Sample Midterm Exam – Fall 2025 (with Answers)

Q1 – Forward Pass of a Neuron
Given: input $x = 2.0$, weight $w = 0.5$, bias $b = -0.3$. Activation is the sigmoid

$$f(z) = \frac{1}{1+e^{-z}}, \qquad z = wx + b.$$

(1) Compute z **and** $\hat{y} = f(z)$.

$$\begin{aligned} z &= wx + b = (0.5)(2.0) + (-0.3) = 1.0 - 0.3 = 0.7, \\ \hat{y} &= f(0.7) = \frac{1}{1+e^{-0.7}}. \end{aligned}$$

Now $e^{-0.7} \approx 0.4965853$, so

$$\hat{y} \approx \frac{1}{1+0.4965853} \approx \frac{1}{1.4965853} \approx 0.6681878 \;\Rightarrow\; \boxed{\hat{y} \approx 0.67.}$$

(2) With target $y = 0.8$, **compute loss** $J = \frac{1}{2}(\hat{y} - y)^2$.

$$\begin{gathered} \hat{y} - y \approx 0.6681878 - 0.8 = -0.1318122, \\ (\hat{y} - y)^2 \approx (-0.1318122)^2 \approx 0.017372, \\ J = \tfrac{1}{2} \times 0.017372 \approx 0.008686 \;\Rightarrow\; \boxed{J \approx 0.01.} \end{gathered}$$

Q2 – Convolution and Pooling Dimensions
An image of size 6×6 is convolved with a 3×3 filter, stride = 2, and no padding.

Given: input image 6×6; convolution with 3×3 filter, stride $S = 2$, padding $P = 0$.

(1) Convolution output size. For each spatial dimension,

$$\text{out} = \left\lfloor \frac{W-F}{S} \right\rfloor + 1 = \left\lfloor \frac{6-3}{2} \right\rfloor + 1 = \left\lfloor \frac{3}{2} \right\rfloor + 1 = 1 + 1 = 2.$$

So the feature map is $\boxed{2 \times 2}$ (with as many channels as filters; here size only was asked).

(2) 2×2 **max pooling, stride** $S = 2$**.** Starting from 2×2,

$$\text{out} = \left\lfloor \frac{2-2}{2} \right\rfloor + 1 = \lfloor 0 \rfloor + 1 = 1.$$

Final pooled map: $\boxed{1 \times 1}$.

(3) Compute the total number of learnable parameters (including biases) if 3 filters were used

Each 3×3 filter has 9 weights + 1 bias = 10 parameters. If there is only one filter, total parameters = 10. If 3 filters were used, parameters = $3 \times (9 + 1) = 30$.

Q3 – PCA and Feature Scaling
A dataset contains three features:

- Feature 1: Building height (in meters), range $[0, 100]$
- Feature 2: Energy use (in kWh), range $[0, 10{,}000]$
- Feature 3: Occupancy rate (in percent), range $[0, 100]$

Solution:
PCA is variance-dominant. The feature with the largest numeric scale typically has the largest variance (without standardization). Here, *Energy use* ranges up to 10,000, far larger than the other features (≤ 100), so it will dominate the covariance matrix and therefore the first principal component.

$$\boxed{\text{Energy use (kWh) most strongly influences PC1 without scaling.}}$$

Q4 – SVM Decision Boundary
For a linear SVM, the separating hyperplane is:

$$3x_1 - 4x_2 + 2 = 0$$

For the point $(x_1, x_2) = (2, 1)$, compute $f(x)$ and determine which side it lies on.

Given hyperplane: $3x_1 - 4x_2 + 2 = 0$. Decision function: $f(\mathbf{x}) = 3x_1 - 4x_2 + 2$.

For $(x_1, x_2) = (2, 1)$:
$$f(2, 1) = 3(2) - 4(1) + 2 = 6 - 4 + 2 = 4.$$
Since $f(2, 1) = 4 > 0$, the point lies on the *positive* side of the boundary.

$$\boxed{f(2, 1) = 4 \text{ (positive side).}}$$

Q5 – Decision Tree Entropy and Gain
Solution:
Given class counts: Play=Yes: 6, Play=No: 2. Total $N = 8$.

(1) Entropy before splitting (base 2).

$$p_{\text{yes}} = \frac{6}{8} = 0.75, \qquad p_{\text{no}} = \frac{2}{8} = 0.25.$$

$$H_{\text{parent}} = -\Big(0.75\log_2 0.75 + 0.25\log_2 0.25\Big) \approx -\Big(0.75(-0.4150) + 0.25(-2)\Big) = -(-0.3113 - 0.5) = 0.8113.$$

$$\boxed{H_{\text{parent}} \approx 0.81.}$$

Now the feature "Wind" splits into two branches:

Branch	Yes	No	Total
Weak	4	1	5
Strong	2	1	3

(2) Entropy of each branch.

