Quiz II

Name: **Roll Number**.....

- Answer precisely in the space given. No overwriting. Make assumptions, if really ambiguous.
- Do rough work in the additional sheet.

Fill in the blank with precise answers. [8 × 5 = 40]

1. Consider a simple neural network with two inputs, one hidden layer with three neurons and one output. No bias. If the input is $x_1 = 3$ and $x_2 = 2$, output is ——— (fact: All weights are 1.0. Activation functions in the hidden layer is ReLU and activation function in the output is linear.)

Solution:-

The neural network processes the inputs as follows:

1. Hidden Layer Calculation:

- Each hidden neuron receives inputs $x_1 = 3$ and $x_2 = 2$ with weights of 1.0.
- Pre-activation: $(3 \times 1) + (2 \times 1) = 5$.
- ReLU activation: $\text{ReLU}(5) = 5$.
- All three hidden neurons output 5.

2. Output Layer Calculation:

- The output neuron sums the three hidden layer outputs (each multiplied by weight 1.0):

$$5 \times 1 + 5 \times 1 + 5 \times 1 = 15.$$

- Linear activation passes the sum directly.

Output: 15

2. Continuing the about question, assume the target (t) of the above sample was 25, and the loss was squared error, $\frac{\partial \mathcal{L}}{\partial w_{11}}$ is ——— (w_{11} is the weight from first input (i.e., x_1) to first neuron in the hidden layer.)

Solution:-

To compute $\frac{\partial \mathcal{L}}{\partial w_{11}}$ using backpropagation:

1. Loss Derivative:

$$\frac{\partial \mathcal{L}}{\partial \hat{y}} = 2(\hat{y} - t) = 2(15 - 25) = -20.$$

2. Output Layer Gradient:

$$\frac{\partial \hat{y}}{\partial h_1} = 1 \quad (\text{since } \hat{y} = h_1 + h_2 + h_3).$$

3. Hidden Layer Activation Gradient:

$$\frac{\partial h_1}{\partial z_1} = 1 \quad \text{(ReLU derivative at } z_1 = 5\text{)}.$$

4. Pre-activation Gradient:

$$\frac{\partial z_1}{\partial w_{11}} = x_1 = 3.$$

Chain Rule:

$$\frac{\partial \mathcal{L}}{\partial w_{11}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_{11}} = (-20) \cdot 1 \cdot 1 \cdot 3 = -60$$

Answer: -60

3. Consider a single layer perceptron with two input and one output. The weights from first and second inputs are w_1 and w_2 respectively. Bias is w_0. Also assume a $\pm, 1$ logic. The activation at the output is: $\phi(x) = +1$ if $x \geq 0$ and -1 else. If $w_0 = 1, w_1 = -1, w_2 = -1$, then this neuron model is equivalent to: ———————- (fill from the gates like: AND, OR, ExOR, NAND, NOR etc.)

Solution:-

Consider the perceptron with two inputs and one output. The net input to the neuron is given by:

$$x = w_0 \cdot 1 + w_1 x_1 + w_2 x_2$$

Given that $w_0 = 1$, $w_1 = -1$, and $w_2 = -1$, we compute x for all possible input values x_1, x_2 (where $x_1, x_2 \in \{+1, -1\}$):

$$x = 1 - x_1 - x_2$$

The activation function is:

$$\phi(x) = \begin{cases} +1, & \text{if } x \geq 0 \\ -1, & \text{otherwise} \end{cases}$$

We now evaluate the truth table:

x_1	x_2	$x = 1 - x_1 - x_2$	$\phi(x)$
$+1$	$+1$	$1 - 1 - 1 = -1$	-1
$+1$	-1	$1 - 1 - (-1) = 1$	$+1$
-1	$+1$	$1 - (-1) - 1 = 1$	$+1$
-1	-1	$1 - (-1) - (-1) = 3$	$+1$

This matches the truth table of the **NAND** gate (in ± 1 logic):

$$\text{NAND}(x_1, x_2) = \begin{cases} +1, & \text{if } (x_1, x_2) \neq (+1, +1) \\ -1, & \text{if } (x_1, x_2) = (+1, +1) \end{cases}$$

Thus, this perceptron implements the **NAND** gate.

4. Consider a binary classification problem (A vs B) with perceptron and bias. Consider two samples from A: $[1,2]^T$, $[4,2]^T$ and one sample from B: $[3,0]^T$. We initialize $\mathbf{w^0} = [1,1,0]^T$. When learning rate $\eta = 0.1$, $\mathbf{w^1}$ (i.e., after one iteration of perceptron algorithm) is ————

Solution:-

We are given:

- Class A samples (label +1): $\mathbf{x}_1 = \begin{bmatrix}1\\2\end{bmatrix}$, $\mathbf{x}_2 = \begin{bmatrix}4\\2\end{bmatrix}$
- Class B sample (label −1): $\mathbf{x}_3 = \begin{bmatrix}3\\0\end{bmatrix}$
- Initial weights: $\mathbf{w}^{(0)} = \begin{bmatrix}1\\1\\0\end{bmatrix}$
- Learning rate: $\eta = 0.1$

We augment each input vector with a bias term of 1:

$$\tilde{\mathbf{x}}_1 = \begin{bmatrix}1\\2\\1\end{bmatrix}, \quad \tilde{\mathbf{x}}_2 = \begin{bmatrix}4\\2\\1\end{bmatrix}, \quad \tilde{\mathbf{x}}_3 = \begin{bmatrix}3\\0\\1\end{bmatrix}$$

Now we apply one iteration of the Perceptron algorithm:

- For $\tilde{\mathbf{x}}_1$, $y = +1$:
$$\mathbf{w}^{(0)} \cdot \tilde{\mathbf{x}}_1 = [1,1,0] \cdot [1,2,1] = 3 > 0 \Rightarrow \text{Correct}$$
- For $\tilde{\mathbf{x}}_2$, $y = +1$:
$$\mathbf{w}^{(0)} \cdot \tilde{\mathbf{x}}_2 = [1,1,0] \cdot [4,2,1] = 6 > 0 \Rightarrow \text{Correct}$$
- For $\tilde{\mathbf{x}}_3$, $y = -1$:
$$\mathbf{w}^{(0)} \cdot \tilde{\mathbf{x}}_3 = [1,1,0] \cdot [3,0,1] = 3 > 0 \Rightarrow \text{Incorrect}$$

  Update rule:
$$\mathbf{w}^{(1)} = \mathbf{w}^{(0)} + \eta \cdot y \cdot \tilde{\mathbf{x}}_3 = \begin{bmatrix}1\\1\\0\end{bmatrix} + 0.1 \cdot (-1) \cdot \begin{bmatrix}3\\0\\1\end{bmatrix} = \begin{bmatrix}0.7\\1\\-0.1\end{bmatrix}$$

Final answer:

$$\boxed{\mathbf{w}^{(1)} = \begin{bmatrix}0.7\\1\\-0.1\end{bmatrix}}$$

5. Consider a 1D binary classification problem (2.5,+), (4.2,+), (5.6,-), (7.8,-). We train a hard margin SVM. The support vector(s) is/are ————————- (if there are more than one SV, write all.)

