid* function for nonlinearity.

   (5+5+5=15)

2. Consider the following fictitious word embedding vectors for the four words in a text sequence as inputs to a self-attention layer of a transformer:

$$V_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \; V_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \; V_3 = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}, \; V_4 = \begin{bmatrix} 3 \\ 1 \\ 4 \end{bmatrix}.$$

Assume the weight matrices for query, key, and value vectors i.e. W_q,W_k, and W_v are all identity matrices. Compute the attention score matrix at the output of the self-attention layer. Use the scaled dot product-based similarity and *softmax* normalization for the attention weights. Show all your steps. (15)

3. (a) Consider a graph G containing six nodes marked as $\{1, 2, 3, 4, 5, 6\}$. The feature matrix of G is,

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

The edge list of G is presented as

$$E = \begin{bmatrix} 1 & 1 & 1 & 2 & 3 & 4 & 4 & 5 & 5 & 5 & 6 \\ 2 & 3 & 4 & 1 & 1 & 1 & 5 & 1 & 4 & 6 & 5 \end{bmatrix}$$

A message-passing layer performs a 2-hop neighborhood aggregation on G with two aggregators (1) mean and (2) max pool. Report the updated features of the node 1 and 4 in the G after the message passing operation for each aggregator.

(b) Consider a graph whose symmetrically normalized Laplacian is $\tilde{L}$. If the maximum eigenvalue of $\tilde{L}$ is $\lambda_{\max} = \frac{p}{3-p^2}$ for $p \in \mathbb{R}$, then show that $-2 \leq p \leq \frac{3}{2}$. Under what condition will the equality hold?

(6+4 = 10)

4. Consider a simple RNN with one neuron, where at each time step t, the hidden state h_t is updated using the formula:

$$h_t = \tanh(W \cdot h_{t-1} + U \cdot x_t),$$

where:

- h_t is the hidden state at time t.
- x_t is the input at time t.
- $W = 0.5$ is the recurrent weight.
- $U = 1.0$ is the input weight.

Suppose the RNN receives an input sequence $x = [1, 2, 1]$ over three time steps $t = 1, 2, 3$, and the initial hidden state is $h_0 = 0$.

- Calculate the hidden states h_1, h_2, and h_3 for each time step, showing all intermediate steps.
- Using Backpropagation Through Time (BPTT), compute $\frac{\partial h_3}{\partial W}$ up to two decimal places, showing all intermediate steps.

(6+9 = 15)

5. (a) Consider a dataset $X = \{x_1, x_2, \ldots, x_n\}$ with n observations, where each observation x_i lies in a d-dimensional space. In Kernel PCA, we map the data to a higher-dimensional feature space using a kernel function $k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$, where $\phi(\cdot)$ is an implicit mapping function. The kernel matrix K is defined as:

$$K_{ij} = k(x_i, x_j).$$

- To center the data in the feature space, the kernel matrix K is transformed into the centered kernel matrix K_c. Prove that K_c is given by:

$$K_c = K - \mathbf{1}_{1/n}K - K\mathbf{1}_{1/n} + \mathbf{1}_{1/n}K\mathbf{1}_{1/n},$$

where $\mathbf{1}_{1/n}$ is an $n \times n$ matrix of ones, scaled by $\frac{1}{n}$.

- Prove that the centered kernel matrix K_c is symmetric.

(b) Consider clustering a dataset $X = \{x_1, x_2, \ldots, x_n\}$ with n observations, each of dimension d, into k clusters using the Kernel K-means algorithm. Kernel K-means utilizes a kernel matrix K (of size $n \times n$), where $K_{ij} = k(x_i, x_j)$ represents the similarity between observations x_i and x_j in a high-dimensional feature space. Assuming that the kernel matrix K is precomputed, show that the time complexity of Kernel K-means across T iterations is $O(T \cdot n^2 \cdot k)$.

((6+4)+5 = 15)

6. (a) Consider a Variational Autoencoder (VAE) with a prior distribution $p(z) = \mathcal{N}(0, I)$ and a Gaussian encoder $q(z|x) = \mathcal{N}(z; \mu(x), \sigma^2(x)I)$, where $\mu(x)$ and $\sigma^2(x)$ are the outputs of the encoder's neural network given the input x. The decoder is defined by the likelihood $p(x|z) = \mathcal{N}(x; \tilde{\mu}(z), \tilde{\sigma}^2(z)I)$, where $\tilde{\mu}(z)$ and $\tilde{\sigma}^2(z)$ are the outputs of the decoder's neural network given the latent variable z.