Weak (4 Yes, 1 No):

$$p_Y = \tfrac{4}{5} = 0.8, \quad p_N = \tfrac{1}{5} = 0.2,$$

$$H_{\text{Weak}} = -(0.8\log_2 0.8 + 0.2\log_2 0.2) \approx -(0.8(-0.3219) + 0.2(-2.3219)) \approx -(-0.2575 - 0.4644) \approx 0.7219.$$

$$\boxed{H_{\text{Weak}} \approx 0.72.}$$

Strong (2 Yes, 1 No):

$$p_Y = \tfrac{2}{3} \approx 0.6667, \quad p_N = \tfrac{1}{3} \approx 0.3333,$$

$$H_{\text{Strong}} = -(\tfrac{2}{3}\log_2 \tfrac{2}{3} + \tfrac{1}{3}\log_2 \tfrac{1}{3}) \approx -\big(0.6667(-0.5850) + 0.3333(-1.5850)\big) \approx -(-0.3900 - 0.5283) \approx 0.9183.$$

$$\boxed{H_{\text{Strong}} \approx 0.92.}$$

(3) Which branch is purer, and why?
Lower entropy $\Rightarrow$ purer.

$$H_{\text{Weak}} \approx 0.72 \; < \; H_{\text{Strong}} \approx 0.92.$$

$$\boxed{\text{The Weak branch is purer (lower entropy).}}$$

Q6 – True or False (Concept Check)

Statement	Answer
1. Increasing the number of filters in a CNN layer decreases the feature map depth.	**False**
2. The sigmoid activation always outputs values between -1 and 1.	**False**
3. In PCA, the first principal component captures the direction of maximum variance in the data.	**True**
4. Gradient descent can converge faster with a larger learning rate, but may overshoot the minimum.	**True**

Q7 – Transformers (Very Short Answer)
Answer each in one short sentence.

1. **What is the main purpose of the attention mechanism in a Transformer?**
Solution: Attention lets each token focus on the most relevant words in a sequence, learning which parts of input matter most.

2. **What is the role of the Feed-Forward Network (FFN) layer after attention in each Transformer block?**
Solution: The FFN refines each token's internal representation independently, adding non-linearity and improving feature mixing after attention.

Q8 – CNN Filter Computation with Stride

We perform a convolution of the 4×4 input X with a 2×2 kernel K, stride = 2, and bias $b = 1$.

$$X = \begin{bmatrix} 1 & 2 & 3 & 0 \\ 0 & 1 & 2 & 3 \\ 3 & 1 & 0 & 2 \\ 2 & 0 & 1 & 1 \end{bmatrix}, \quad K = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}, \quad b = 1$$

Since stride = 2 and padding = 0, the kernel moves two pixels at a time. The output size is computed as:

$$\text{Output size} = \frac{(N - F)}{S} + 1 = \frac{(4 - 2)}{2} + 1 = 2$$

So the output feature map will be 2×2.

Each output = $(X_{\text{patch}} \cdot K) + b$

—

Patch 1: top-left (1,1)

$$\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} * \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} = (1 \times 1) + (2 \times 0) + (0 \times -1) + (1 \times 1) = 2$$

Add bias: $2 + 1 = 3$

Patch 2: top-right (1,3)

$$\begin{bmatrix} 3 & 0 \\ 2 & 3 \end{bmatrix} * \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} = (3 \times 1) + (0 \times 0) + (2 \times -1) + (3 \times 1) = 4$$

Add bias: $4 + 1 = 5$

Patch 3: bottom-left (3,1)

$$\begin{bmatrix} 3 & 1 \\ 2 & 0 \end{bmatrix} * \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} = (3 \times 1) + (1 \times 0) + (2 \times -1) + (0 \times 1) = 1$$

Add bias: $1 + 1 = 2$

Patch 4: bottom-right (3,3)

$$\begin{bmatrix} 0 & 2 \\ 1 & 1 \end{bmatrix} * \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} = (0 \times 1) + (2 \times 0) + (1 \times -1) + (1 \times 1) = 0$$

Add bias: $0 + 1 = 1$

—

Final Output Feature Map:

$$Y = \begin{bmatrix} 3 & 5 \\ 2 & 1 \end{bmatrix}$$