Question2

Proof

Given:

$$K_j(x,z) = \langle \Phi_j(x), \Phi_j(z) \rangle \tag{1}$$

for $j = 1, 2$.

We want to find if:

$$K(x,z) = K_1(x,z)K_2(x,z) \tag{2}$$

is a valid kernel.

Using the given information, we can express $K(x,z)$ as:

$$K(x,z) = \langle \Phi_1(x), \Phi_1(z) \rangle \langle \Phi_2(x), \Phi_2(z) \rangle \tag{3}$$

Now, let's expand the inner products:

$$\langle \Phi_1(x), \Phi_1(z) \rangle = \sum_{i=1}^{D} \Phi_1(x)^i \Phi_1(z)^i \tag{4}$$

$$\langle \Phi_2(x), \Phi_2(z) \rangle = \sum_{i=1}^{D} \Phi_2(x)^i \Phi_2(z)^i \tag{5}$$

The product of these two sums is:

$$K(x,z) = \left(\sum_{i=1}^{D} \Phi_1(x)^i \Phi_1(z)^i \right) \left(\sum_{j=1}^{D} \Phi_2(x)^j \Phi_2(z)^j \right) \tag{6}$$

Expanding this product, we get:

$$K(x,z) = \sum_{i=1}^{D} \sum_{j=1}^{D} \Phi_1(x)^i \Phi_1(z)^i \Phi_2(x)^j \Phi_2(z)^j \tag{7}$$

Now, consider a new feature map $\Psi : R^d \to R^{D^2}$ defined as:

$$\Psi(x) = \left[\Phi_1(x)^1 \Phi_2(x)^1, \Phi_1(x)^2 \Phi_2(x)^2, \ldots, \Phi_1(x)^D \Phi_2(x)^D \right] \tag{8}$$

The inner product in this new space is:

$$\langle \Psi(x), \Psi(z) \rangle = \sum_{i=1}^{D} \sum_{j=1}^{D} \Phi_1(x)^i \Phi_1(z)^i \Phi_2(x)^i \Phi_2(z)^i \tag{9}$$

This is exactly the expression for $K(x,z)$ that we derived earlier!

Thus, $K(x,z) = K_1(x,z)K_2(x,z)$ is indeed a valid kernel function, as it can be expressed as an inner product in a new high-dimensional space.

Question7

Solve

Given:

$$J(w,b) = \sum_{i=1}^{N} ||x_i^T w + b - y_i||_2^2 + \lambda(||w||_2^2 + b^2) \tag{10}$$

To find the optimal values of w and b, we'll differentiate the objective function with respect to w and b, and set the derivatives to zero.

1. Differentiating with respect to w:

$$\frac{\partial J}{\partial w} = 2\sum_{i=1}^{N} x_i(x_i^T w + b - y_i) + 2\lambda w \tag{11}$$

Setting this to zero, we get:

$$\sum_{i=1}^{N} x_i(x_i^T w + b - y_i) + \lambda w = 0 \tag{12}$$

2. Differentiating with respect to b:

$$\frac{\partial J}{\partial b} = 2\sum_{i=1}^{N}(x_i^T w + b - y_i) + 2\lambda b \tag{13}$$

Setting this to zero, we get:

$$\sum_{i=1}^{N}(x_i^T w + b - y_i) + \lambda b = 0 \tag{14}$$

Rewrite the above equations in matrix form.

Let X be the data matrix of size $N \times D$ where each row is a data sample x_i, and let y be the column vector of target values.

From (12), we can rewrite the equation in matrix form as:

$$X^T(Xw + b\mathbf{1} - y) + \lambda w = 0 \tag{15}$$

Where $\mathbf{1}$ is a column vector of ones of size $N \times 1$

From (14), summing over all data points, we get:

$$\mathbf{1}^T(Xw + b\mathbf{1} - y) + \lambda b = 0 \tag{16}$$

These are the normal equations for this regularized linear regression problem. Solve them simultaneously to get the values of w and b.

Question8

Solve

Layer 1: Convolutional Layer

- Filter size:
- $5 \times 5 \times (height \times width)$
- Number of filters: 100
- Input channels: 3
- Bias terms: There is one bias term per filter.

Parameters per filter $= (5 \times 5 \times 3) + 1 = 76$
Total parameters in Layer $1 = 76 \times 100 = 7600$

Layer 2: Max-Pooling Layer

Max-pooling layers do not have parameters that are learned during training. Their purpose is to reduce the spatial dimensions of the input.

Layer 3: Convolutional Layer

- Filter size: 3×3
- Number of filters: 50
- Input channels: This depends on the output of Layer 1. Each filter in Layer 1 produces one feature map, so there are 100 feature maps feeding into Layer 3.
- Bias terms: One per filter.

Parameters per filter in Layer $3 = (3 \times 3 \times 100) + 1 = 901$
Total parameters in Layer $3 = 901 \times 50 = 45050$

Summary

- Layer 1: 7600 parameters
- Layer 2: 0 parameters
- Layer 3: 45050 parameters