Solution:-

We are given the one-dimensional dataset:

$$\{(2.5, +), (4.2, +), (5.6, -), (7.8, -)\}.$$

For a hard-margin SVM with decision function

$$f(x) = wx + b,$$

we must satisfy the constraint

$$y_i(wx_i + b) \geq 1 \quad \text{for all } i.$$

Analysis:

- The positive class consists of points at $x = 2.5$ and $x = 4.2$
- The negative class consists of points at $x = 5.6$ and $x = 7.8$
- The decision boundary should lie between the rightmost positive point and leftmost negative point

Identifying Support Vectors:

- The rightmost positive point: 4.2
- The leftmost negative point: 5.6
- These are the closest points between the two classes and will satisfy:

$$w \cdot 4.2 + b = +1 \quad \text{and} \quad w \cdot 5.6 + b = -1.$$

- Solving these gives the optimal hyperplane parameters

Verification:

- Point at 2.5: $w \cdot 2.5 + b \geq 1$ (satisfied with margin)
- Point at 7.8: $w \cdot 7.8 + b \leq -1$ (satisfied with margin)
- Only 4.2 and 5.6 lie exactly on the margin boundaries

Thus, the support vectors that define the optimal margin are:

$$\boxed{(4.2, +) \quad \text{and} \quad (5.6, -).}$$

6. Consider the following dataset $\mathcal{D}$ of 5 samples in 2D, $\mathcal{D} = \{[1,1]^T, [2,2]^T, [3,3]^T, [4,4]^T, [5,5]^T\}$. The covariance of $\mathcal{D}$ is ————

Solution:-

Given the 2D dataset:

$$D = \left\{\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}2\\2\end{bmatrix}, \begin{bmatrix}3\\3\end{bmatrix}, \begin{bmatrix}4\\4\end{bmatrix}, \begin{bmatrix}5\\5\end{bmatrix}\right\}$$

Step 1: Compute the Mean Vector

$$\mu = \frac{1}{5}\sum_{i=1}^{5} \mathbf{x}_i = \frac{1}{5}\left(\begin{bmatrix}1\\1\end{bmatrix} + \begin{bmatrix}2\\2\end{bmatrix} + \begin{bmatrix}3\\3\end{bmatrix} + \begin{bmatrix}4\\4\end{bmatrix} + \begin{bmatrix}5\\5\end{bmatrix}\right) = \begin{bmatrix}3\\3\end{bmatrix}$$

Step 2: Center the Data Subtract the mean from each data point:

$$\mathbf{x}_i' = \mathbf{x}_i - \mu$$

The centered data becomes:

$$\left\{\begin{bmatrix}-2\\-2\end{bmatrix}, \begin{bmatrix}-1\\-1\end{bmatrix}, \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}2\\2\end{bmatrix}\right\}$$

Step 3: Compute the Covariance Matrix For a 2D dataset, the covariance matrix is:

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i' \mathbf{x}_i'^T$$

Calculating each term:

$$\frac{1}{5}\left(\begin{bmatrix}-2\\-2\end{bmatrix}\begin{bmatrix}-2 & -2\end{bmatrix} + \begin{bmatrix}-1\\-1\end{bmatrix}\begin{bmatrix}-1 & -1\end{bmatrix} + \begin{bmatrix}0\\0\end{bmatrix}\begin{bmatrix}0 & 0\end{bmatrix} + \begin{bmatrix}1\\1\end{bmatrix}\begin{bmatrix}1 & 1\end{bmatrix} + \begin{bmatrix}2\\2\end{bmatrix}\begin{bmatrix}2 & 2\end{bmatrix}\right)$$

$$= \frac{1}{5}\left(\begin{bmatrix}4 & 4\\4 & 4\end{bmatrix} + \begin{bmatrix}1 & 1\\1 & 1\end{bmatrix} + \begin{bmatrix}0 & 0\\0 & 0\end{bmatrix} + \begin{bmatrix}1 & 1\\1 & 1\end{bmatrix} + \begin{bmatrix}2 & 2\\2 & 2\end{bmatrix}\right)$$

$$= \frac{1}{5}\begin{bmatrix}10 & 10\\10 & 10\end{bmatrix} = \begin{bmatrix}2 & 2\\2 & 2\end{bmatrix}$$

Final Answer: The covariance matrix of dataset D is:

$$\boxed{\begin{bmatrix}2 & 2\\2 & 2\end{bmatrix}}$$

Interpretation: The identical diagonal elements (2) indicate equal variance in both dimensions, and the off-diagonal elements (2) show perfect positive correlation between the two dimensions, as all points lie exactly on the line y=x.

7. A dataset $\mathcal{D}$ consists of six feature measurements. $\mathcal{D} = \{2.2, 3.5, 4.8, 1.7, 2.8, 3.0\}$. $\mathcal{D}$ needs to be normalized such that mean is zero and variance is unity. The normalized $\mathcal{D}$ is _____

Solution:-

Given the dataset:

$$D = \{2.2,\ 3.5,\ 4.8,\ 1.7,\ 2.8,\ 3.0\}$$

Step 1: Compute the Mean (μ)

$$\mu = \frac{1}{6}\sum_{i=1}^{6} x_i = \frac{2.2 + 3.5 + 4.8 + 1.7 + 2.8 + 3.0}{6} = \frac{18.0}{6} = 3.0$$

