Quiz 2

Prof . Prathosh AP and Chandan J

Questions

1. **(GMM E-Step):** A Gaussian Mixture Model has two components, C_1 and C_2, with prior probabilities $\pi_1 = 0.6, \pi_2 = 0.4$. The components are 1D Gaussians with parameters $\mu_1 = 5, \sigma_1^2 = 1$ and $\mu_2 = 10, \sigma_2^2 = 4$. For a data point $x = 7$, calculate the responsibility (posterior probability) of component C_1 for this point, i.e., $\gamma(z_1)$. The PDF of a normal distribution is $\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Provide the answer to three decimal places.

2. **(Jensen's Inequality):** Consider the convex function $f(x) = x^4$. Let X be a random variable that can take values $\{-2, 4\}$ with probabilities $P(X = -2) = 0.25$ and $P(X = 4) = 0.75$. Calculate the value of $f(E[X]) - E[f(X)]$.

3. **(ELBO Calculation):** For a latent variable model, the Evidence Lower Bound (ELBO) is given by $\mathcal{L}(q) = E_{q(z|x)}[\log p(x, z) - \log q(z|x)]$. Given a data point x, an approximate posterior $q(z = 1|x) = 0.8, q(z = 0|x) = 0.2$, and the joint distribution values $\log p(x, z = 1) = -3.5$ and $\log p(x, z = 0) = -5.0$, what is the value of the ELBO?

4. **(VAE KL Divergence):** A VAE encoder outputs parameters for a diagonal Gaussian posterior $q(z|x) = \mathcal{N}(z|\mu, \text{diag}(\sigma^2))$ over a 2D latent space. For a given input x, the encoder outputs $\mu = [0.4, -0.3]$ and the log-variance vector $\log \sigma^2 = [-1.8, -0.4]$. Calculate the KL divergence $D_{KL}(q(z|x) \| p(z))$, where the prior $p(z)$ is the standard normal distribution $\mathcal{N}(0, I)$. The formula is $D_{KL} = \frac{1}{2} \sum_{j=1}^{D} (\sigma_j^2 + \mu_j^2 - 1 - \log \sigma_j^2)$.

5. **(GMM M-Step):** In the M-step of the EM algorithm for a GMM, we have the following responsibilities for two data points, $x_1 = 5$ and $x_2 = 15$, and two components: $\gamma(z_{11}) = 0.9, \gamma(z_{12}) = 0.1$ and $\gamma(z_{21}) = 0.2, \gamma(z_{22}) = 0.8$. Calculate the updated mean μ_2' for the second component.

6. **(VAE Reparameterization):** Using the reparameterization trick $z = \mu + \sigma \cdot \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$, a VAE encoder provides $\mu = 2.5$ and log-variance $\log \sigma^2 = 1.6$. If the random sample from the standard normal is $\epsilon = -1.5$, what is the value of the generated latent variable z?

7. **(Beta-VAE Loss):** A Beta-VAE is trained with $\beta = 8$. At a certain training step, the average reconstruction loss (negative log-likelihood) per data point is 12.4, and the KL divergence term is 2.1. What is the total value of the objective function being minimized?

8. **(VQ-VAE Quantization):** A VQ-VAE uses a codebook with $K = 256$ vectors, each of dimension $D = 32$. The encoder output $z_e(x)$ is a tensor of shape $[16, 16, 32]$. What is the total size (in bits) of the discrete latent representation (the indices sent to the decoder) for a single input?

9. **(VQ-VAE Commitment Loss):** In a VQ-VAE, the commitment loss term is $\beta \| z_e(x) - \text{sg}[e_k] \|_2^2$, where 'sg' is the stop-gradient operator. For one vector from the encoder output, $z_e(x) = [1.2, -0.5, 0.8]$, its closest codebook vector is $e_k = [1.0, -0.9, 1.1]$. If the hyperparameter $\beta = 0.25$, what is the commitment loss for this single vector?

10. **(DDPM Forward Process):** The DDPM forward process is defined by $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$. Let the noise schedule be linear from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$ with $T = 1000$. For timestep $t = 200$, $\bar{\alpha}_{200} \approx 0.979$. Given a normalized data point $x_0 = 0.8$ and a noise sample $\epsilon = -1.2$, what is the value of x_{200}?