Prove that the Evidence Lower Bound (ELBO) for the VAE objective can be expressed as:

$$\mathcal{L}(x; \theta, \phi) = -\frac{1}{2}\sum_{i=1}^{D}\left(1 + \log(\sigma_i^2(x)) - \mu_i(x)^2 - \sigma_i^2(x)\right) + \mathbb{E}_{q(z|x)}\left[\log p(x|z)\right],$$

where θ and ϕ represent the parameters of the decoder and encoder networks, respectively, and D is the dimensionality of the latent space z.

You may use the fact that the KL divergence between two Gaussian distributions $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ is given by:

$$\mathrm{KL}(\mathcal{N}(\mu_1, \Sigma_1) \| \mathcal{N}(\mu_2, \Sigma_2)) = \frac{1}{2}\left[\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T\Sigma_2^{-1}(\mu_2 - \mu_1) - k + \log\frac{\det\Sigma_2}{\det\Sigma_1}\right].$$

(b) Consider a Denoising Diffusion Probabilistic Model (DDPM) with a forward process that adds Gaussian noise to the data over T time steps. The forward process is defined by:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I),$$

where x_0 is the original data point, x_t is the noisy data at step t, β_t is the variance schedule, and $\mathcal{N}(\cdot; \mu, \Sigma)$ represents a Gaussian distribution with mean μ and covariance Σ.

Given a simple linear variance schedule $\beta_t = \frac{t}{T}$ for $t = 1, 2, \ldots, T$, prove that the mean of the distribution $q(x_t|x_0)$ is:

$$\mathbb{E}[x_t|x_0] \approx \exp\left(-\frac{t(t+1)}{4T}\right)x_0.$$

(8+12= 20)

7. (a) Consider a Generative Adversarial Network (GAN) where the discriminator D and generator G are parameterized by neural networks. Let $P_{\text{data}}(x)$ denote the true data distribution and $P_G(x)$ denote the distribution of generated samples from G.

The f-divergence between two probability distributions P and Q for a convex function $f : (0, \infty) \to \mathbb{R}$ is defined as:

$$D_f(P||Q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right) dx,$$

where $p(x)$ and $q(x)$ are the probability density functions of P and Q respectively.
In the case of GANs, where we aim to minimize the divergence between P_{data} and P_G, the Jensen-Shannon divergence (JS divergence) is often used. However, it can be generalized to other f-divergences.

- Show that if f is chosen such that $f(t) = t \log t - (t+1) \log(t+1)$, the f-divergence becomes the Jensen-Shannon divergence.
- Prove that minimizing the f-divergence $D_f(P_{\text{data}}||P_G)$ with respect to the generator G can be framed as a game where the objective of the discriminator is to maximize the following expression:

  $$\mathbb{E}_{x \sim P_{\text{data}}}[T(x)] - \mathbb{E}_{x \sim P_G}[f^*(T(x))],$$

  where $T(x)$ is a function parameterized by the discriminator, and f^* is the Fenchel conjugate of f.

(5+10= 15)

8. Consider a simple neural network with a single hidden layer and linear activation functions. Let:

- $\mathbf{W}$ be the weight matrix of the hidden layer.
- $\mathbf{b}$ be the bias vector of the hidden layer.
- $\mathbf{u}$ be the weight vector for the output layer.
- $\mathbf{x}$ be the input vector.
- y be the target output.

The output of the network without dropout is given by:

$$\hat{y} = \mathbf{u}^T(\mathbf{W}\mathbf{x} + \mathbf{b}).$$

The loss function (e.g., mean squared error) without dropout is:

$$L(\mathbf{u}, \mathbf{W}, \mathbf{b}) = \frac{1}{2}(\hat{y} - y)^2.$$

During training with dropout, a dropout mask $\mathbf{d}$ is applied to the hidden layer activations. Each element d_i of the dropout mask $\mathbf{d}$ is a Bernoulli random variable with parameter p (the probability of keeping a unit).

- Derive an expression for the expected cost function with dropout applied during training. Assume that each element of the dropout mask $\mathbf{d}$ is an independent random variable.
- Prove that applying dropout during training results in an expected cost function with a regularization term. Specifically, show that the expected cost function with dropout can be interpreted as the original cost function plus a regularization term. In other words, prove that:

$$\mathbb{E}_{\mathbf{d}}[L_{\text{dropout}}(\mathbf{u}, \mathbf{W}, \mathbf{b})] \approx L(\mathbf{u}, \mathbf{W}, \mathbf{b}) + \lambda R(\mathbf{W}),$$

where λ is a regularization parameter and $R(\mathbf{W})$ is a regularization term dependent on the weights.

(4+6= 10)

Best of luck!