Step 2: Compute the Variance (σ^2) First, calculate the squared differences from the mean:

$$(2.2 - 3.0)^2 = 0.64$$
$$(3.5 - 3.0)^2 = 0.25$$
$$(4.8 - 3.0)^2 = 3.24$$
$$(1.7 - 3.0)^2 = 1.69$$
$$(2.8 - 3.0)^2 = 0.04$$
$$(3.0 - 3.0)^2 = 0.00$$

$$\sigma^2 = \frac{0.64 + 0.25 + 3.24 + 1.69 + 0.04 + 0.00}{6} = \frac{5.86}{6} \approx 0.9767$$
$$\sigma = \sqrt{0.9767} \approx 0.9883$$

Step 3: Normalize Each Data Point Using the formula:

$$z_i = \frac{x_i - \mu}{\sigma}$$
$$z_1 = \frac{2.2 - 3.0}{0.9883} \approx -0.809$$
$$z_2 = \frac{3.5 - 3.0}{0.9883} \approx 0.506$$
$$z_3 = \frac{4.8 - 3.0}{0.9883} \approx 1.822$$
$$z_4 = \frac{1.7 - 3.0}{0.9883} \approx -1.316$$
$$z_5 = \frac{2.8 - 3.0}{0.9883} \approx -0.202$$
$$z_6 = \frac{3.0 - 3.0}{0.9883} = 0.000$$

Final Answer: The normalized dataset is:

$$\boxed{\{-0.809,\ 0.506,\ 1.822,\ -1.316,\ -0.202,\ 0.000\}}$$

Verification:

- Mean of normalized data: $(-0.809 + 0.506 + 1.822 - 1.316 - 0.202 + 0.000)/6 \approx 0$
- Variance of normalized data: ≈ 1 (by construction)

8. We know for a valid kernel $\kappa(\mathbf{p}, \mathbf{q})$, there exists a $\phi(\mathbf{p})^T \phi(\mathbf{q})$ form where ϕ does an appropriate feature transformation on $\mathbf{p}$ and $\mathbf{q}$. When $\mathbf{p} = [1, 1]^T$; $\mathbf{q} = [2, 2]^T$ and $\kappa(\mathbf{p}, \mathbf{q}) = (\mathbf{p}^T \mathbf{q} + 1)^2$, find the corresponding ϕ and write as $\phi(\mathbf{p})^T \phi(\mathbf{q})$ ———————— (such as $[- - - - - - - - -]^T [- - - - - - - - - -]$)

Solution:-

Given the kernel function:

$$\kappa(\mathbf{p}, \mathbf{q}) = (\mathbf{p}^T \mathbf{q} + 1)^2$$

with input vectors:

$$\mathbf{p} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{q} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

Step 1: Expand the Kernel Function

$$\kappa(\mathbf{p}, \mathbf{q}) = (\mathbf{p}^T \mathbf{q} + 1)^2 = (1 \cdot 2 + 1 \cdot 2 + 1)^2 = (2 + 2 + 1)^2 = 25$$

Step 2: Identify the Feature Mapping The polynomial kernel $(x^T y + 1)^2$ corresponds to the feature mapping:

$$\phi(\mathbf{x}) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2} x_1 x_2 \\ \sqrt{2} x_1 \\ \sqrt{2} x_2 \\ 1 \end{bmatrix}$$

Step 3: Compute (p) and (q) For $\mathbf{p} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$:

$$\phi(\mathbf{p}) = \begin{bmatrix} 1^2 \\ 1^2 \\ \sqrt{2} \cdot 1 \cdot 1 \\ \sqrt{2} \cdot 1 \\ \sqrt{2} \cdot 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \sqrt{2} \\ \sqrt{2} \\ \sqrt{2} \\ 1 \end{bmatrix}$$

For $\mathbf{q} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$:

$$\phi(\mathbf{q}) = \begin{bmatrix} 2^2 \\ 2^2 \\ \sqrt{2} \cdot 2 \cdot 2 \\ \sqrt{2} \cdot 2 \\ \sqrt{2} \cdot 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \\ 4\sqrt{2} \\ 2\sqrt{2} \\ 2\sqrt{2} \\ 1 \end{bmatrix}$$

Step 4: Verify $(\mathbf{p})^T(q) = (p, q)\phi(\mathbf{p})^T\phi(\mathbf{q}) = 1 \cdot 4 + 1 \cdot 4 + \sqrt{2} \cdot 4\sqrt{2} + \sqrt{2} \cdot 2\sqrt{2} + \sqrt{2} \cdot 2\sqrt{2} + 1 \cdot 1$

$$= 4 + 4 + 8 + 4 + 4 + 1 = 25 \quad \text{(matches kernel calculation)}$$

Final Answer: The feature transformation is:

$$\phi(\mathbf{p})^T\phi(\mathbf{q}) = \boxed{\begin{bmatrix} 1 & 1 & \sqrt{2} & \sqrt{2} & \sqrt{2} & 1 \end{bmatrix} \begin{bmatrix} 4 \\ 4 \\ 4\sqrt{2} \\ 2\sqrt{2} \\ 2\sqrt{2} \\ 1 \end{bmatrix}}$$

Quiz II

Name: **Roll Number**.....

- Answer precisely in the space given. No overwriting. Make assumptions, if really ambiguous.
- Do rough work in the additional sheet.

Fill in the blank with precise answers; [8 × 5 = 40]

1. Consider a simple neural network with two inputs, one hidden layer with three neurons and one output. No bias. If the input is $x_1 = 3$ and $x_2 = 3$, output is ——— (fact: All weights are 1.0. Activation functions in the hidden layer is ReLU and activation function in the output is linear.)

   Answer:

   Each hidden neuron computes the following.

   $$\begin{aligned} z_i &= w_{i1}x_1 + w_{i2}x_2 && i \in \{1,2,3\} \\ &= 3 + 3 \\ &= 6 \\ a_i &= \max(0, z_i) && i \in \{1,2,3\} \\ &= 6 \end{aligned}$$

   The final output is a linear combination of the hidden neuron outputs with weights 1.

   $$\begin{aligned} \hat{y} &= \sum_{i=1}^{3} a_i \\ &= 6 + 6 + 6 \\ &= \boxed{18} \end{aligned}$$

2. Continuing the about question, assume the target (t) of the above sample was 25, and the loss was squared error, $\frac{\partial \mathcal{L}}{\partial w_{11}}$ is ——— (w_{11} is the weight from first input (i.e., x_1) to first neuron in the hidden layer.)