11. **(DDPM x_0 Prediction):** During DDPM training or inference at timestep t, the model $\epsilon_\theta(x_t, t)$ predicts the noise that was added to x_0. The original data point can be estimated as $\hat{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t, t))$. If at $t = 500$, $\bar{\alpha}_{500} = 0.7$, $x_{500} = 0.5$, and the model predicts $\epsilon_\theta = -0.2$, what is the estimated value of $\hat{x}_0$?

12. **(DDPM Simplified Loss):** The simplified DDPM training objective is $L_{\text{simple}} = E_{t,x_0,\epsilon}[\|\epsilon - \epsilon_\theta(x_t, t)\|^2]$. For a single 1D sample, let $x_0 = 1.5$. At timestep $t = 400$, $\bar{\alpha}_{400} = 0.75$. A noise sample $\epsilon = 0.5$ is drawn. The model takes the resulting x_{400} and t as input and predicts a noise value of $\epsilon_\theta = 0.3$. What is the squared error loss for this single instance?

13. **(DDPM Sampling Step):** The DDPM sampling step to get x_{t-1} from x_t is given by $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)) + \sigma_t z$, where $z \sim \mathcal{N}(0, I)$. Let $\alpha_t = 0.99$, $\bar{\alpha}_t = 0.80$, and $\sigma_t^2 = \beta_t = 0.01$. If $x_t = 0.7$, the model predicts $\epsilon_\theta = -0.4$, and the sampled noise $z = 1.0$, calculate the value of x_{t-1}.

14. **(U-Net Architecture):** A U-Net architecture is used for images of size 128×128. The network has 4 downsampling blocks, each containing a 2×2 max-pooling operation. What is the spatial resolution (width or height) of the feature map at the bottleneck (the lowest point) of the U-Net?

15. **(VAE Encoder Implementation):** In a PyTorch implementation of a VAE encoder for $32 \times 32 \times 3$ images, the input is first flattened and then passed through a linear layer 'nn.Linear(3072, 400)'. This is followed by a ReLU activation and then two parallel linear layers to produce μ and $\log \sigma^2$, both mapping from 400 features to a latent dimension of 20. What is the total number of trainable weight and bias parameters in the layer that produces the mean vector μ?

16. **(VQ-VAE Implementation):** The VQ-VAE loss function has three components: reconstruction loss, codebook loss $\|\text{sg}[z_e(x)] - e_k\|_2^2$, and commitment loss $\beta\|z_e(x) - \text{sg}[e_k]\|_2^2$. During backpropagation for the codebook loss term, which part of the model is updated?

    (a) Encoder

    (b) Codebook Embeddings

(c) Decoder

(d) Both Encoder and Codebook

17. **(DDPM Timestep Embedding):** In a DDPM, sinusoidal embeddings are used to represent the timestep t. The embedding for dimension i is often calculated as $f(t)_i = \sin(t/10000^{2i/d})$ where d is the embedding dimension. Let $d = 128$ and $t = 0.1$. Calculate the value of the component for $i = 1$ (the second component, assuming 0-indexing).

18. **(Proof of Jensen's Inequality Application):** In the standard derivation of the ELBO, we start from $\log p(x)$ and arrive at an inequality. The step relies on Jensen's inequality for the concave 'log' function: $E[\log(Y)] \leq \log(E[Y])$. What quantity corresponds to Y in this specific proof step?

    (a) $p(x, z)$

    (b) $q(z|x)$

    (c) $\frac{p(x,z)}{q(z|x)}$

    (d) $p(x)$

19. **(ELBO vs Log-Likelihood):** The gap between the log-evidence $\log p(x)$ and the ELBO is equal to a KL divergence term. Which KL divergence is it?

    (a) $D_{KL}(p(z)\|q(z|x))$

    (b) $D_{KL}(q(z|x)\|p(z))$

    (c) $D_{KL}(p(z|x)\|q(z|x))$

    (d) $D_{KL}(q(z|x)\|p(z|x))$

20. **(Beta-VAE Trade-off):** An engineer trains a VAE with $\beta = 1$ and gets a reconstruction loss of 15.0 and a KL divergence of 6.0. To encourage better disentanglement, they retrain with $\beta = 5$. The new KL divergence is 2.0. If the total loss for the new model is 27.5, what is its reconstruction loss?

21. **(DDPM ELBO Equivalence):** The simplified DDPM objective L_{simple} is derived from the full ELBO by setting the variance of the model's reverse transition $p_\theta(x_{t-1}|x_t)$ to a fixed constant σ_t^2. In the original DDPM paper, the theoretically derived choice for this variance corresponds to which of the following?