   Answer:

   We know $y = 25$, $\hat{y} = 18$. We can write the squared error as follows.

   $$\mathcal{L} = (y - \hat{y})^2,$$

   The gradient with respect to the final output can be computed as follows.

   $$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \hat{y}} &= -2(y - \hat{y}) \\ &= -2(25 - 18) \\ &= -14 \end{aligned}$$

   It is easy to see that the following holds, when the output layer weights are tied to one.

   $$\begin{aligned} \frac{\partial \mathcal{L}}{\partial} &= \frac{\partial \mathcal{L}}{\partial a_i} && i \in \{1,2,3\} \\ &= \frac{\partial \mathcal{L}}{\partial z_i} && i \in \{1,2,3\} \end{aligned}$$

   Finally, the gradient with respect to the linear layer weights can be computed as follows.

   $$\begin{aligned} \frac{\partial \mathcal{L}}{\partial w_{11}} &= \frac{\partial \mathcal{L}}{\partial z_i} \times \frac{\partial z_i}{\partial w_{11}} \\ &= -14 \times x_1 \\ &= -14 \times 3 \\ &= \boxed{-42} \end{aligned}$$

3. Consider a single layer perceptron with two input and one output. The weights from first and second inputs are w_1 and w_2 respectively. Bias is w_0. Also assume a $\pm, 1$ logic. The activation at the output is: $\phi(x) = +1$ if $x \geq 0$ and -1 else. If $w_0 = 1, w_1 = -1, w_2 = -1$, then this neuron model is equivalent to: ————————- (fill from the gates like: AND, OR, ExOR, NAND, NOR etc.)

Answer:

The output is computed as follows.

$$z = 1 - x_1 - x_2$$
$$\phi(x) = \begin{cases} +1, & \text{if } z \geq 0 \\ -1, & \text{if } z < 0 \end{cases}$$

Computing the truth table for all input combinations $(x_1, x_2) \in \{+1, -1\}^2$ shows the model resembles $\boxed{\text{NAND}}$ gate.

x_1	x_2	z	$\phi(x)$
1	1	-1	-1
1	-1	1	1
-1	1	1	1
-1	-1	3	1

4. Consider a binary classification problem (A vs B) with perceptron and bias. Consider two samples from A: $[1, 2]^T, [4, 2]^T$ and one sample from B: $[3, 0]^T$. We initialize $\mathbf{w^0} = [1, 1, 0]^T$. When learning rate $\eta = 0.1$, $\mathbf{w^1}$ (i.e., after one iteration of perceptron algorithm) is ————————

Answer:

The output of the perceptron can be computed as follows. Note that the input has to be preprended with a 1 for accommodating the bias weight.

$$\hat{y}_i = w^T x_i$$
$$\hat{y}_1 = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} = 2$$
$$\hat{y}_2 = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} = 5$$
$$\hat{y}_3 = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 0 \end{bmatrix} = 4$$

The initial weight can be updated for the misclassified sample 3 (of class B) as follows.

$$\mathbf{w}^1 = \mathbf{w}^0 - \eta \mathbf{x}_3$$
$$= \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} - 0.1 \begin{bmatrix} 1 \\ 3 \\ 0 \end{bmatrix}$$
$$= \boxed{\begin{bmatrix} 0.9 \\ 0.7 \\ 0 \end{bmatrix}}$$

Some students have assumed that the input is appended by 1, rather than prepended. The weight update in that case would be as follows.

$$\begin{aligned}\mathbf{w}^1 &= \mathbf{w}^0 - \eta \mathbf{x}_3 \\ &= \begin{bmatrix}1\\1\\0\end{bmatrix} - 0.1\begin{bmatrix}3\\0\\1\end{bmatrix} \\ &= \boxed{\begin{bmatrix}0.7\\1\\-0.1\end{bmatrix}}\end{aligned}$$

5. Consider a 1D binary classification problem (2.5,+), (4.2,-), (5.6,-), (7.8,-). We train a hard margin SVM. The support vector(s) is/are ———————- (if there are more than one SV, write all.)

   Answer:

   The hard margin SVM places the decision boundary midway between the lone positive and the closest negative. Thus the support vectors are at $\boxed{2.5, 4.2}$.

6. Consider the following dataset $\mathcal{D}$ of 5 samples in 2D, $\mathcal{D} = \{[1,2]^T, [2,1]^T, [3,3]^T, [4,4]^T, [5,5]^T\}$. The covariance of $\mathcal{D}$ is ————————

   Answer:

   The mean and variance of the sequence $\{1, 2, \ldots, 5\}$ are 3 and 2 respectively (easy to see). The covariance can be computed as follows.

   $$\text{Cov}(x, y) = \frac{2 + 2 + 0 + 1 + 4}{5} = 1.8$$

   The covariance matrix can thus be written as follows.

   $$\Sigma(\mathcal{D}) = \boxed{\begin{bmatrix}2 & 1.8\\1.8 & 2\end{bmatrix}}$$

7. A dataset $\mathcal{D}$ consists of six feature measurements. $\mathcal{D} = \{3.2, 4.5, 5.8, 2.7, 3.8, 4.0\}$. $\mathcal{D}$ needs to be normalized such that mean is zero and variance is unity. The normalized $\mathcal{D}$ is ————————————

   Answer:

   The mean and variance can be computed as follows.

   $$\mu = \frac{3.2 + 4.5 + 5.8 + 2.7 + 3.8 + 4.0}{6} = 4$$

   $$\sigma^2 = \frac{(3.2-4)^2 + (4.5-4)^2 + (5.8-4)^2 + (2.7-4)^2 + (3.8-4)^2 + (4.0-4)^2}{6} \approx 0.98$$

   Finally, the normalized dataset can be computed as follows.

   $$\left\{\frac{3.2-4}{0.99}, \frac{4.5-4}{0.99}, \frac{5.8-4}{0.99}, \frac{2.7-4}{0.99}, \frac{3.8-4}{0.99}, \frac{4.0-4}{0.99}\right\} \approx \boxed{\{-0.80, 0.51, 1.82, -1.31, -0.20, 0\}}$$

8. We know for a valid kernel $\kappa(\mathbf{p}, \mathbf{q})$, there exists a $\phi(\mathbf{p})^T\phi(\mathbf{q})$ form where ϕ does an appropriate feature transformation on $\mathbf{p}$ and $\mathbf{q}$. When $\mathbf{p} = [1,2]^T$; $\mathbf{q} = [2,1]^T$ and $\kappa(\mathbf{p}, \mathbf{q}) = (\mathbf{p}^T\mathbf{q} + 1)^2$, find the corresponding ϕ and write as $\phi(\mathbf{p})^T\phi(\mathbf{q})$ ———————— (such as $[-----------]^T[-----------]$)

   Answer:

The kernel can be expanded as follows.

$$\begin{aligned}\kappa(\mathbf{p},\mathbf{q}) &= (\mathbf{p}^T\mathbf{q}+1)^2 \\ &= (p_1q_1+p_2q_2+1)^2 \\ &= p_1^2q_1^2+2p_1p_2q_1q_2+p_2^2q_2^2+2p_1q_1+2p_2q_2+1\end{aligned}$$

We would like to separate the function $\kappa(\mathbf{p},\mathbf{q})$ as $\phi(\mathbf{p})^T\phi(\mathbf{q})$. Since the function is symmetric in $\mathbf{p}$ and $\mathbf{q}$, this can be achieved by matching the terms in the sum as follows.

$$\phi(\mathbf{p}) = \boxed{\begin{bmatrix} p_1^2 \\ \sqrt{2}\,p_1p_2 \\ p_2^2 \\ \sqrt{2}\,p_1 \\ \sqrt{2}\,p_2 \\ 1 \end{bmatrix}}$$

Note that since inner product is invariant to permutation, any permutation of the feature transformation would meet the requirement.

Quiz II

Name: **Roll Number**.....

- Answer precisely in the space given. No overwriting. Make assumptions, if really ambiguous.
- Do rough work in the additional sheet.

Fill in the blank with precise answers, [8 × 5 = 40]

1. Consider a simple neural network with two inputs, one hidden layer with three neurons and one output. No bias. If the input is $x_1 = 2$ and $x_2 = 2$, output is ——— (fact: All weights are 1.0. Activation functions in the hidden layer is ReLU and activation function in the output is linear.)

Solution:-

The neural network processes the inputs as follows:

1. Hidden Layer Calculation:

- Each hidden neuron receives inputs $x_1 = 2$ and $x_2 = 2$ with weights of 1.0.
- Pre-activation: $(2 \times 1) + (2 \times 1) = 4$.
- ReLU activation: $\text{ReLU}(4) = 4$.
- All three hidden neurons output 4.

2. Output Layer Calculation:

- The output neuron sums the three hidden layer outputs (each multiplied by weight 1.0):

$$4 \times 1 + 4 \times 1 + 4 \times 1 = 12.$$

- Linear activation passes the sum directly.

Output: 12

---

2. Continuing the about question, assume the target (t) of the above sample was 25, and the loss was squared error, $\frac{\partial \mathcal{L}}{\partial w_{11}}$ is ——— (w_{11} is the weight from first input (i.e., x_1) to first neuron in the hidden layer.)

Solution:-

To compute $\frac{\partial \mathcal{L}}{\partial w_{11}}$ using backpropagation:

1. Loss Derivative:

$$\frac{\partial \mathcal{L}}{\partial \hat{y}} = 2(\hat{y} - t) = 2(12 - 25) = -26.$$

2. Output Layer Gradient:

$$\frac{\partial \hat{y}}{\partial h_1} = 1 \quad (\text{since } \hat{y} = h_1 + h_2 + h_3).$$

3. Hidden Layer Activation Gradient:

$$\frac{\partial h_1}{\partial z_1} = 1 \quad \text{(ReLU derivative at } z_1 = 4\text{)}.$$

4. Pre-activation Gradient:

$$\frac{\partial z_1}{\partial w_{11}} = x_1 = 2.$$

Chain Rule:

$$\frac{\partial \mathcal{L}}{\partial w_{11}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_{11}} = (-26) \cdot 1 \cdot 1 \cdot 2 = -52$$

Answer: -52

---

3. Consider a single layer perceptron with two input and one output. The weights from first and second inputs are w_1 and w_2 respectively. Bias is w_0. Also assume a $\pm, 1$ logic. The activation at the output is: $\phi(x) = +1$ if $x \geq 0$ and -1 else. If $w_0 = 1, w_1 = -1, w_2 = -1$, then this neuron model is equivalent to: ————- (fill from the gates like: AND, OR, ExOR, NAND, NOR etc.)

Solution:-

Consider the perceptron with two inputs and one output. The net input to the neuron is given by:

$$x = w_0 \cdot 1 + w_1 x_1 + w_2 x_2$$

Given that $w_0 = 1$, $w_1 = -1$, and $w_2 = -1$, we compute x for all possible input values x_1, x_2 (where $x_1, x_2 \in \{+1, -1\}$):

$$x = 1 - x_1 - x_2$$

The activation function is:

$$\phi(x) = \begin{cases} +1, & \text{if } x \geq 0 \\ -1, & \text{otherwise} \end{cases}$$

We now evaluate the truth table:

x_1	x_2	$x = 1 - x_1 - x_2$	$\phi(x)$
$+1$	$+1$	$1 - 1 - 1 = -1$	-1
$+1$	-1	$1 - 1 - (-1) = 1$	$+1$
-1	$+1$	$1 - (-1) - 1 = 1$	$+1$
-1	-1	$1 - (-1) - (-1) = 3$	$+1$

This matches the truth table of the **NAND** gate (in ± 1 logic):

$$\text{NAND}(x_1, x_2) = \begin{cases} +1, & \text{if } (x_1, x_2) \neq (+1, +1) \\ -1, & \text{if } (x_1, x_2) = (+1, +1) \end{cases}$$

Thus, this perceptron implements the **NAND** gate.