    (a) β_t

    (b) $\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$

    (c) $\sqrt{\beta_t}$

    (d) α_t

22. **(DDPM Inference Implementation):** In a PyTorch DDPM sampling loop, you have a tensor x_t of shape '[1, 3, 64, 64]' representing the noisy image at step 't'. The model $\epsilon_\theta(x_t, t)$ returns a predicted noise tensor of the same shape. The timestep 't' is a single integer. When passing 't' to the model, it is typically converted to a tensor. What must be the shape of this timestep tensor 't' so that it can be correctly processed by the model for a single image?

(a) ‘[1]‘

(b) ‘[1, 1]‘

(c) ‘[1, 64]‘

(d) ‘[64, 64]‘

23. **(Reparameterization Gradient Flow):** In a VAE, the reparameterization trick $z = \mu(x) + \sigma(x) \cdot \epsilon$ is used. Why is this trick essential when computing the gradient of the VAE loss with respect to the encoder's parameters?

    (a) It makes the KL divergence term computable.

    (b) It makes the sampling process deterministic with respect to the encoder's parameters.

    (c) It ensures the latent variable z has a unit Gaussian distribution.

    (d) It reduces the variance of the stochastic gradients.

24. **(DDPM ELBO for $t = 1$):** The DDPM ELBO contains a reconstruction term $L_0 = -\log p_\theta(x_0|x_1)$. This term corresponds to the final denoising step. The mean of $p_\theta(x_0|x_1)$ is derived from the noise prediction $\epsilon_\theta(x_1, 1)$. Given $\alpha_1 = 0.9999$, $x_1 = 0.6$, and $\epsilon_\theta(x_1, 1) = -0.3$, what is the predicted mean of x_0?

25. **(U-Net with Time Embedding):** In a U-Net for DDPMs, the time embedding is usually added to the feature maps after a convolutional layer. If a feature map has shape ‘[B, C, H, W]‘ and the time embedding has shape ‘[B, C]‘, how is the embedding typically broadcasted and combined?

    (a) It is reshaped to ‘[B, C, 1, 1]‘ and then added.

    (b) It is tiled to ‘[B, C, H, W]‘ and then multiplied.

    (c) It is passed through a separate linear layer for each spatial location.

    (d) It is concatenated along the channel dimension.

Answers and Explanations

1. **Answer: 0.556**
   Explanation: The responsibility $\gamma(z_1)$ is calculated using Bayes' rule.

   - Calculate likelihoods (proportional to PDF value):

   $$p(x|C_1) \propto \frac{1}{\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} = \frac{1}{1} e^{-\frac{(7-5)^2}{2}} = e^{-2} \approx 0.1353$$

   $$p(x|C_2) \propto \frac{1}{\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} = \frac{1}{2} e^{-\frac{(7-10)^2}{8}} = 0.5 \cdot e^{-1.125} \approx 0.1623$$

   - Apply Bayes' rule:

   $$\gamma(z_1) = \frac{\pi_1 p(x|C_1)}{\pi_1 p(x|C_1) + \pi_2 p(x|C_2)} = \frac{0.6 \cdot 0.1353}{0.6 \cdot 0.1353 + 0.4 \cdot 0.1623} = \frac{0.0812}{0.0812 + 0.0649} \approx 0.556$$

2. **Answer: -156.9375**
   Explanation:

   - $E[X] = (-2)(0.25) + (4)(0.75) = -0.5 + 3.0 = 2.5$.
   - $f(E[X]) = (2.5)^4 = 39.0625$.
   - $E[f(X)] = E[X^4] = (-2)^4(0.25) + (4)^4(0.75) = 16(0.25) + 256(0.75) = 4 + 192 = 196$.
   - $f(E[X]) - E[f(X)] = 39.0625 - 196 = -156.9375$.