---

4. Consider a binary classification problem (A vs B) with perceptron and bias. Consider two samples from A: $[1,2]^T$, $[4,2]^T$ and one sample from B: $[3,0]^T$. We initialize $\mathbf{w^0} = [1,1,0]^T$. When learning rate $\eta = 0.1$, $\mathbf{w^1}$ (i.e., after one iteration of perceptron algorithm) is ————

Solution:-

We are given:

- Class A samples (label +1): $\mathbf{x}_1 = \begin{bmatrix}1\\2\end{bmatrix}$, $\mathbf{x}_2 = \begin{bmatrix}4\\2\end{bmatrix}$
- Class B sample (label -1): $\mathbf{x}_3 = \begin{bmatrix}3\\0\end{bmatrix}$
- Initial weights: $\mathbf{w}^{(0)} = \begin{bmatrix}1\\1\\0\end{bmatrix}$
- Learning rate: $\eta = 0.1$

We augment each input vector with a bias term of 1:

$$\tilde{\mathbf{x}}_1 = \begin{bmatrix}1\\2\\1\end{bmatrix}, \quad \tilde{\mathbf{x}}_2 = \begin{bmatrix}4\\2\\1\end{bmatrix}, \quad \tilde{\mathbf{x}}_3 = \begin{bmatrix}3\\0\\1\end{bmatrix}$$

Now we apply one iteration of the Perceptron algorithm:

- For $\tilde{\mathbf{x}}_1$, $y = +1$:
$$\mathbf{w}^{(0)} \cdot \tilde{\mathbf{x}}_1 = [1,1,0] \cdot [1,2,1] = 3 > 0 \Rightarrow \text{Correct}$$
- For $\tilde{\mathbf{x}}_2$, $y = +1$:
$$\mathbf{w}^{(0)} \cdot \tilde{\mathbf{x}}_2 = [1,1,0] \cdot [4,2,1] = 6 > 0 \Rightarrow \text{Correct}$$
- For $\tilde{\mathbf{x}}_3$, $y = -1$:
$$\mathbf{w}^{(0)} \cdot \tilde{\mathbf{x}}_3 = [1,1,0] \cdot [3,0,1] = 3 > 0 \Rightarrow \text{Incorrect}$$

  Update rule:
$$\mathbf{w}^{(1)} = \mathbf{w}^{(0)} + \eta \cdot y \cdot \tilde{\mathbf{x}}_3 = \begin{bmatrix}1\\1\\0\end{bmatrix} + 0.1 \cdot (-1) \cdot \begin{bmatrix}3\\0\\1\end{bmatrix} = \begin{bmatrix}0.7\\1\\-0.1\end{bmatrix}$$

Final answer:

$$\boxed{\mathbf{w}^{(1)} = \begin{bmatrix}0.7\\1\\-0.1\end{bmatrix}}$$

---

5. Consider a 1D binary classification problem (2.5,+), (4.2,+), (5.6,+), (7.8,-). We train a hard margin SVM. The support vector(s) is/are ————- (if there are more than one SV, write all.)

Solution:-

We are given the one-dimensional dataset:

$$\{(2.5, +), (4.2, +), (5.6, +), (7.8, -)\}.$$

For a hard-margin SVM with decision function

$$f(x) = wx + b,$$

we must satisfy the constraint

$$y_i(wx_i + b) \geq 1 \quad \text{for all } i.$$

Intuition:

- The positive points 2.5, 4.2, and 5.6 lie to the left, while the negative point 7.8 is to the right.
- To maximize the margin while correctly separating the classes, the decision boundary is determined by the closest points to it—namely, the rightmost positive example and the negative example.
- This implies that the point at $x = 5.6$ (with label $+$) and the point at $x = 7.8$ (with label $-$) will be the support vectors, satisfying:

$$w \cdot 5.6 + b = +1 \quad \text{and} \quad w \cdot 7.8 + b = -1.$$

Thus, the support vectors that define the optimal margin are:

$$\boxed{(5.6, +) \quad \text{and} \quad (7.8, -).}$$

---

6. Consider the following dataset $\mathcal{D}$ of 5 samples in 2D, $\mathcal{D} = \{[1,1]^T, [2,2]^T, [3,3]^T, [4,5]^T, [5,4]^T\}$. The covariance of $\mathcal{D}$ is ————

Solution: Computing the 2×2 Sample Covariance Matrix

Given the data set

$$D = \{(1,1), (2,2), (3,3), (4,5), (5,4)\},$$

we will compute the 2×2 sample covariance matrix.

Step 1: Compute the Sample Mean

Let the number of data points be $N = 5$. The sample mean is given by

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

Writing out the coordinates:

$$\bar{x} = \frac{1}{5}\Big[(1,1) + (2,2) + (3,3) + (4,5) + (5,4)\Big].$$

Compute the sum component-wise:

$$\sum_{i=1}^{5} x_i = (1 + 2 + 3 + 4 + 5,\ 1 + 2 + 3 + 5 + 4) = (15, 15).$$

Thus,

$$\bar{x} = \frac{1}{5}(15, 15) = (3, 3).$$

Step 2: Center Each Data Point

Subtract the mean from each point:

$$\begin{aligned}(1,1)-(3,3)&=(-2,-2),\\(2,2)-(3,3)&=(-1,-1),\\(3,3)-(3,3)&=(0,0),\\(4,5)-(3,3)&=(1,2),\\(5,4)-(3,3)&=(2,1).\end{aligned}$$

Step 3: Compute the Sum of the Outer Products

The sample covariance matrix is given by

$$\Sigma=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})(x_i-\bar{x})^T.$$

We compute the outer products one-by-one:

(a) For $x_1-\bar{x}=(-2,-2)$:

$$(-2,-2)^T(-2,-2)=\begin{pmatrix}(-2)(-2) & (-2)(-2)\\(-2)(-2) & (-2)(-2)\end{pmatrix}=\begin{pmatrix}4 & 4\\4 & 4\end{pmatrix}.$$

(b) For $x_2-\bar{x}=(-1,-1)$:

$$(-1,-1)^T(-1,-1)=\begin{pmatrix}(-1)(-1) & (-1)(-1)\\(-1)(-1) & (-1)(-1)\end{pmatrix}=\begin{pmatrix}1 & 1\\1 & 1\end{pmatrix}.$$

(c) For $x_3-\bar{x}=(0,0)$:

$$(0,0)^T(0,0)=\begin{pmatrix}0 & 0\\0 & 0\end{pmatrix}.$$

(d) For $x_4-\bar{x}=(1,2)$:

$$(1,2)^T(1,2)=\begin{pmatrix}1\cdot 1 & 1\cdot 2\\2\cdot 1 & 2\cdot 2\end{pmatrix}=\begin{pmatrix}1 & 2\\2 & 4\end{pmatrix}.$$

(e) For $x_5-\bar{x}=(2,1)$:

$$(2,1)^T(2,1)=\begin{pmatrix}2\cdot 2 & 2\cdot 1\\1\cdot 2 & 1\cdot 1\end{pmatrix}=\begin{pmatrix}4 & 2\\2 & 1\end{pmatrix}.$$