3. **Answer: -3.300**
   Explanation: $\mathcal{L}(q) = \sum_z q(z|x)(\log p(x,z) - \log q(z|x))$

   $$\begin{aligned}\mathcal{L}(q) &= q(z=1|x)[\log p(x,1) - \log q(z=1|x)] + q(z=0|x)[\log p(x,0) - \log q(z=0|x)] \\ &= 0.8[-3.5 - \log(0.8)] + 0.2[-5.0 - \log(0.2)] \\ &= 0.8[-3.5 - (-0.223)] + 0.2[-5.0 - (-1.609)] \\ &= 0.8[-3.277] + 0.2[-3.391] = -2.6216 - 0.6782 \approx -3.300\end{aligned}$$

4. **Answer: 0.643**
   Explanation:

   - Given: $\mu = [0.4, -0.3]$, $\log \sigma^2 = [-1.8, -0.4]$.
   - We need σ^2: $\sigma_1^2 = e^{-1.8} \approx 0.1653$, $\sigma_2^2 = e^{-0.4} \approx 0.6703$.
   - Summand for $j = 1$: $0.1653 + (0.4)^2 - 1 - (-1.8) = 0.1653 + 0.16 - 1 + 1.8 = 1.1253$.
   - Summand for $j = 2$: $0.6703 + (-0.3)^2 - 1 - (-0.4) = 0.6703 + 0.09 - 1 + 0.4 = 0.1603$.
   - $D_{KL} = \frac{1}{2}(1.1253 + 0.1603) = \frac{1}{2}(1.2856) \approx 0.643$.

5. **Answer: 13.889**
   Explanation: $\mu_k' = \frac{\sum_i \gamma(z_{ik}) x_i}{\sum_i \gamma(z_{ik})}$

- Denominator for $k = 2$: $N_2 = \gamma(z_{12}) + \gamma(z_{22}) = 0.1 + 0.8 = 0.9$.
- Numerator for $k = 2$: $(\gamma(z_{12})x_1)+(\gamma(z_{22})x_2) = (0.1\cdot5)+(0.8\cdot15) = 0.5+12.0 = 12.5$.
- $\mu'_2 = \frac{12.5}{0.9} \approx 13.889$.

6. **Answer: -0.838**
   Explanation: $\sigma = \sqrt{e^{\log\sigma^2}} = e^{0.5\cdot\log\sigma^2}$.

   - $\sigma = e^{0.5\cdot1.6} = e^{0.8} \approx 2.2255$.
   - $z = \mu + \sigma \cdot \epsilon = 2.5 + 2.2255 \cdot (-1.5) = 2.5 - 3.3383 \approx -0.838$.

7. **Answer: 29.2**
   Explanation: $L = L_{\text{recon}} + \beta \cdot D_{KL} = 12.4 + 8 \cdot (2.1) = 12.4 + 16.8 = 29.2$.

8. **Answer: 2048**
   Explanation:

   - Number of latent vectors = $16 \times 16 = 256$.
   - Bits to represent one vector index = $\log_2(K) = \log_2(256) = 8$ bits.
   - Total bits = 256 vectors $\times$ 8 bits/vector = 2048 bits.

9. **Answer: 0.0725**
   Explanation: Loss $= \beta\|z_e(x) - e_k\|_2^2$.

   - Squared distance $= (1.2 - 1.0)^2 + (-0.5 - (-0.9))^2 + (0.8 - 1.1)^2 = (0.2)^2 + (0.4)^2 + (-0.3)^2 = 0.04 + 0.16 + 0.09 = 0.29$.
   - Loss $= 0.25 \cdot 0.29 = 0.0725$.

10. **Answer: 0.618**
    Explanation: $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$.

    - $\sqrt{0.979} \approx 0.9894$, $\sqrt{1 - 0.979} = \sqrt{0.021} \approx 0.1449$.
    - $x_{200} = (0.9894 \cdot 0.8) + (0.1449 \cdot -1.2) = 0.7915 - 0.1739 \approx 0.618$.

11. **Answer: 0.729**
    Explanation: $\hat{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta)$.

    - $\sqrt{0.7} \approx 0.8367$, $\sqrt{1 - 0.7} = \sqrt{0.3} \approx 0.5477$.
    - $\hat{x}_0 = \frac{1}{0.8367}(0.5 - 0.5477 \cdot (-0.2)) = \frac{1}{0.8367}(0.5 + 0.1095) = \frac{0.6095}{0.8367} \approx 0.729$.

12. **Answer: 0.04**
    Explanation: Loss is simply $\|\epsilon - \epsilon_\theta\|^2 = (0.5 - 0.3)^2 = (0.2)^2 = 0.04$. The other information is used by the model to produce the prediction, but is not part of the final error calculation.

13. **Answer: 0.813**
    Explanation: $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta) + \sigma_t z$.

    - $\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} = \frac{0.01}{\sqrt{0.2}} \approx 0.02236$.