Summing all these outer products:

$$\begin{aligned}S&=\begin{pmatrix}4 & 4\\4 & 4\end{pmatrix}+\begin{pmatrix}1 & 1\\1 & 1\end{pmatrix}+\begin{pmatrix}0 & 0\\0 & 0\end{pmatrix}+\begin{pmatrix}1 & 2\\2 & 4\end{pmatrix}+\begin{pmatrix}4 & 2\\2 & 1\end{pmatrix}\\&=\begin{pmatrix}4+1+0+1+4 & 4+1+0+2+2\\4+1+0+2+2 & 4+1+0+4+1\end{pmatrix}\\&=\begin{pmatrix}10 & 9\\9 & 10\end{pmatrix}.\end{aligned}$$

Step 4: Divide by $N-1$

Since $N=5$, we have $N-1=4$. The sample covariance matrix is:

$$\Sigma=\frac{1}{4}\begin{pmatrix}10 & 9\\9 & 10\end{pmatrix}=\begin{pmatrix}\frac{10}{4} & \frac{9}{4}\\\frac{9}{4} & \frac{10}{4}\end{pmatrix}=\begin{pmatrix}2.5 & 2.25\\2.25 & 2.5\end{pmatrix}.$$

Final Answer

$$\boxed{\begin{pmatrix} 2.5 & 2.25 \\ 2.25 & 2.5 \end{pmatrix}}$$

---

7. A dataset $\mathcal{D}$ consists of six feature measurements. $\mathcal{D} = \{1.2, 2.5, 3.8, 0.7, 1.8, 2.0\}$. $\mathcal{D}$ needs to be normalized such that mean is zero and variance is unity. The normalized $\mathcal{D}$ is _____

Solution :- Normalization of Dataset D

Let

$$D = \{\, 1.2,\ 2.5,\ 3.8,\ 0.7,\ 1.8,\ 2.0\}.$$

Step 1: Compute the Mean

$$\mu = \frac{1.2 + 2.5 + 3.8 + 0.7 + 1.8 + 2.0}{6} = 2.0.$$

Step 2: Compute the Variance and Standard Deviation

$$\begin{aligned}
\sigma^2 &= \frac{1}{6}\sum_{i=1}^{6}(x_i - \mu)^2 \\
&= \frac{(1.2-2.0)^2 + (2.5-2.0)^2 + (3.8-2.0)^2 + (0.7-2.0)^2 + (1.8-2.0)^2 + (2.0-2.0)^2}{6} \\
&= \frac{0.64 + 0.25 + 3.24 + 1.69 + 0.04 + 0}{6} \approx 0.9767, \\
\sigma &= \sqrt{\sigma^2} \approx 0.9883.
\end{aligned}$$

Step 3: Compute the Normalized Values

For each $x_i \in D$, define

$$z_i = \frac{x_i - \mu}{\sigma}.$$

The computations are summarized in the table below:

x_i	$x_i - \mu$	$z_i = \dfrac{x_i - \mu}{\sigma}$
1.2	-0.8	-0.81
2.5	$+0.5$	$+0.51$
3.8	$+1.8$	$+1.82$
0.7	-1.3	-1.32
1.8	-0.2	-0.20
2.0	0.0	0.00

Hence, the normalized dataset is

$$\{\, -0.81,\ 0.51,\ 1.82,\ -1.32,\ -0.20,\ 0.00\}.$$

---

8. We know for a valid kernel $\kappa(\mathbf{p}, \mathbf{q})$, there exists a $\phi(\mathbf{p})^T \phi(\mathbf{q})$ form where ϕ does an appropriate feature transformation on $\mathbf{p}$ and $\mathbf{q}$. When $\mathbf{p} = [3, 1]^T$; $\mathbf{q} = [2, 3]^T$ and $\kappa(\mathbf{p}, \mathbf{q}) = (\mathbf{p}^T \mathbf{q} + 1)^2$, find the corresponding ϕ and write as $\phi(\mathbf{p})^T \phi(\mathbf{q})$ ———————— (such as $[---------]^T[-----------]$)

Solution :-

We are given the kernel:

$$\kappa(\mathbf{p}, \mathbf{q}) = (\mathbf{p}^T \mathbf{q} + 1)^2$$

This is a polynomial kernel of degree 2 with bias term 1. We aim to find a feature transformation $\phi(\cdot)$ such that:

$$\kappa(\mathbf{p}, \mathbf{q}) = \phi(\mathbf{p})^T \phi(\mathbf{q})$$

Let:

$$\mathbf{p} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \quad \mathbf{q} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

The expansion of the kernel is:

$$(\mathbf{p}^T \mathbf{q} + 1)^2 = (p_1 q_1 + p_2 q_2 + 1)^2$$

We can define the feature map ϕ as:

$$\phi(\mathbf{x}) = \begin{bmatrix} x_1^2 \\ \sqrt{2} x_1 x_2 \\ x_2^2 \\ \sqrt{2} x_1 \\ \sqrt{2} x_2 \\ 1 \end{bmatrix}$$

Thus, we compute:

$$\phi(\mathbf{p}) = \begin{bmatrix} 9 \\ 3\sqrt{2} \\ 1 \\ 3\sqrt{2} \\ \sqrt{2} \\ 1 \end{bmatrix}, \quad \phi(\mathbf{q}) = \begin{bmatrix} 4 \\ 6\sqrt{2} \\ 9 \\ 2\sqrt{2} \\ 3\sqrt{2} \\ 1 \end{bmatrix}$$

Finally, the dot product gives:

$$\phi(\mathbf{p})^T \phi(\mathbf{q}) = (3 \cdot 2 + 1 \cdot 3 + 1)^2 = (6 + 3 + 1)^2 = 10^2 = 100$$

To find the feature transformation ϕ such that $\kappa(\mathbf{p}, \mathbf{q}) = (\mathbf{p}^T \mathbf{q} + 1)^2 = \phi(\mathbf{p})^T \phi(\mathbf{q})$, we expand the kernel expression:

$$(\mathbf{p}^T \mathbf{q} + 1)^2 = (p_1 q_1 + p_2 q_2 + 1)^2 = p_1^2 q_1^2 + p_2^2 q_2^2 + 1 + 2 p_1 q_1 p_2 q_2 + 2 p_1 q_1 + 2 p_2 q_2.$$

This corresponds to the dot product of the transformed feature vectors:

$$\phi(\mathbf{p}) = \begin{bmatrix} p_1^2 \\ p_2^2 \\ \sqrt{2} p_1 p_2 \\ \sqrt{2} p_1 \\ \sqrt{2} p_2 \\ 1 \end{bmatrix}, \quad \phi(\mathbf{q}) = \begin{bmatrix} q_1^2 \\ q_2^2 \\ \sqrt{2} q_1 q_2 \\ \sqrt{2} q_1 \\ \sqrt{2} q_2 \\ 1 \end{bmatrix}.$$

Thus, the required form is:

$$\boxed{\begin{bmatrix} p_1^2 & p_2^2 & \sqrt{2}p_1p_2 & \sqrt{2}p_1 & \sqrt{2}p_2 & 1 \end{bmatrix}^T \begin{bmatrix} q_1^2 & q_2^2 & \sqrt{2}q_1q_2 & \sqrt{2}q_1 & \sqrt{2}q_2 & 1 \end{bmatrix}}.$$

Quiz II

Name: **Roll Number**.....

- Answer precisely in the space given. No overwriting. Make assumptions, if really ambiguous.
- Do rough work in the additional sheet.

Fill in the blank with precise answers: $[8 \times 5 = 40]$

1. Consider a simple neural network with two inputs, one hidden layer with three neurons and one output. No bias. If the input is $x_1 = 4$ and $x_2 = 2$, output is 18 (fact: All weights are 1.0. Activation functions in the hidden layer is ReLU and activation function in the output is linear.)

2. Continuing the about question, assume the target (t) of the above sample was 25, and the loss was squared error, $\frac{\partial \mathcal{L}}{\partial w_{11}}$ is -56 (w_{11} is the weight from first input (i.e., x_1) to first neuron in the hidden layer.)

3. Consider a single layer perceptron with two input and one output. The weights from first and second inputs are w_1 and w_2 respectively. Bias is w_0. Also assume a $\pm, 1$ logic. The activation at the output is: $\phi(x) = +1$ if $x \geq 0$ and -1 else. If $w_0 = 1, w_1 = -1, w_2 = -1$, then this neuron model is equivalent to: NAND.

4. Consider a binary classification problem (A vs B) with perceptron and bias. Consider two samples from A: $[1, 2]^T, [4, 2]^T$ and one sample from B: $[3, 0]^T$. We initialize $\mathbf{w^0} = [1, 1, 0]^T$. When learning rate $\eta = 0.1$, $\mathbf{w^1}$ (i.e., after one iteration of perceptron algorithm) is $[0.7, 1, -0.1]^T$

5. Consider a 1D binary classification problem (2.5,+), (4.8,+), (5.8,-), (7.8,-). We train a hard margin SVM. The support vector(s) is/are (4.8, +) and (5.8, -).

6. Consider the following dataset $\mathcal{D}$ of 5 samples in 2D, $\mathcal{D} = \{[1, 2]^T, [2, 1]^T, [3, 3]^T, [4, 5]^T, [5, 4]^T\}$. The covariance of $\mathcal{D}$ is

$$\boxed{\begin{pmatrix} 2.5 & 2.0 \\ 2.0 & 2.5 \end{pmatrix}}$$

or

$$\boxed{\begin{pmatrix} 2.0 & 1.6 \\ 1.6 & 2.0 \end{pmatrix}}$$

7. A dataset $\mathcal{D}$ consists of six feature measurements. $\mathcal{D} = \{4.2, 5.5, 6.8, 3.7, 4.8, 5.0\}$. $\mathcal{D}$ needs to be normalized such that mean is zero and variance is unity. The normalized $\mathcal{D}$ is

$$\{-0.81,\ 0.51,\ 1.82,\ -1.32,\ -0.20,\ 0.00\}.$$

8. We know for a valid kernel $\kappa(\mathbf{p}, \mathbf{q})$, there exists a $\phi(\mathbf{p})^T \phi(\mathbf{q})$ form where ϕ does an appropriate feature transformation on $\mathbf{p}$ and $\mathbf{q}$. When $\mathbf{p} = [3, 3]^T$; $\mathbf{q} = [4, 4]^T$ and $\kappa(\mathbf{p}, \mathbf{q}) = (\mathbf{p}^T \mathbf{q} + 1)^2$, find the corresponding ϕ and write as $\phi(\mathbf{p})^T \phi(\mathbf{q})$ ———————— (such as $[---------]^T[----------]$)

Answer:

The kernel can be expanded as follows.

$$\begin{aligned}\kappa(\mathbf{p}, \mathbf{q}) &= (\mathbf{p}^T \mathbf{q} + 1)^2 \\ &= (p_1 q_1 + p_2 q_2 + 1)^2 \\ &= p_1^2 q_1^2 + 2p_1 p_2 q_1 q_2 + p_2^2 q_2^2 + 2p_1 q_1 + 2p_2 q_2 + 1\end{aligned}$$

We would like to separate the function $\kappa(\mathbf{p}, \mathbf{q})$ as $\phi(\mathbf{p})^T \phi(\mathbf{q})$. Since the function is symmetric in $\mathbf{p}$ and $\mathbf{q}$, this can be achieved by matching the terms in the sum as follows.

$$\phi(\mathbf{p}) = \boxed{\begin{bmatrix} p_1^2 \\ \sqrt{2}\, p_1 p_2 \\ p_2^2 \\ \sqrt{2}\, p_1 \\ \sqrt{2}\, p_2 \\ 1 \end{bmatrix}}$$

i.e.

$$\phi(\mathbf{p}) = \boxed{\begin{bmatrix} 9 \\ 9\sqrt{2} \\ 9 \\ 3\sqrt{2} \\ 3\sqrt{2} \\ 1 \end{bmatrix}}$$

and

$$\phi(\mathbf{q}) = \boxed{\begin{bmatrix} q_1^2 \\ \sqrt{2}\, q_1 q_2 \\ q_2^2 \\ \sqrt{2}\, q_1 \\ \sqrt{2}\, q_2 \\ 1 \end{bmatrix}}$$

i.e.

$$\phi(\mathbf{q}) = \boxed{\begin{bmatrix} 16 \\ 16\sqrt{2} \\ 16 \\ 4\sqrt{2} \\ 4\sqrt{2} \\ 1 \end{bmatrix}}$$

Note that since inner product is invariant to permutation, any permutation of the feature transformation would meet the requirement.