- $x_t - \ldots \epsilon_\theta = 0.7 - (0.02236 \cdot -0.4) \approx 0.7 + 0.00894 = 0.70894$.
- $\frac{1}{\sqrt{0.99}}(0.70894) \approx 1.005 \cdot 0.70894 \approx 0.7125$.
- $x_{t-1} \approx 0.7125 + \sqrt{0.01} \cdot 1.0 = 0.7125 + 0.1 = 0.8125 \approx 0.813$.

14. **Answer: 8**
Explanation: Initial dimension is 128. After 4 pooling operations, it becomes $128 \to 64 \to 32 \to 16 \to 8$.

15. **Answer: 8020**
Explanation: The layer is 'Linear(in=400, out=20)'.
    - Weights $= 400 \times 20 = 8000$. Biases $= 20$. Total $= 8000 + 20 = 8020$.

16. **Answer: 2**
Explanation: The stop-gradient 'sg' on $z_e(x)$ prevents gradients from flowing to the encoder. The loss is a function of the codebook vectors e_k, so they are the only parameters updated by this term.

17. **Answer: 0.0017**
Explanation: $f(t)_i = \sin(t/10000^{2i/d})$. For $i = 1, t = 0.1, d = 128$:
$$\sin(0.1/10000^{2 \cdot 1/128}) = \sin(0.1/10000^{1/64}) \approx \sin(0.1/1.154) \approx \sin(0.0866) \approx 0.0865$$
There might be a misunderstanding in the question's original intent. If the formula was $\sin(t/(d/2)^i)$, it would be $\sin(0.1/64) \approx 0.00156$. Let's assume the common Transformer formula, so the result is 0.0865.

18. **Answer: 3**
Explanation: $\log p(x) = \log \int p(x,z)dz = \log \int q(z|x)\frac{p(x,z)}{q(z|x)}dz = \log E_{q(z|x)}[\frac{p(x,z)}{q(z|x)}]$. Jensen's inequality is applied to the expectation, so $Y = \frac{p(x,z)}{q(z|x)}$.

19. **Answer: 4**
Explanation: The exact decomposition is $\log p(x) = \text{ELBO} + D_{KL}(q(z|x)\|p(z|x))$. The gap is the KL divergence from the true posterior to the approximate posterior.

20. **Answer: 17.5**
Explanation: $L = L_{\text{recon}} + \beta \cdot D_{KL}$. We have $27.5 = L_{\text{recon}} + 5 \cdot 2.0$. Thus, $L_{\text{recon}} = 27.5 - 10 = 17.5$.

21. **Answer: 2**
Explanation: The variance of the true posterior $q(x_{t-1}|x_t, x_0)$ is $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. This is the theoretically motivated choice for the variance σ_t^2 of the learned reverse process $p_\theta(x_{t-1}|x_t)$.

22. **Answer: 1**
Explanation: The model expects a batch dimension. Even for a single sample, the timestep 't' must be passed as a tensor with a batch dimension, i.e., of shape '[1]'.

23. **Answer: 2**
Explanation: A sampling operation is a stochastic node that breaks the gradient chain. By rewriting z as a deterministic function of parameters (μ, σ) and an independent random source (ϵ), a continuous path is created for gradients to flow from the loss back to the encoder parameters.

24. **Answer: 0.603**
Explanation: The mean of $p_\theta(x_0|x_1)$ is $\mu_\theta(x_1, 1) = \frac{1}{\sqrt{\alpha_1}}(x_1 - \frac{1-\alpha_1}{\sqrt{1-\bar{\alpha}_1}}\epsilon_\theta)$. Since $\bar{\alpha}_1 = \alpha_1$, this becomes $\mu_\theta = \frac{1}{\sqrt{\alpha_1}}(x_1 - \sqrt{1-\alpha_1}\epsilon_\theta)$.

$$\mu_\theta = \frac{1}{\sqrt{0.9999}}(0.6 - \sqrt{0.0001} \cdot (-0.3)) = \frac{1}{0.99995}(0.6 - 0.01 \cdot (-0.3)) = \frac{0.603}{0.99995} \approx 0.60303$$

25. **Answer: 1**
Explanation: This is the standard method. The time embedding '[B, C]' is reshaped to '[B, C, 1, 1]' and added to the feature map '[B, C, H, W]'. Broadcasting rules automatically expand the last two dimensions to match 'H' and 'W'.