The problem statement can be found as Homework 4.pdf in the same folder.

Answer to Problem 1 - Support Vector Machine

a) No such linear separator. If the classifier is correct without errors, the data points belonging to the same class should be posed one side in the line. However, the point $(2, 1)$ is not on the left of points (-1,-1), so it is not separable.

b) Yes, there are half-space that achieves 0 classification errors. $\hat{S} = (1, -1), (1, -1), (4, 1), (4, 1)$. The equation of the max-margin linear separator is $x_1 = 2.5$, where the corresponding margin is 1.5.

c) According to the definition of kernel, we have

$$K(x, z) = \phi(x)^T \phi(z) = (x, x^2)^T (z, z^2) = xz + x^2 z^2$$

Answer to Problem 2

1) $\sum_i \xi_i$

2) The role of C is a trade-off to balance the correctness of points being classified and a large margin. When $C \to 0$, no penalty is introduced, which means that a large margin will emerge. Otherwise, $C \to \infty$, complex optimized objective function will ensure that a high accurate classification result.

3) We not calculate $\phi(x))$ directly as the large time complexity of $\mathcal{O}(d^2)$. We instead use a kernel trick, $K(x, z) = \phi(x)^T \phi(z) = (x^T z)^2$ with a compared small time complexity of $\mathcal{O}(d)$, where $\phi(x) \in \mathbb{R}^{d^2}$. When we predict the category for a new data point, we have $\hat{y} = sign(W^T \phi(X)) = sign(\sum_{i=1}^{n} a_i y_i \phi(x_i)^T \phi(X))$, where the $\phi(x_i)^T \phi(X))$ can be obtained as a kernel calculation.

Answer to Problem 3 - Prove if kernels

1) A valid kernel function, $K(x_i, x_j)$, should follows:

$$n \in \mathbb{N}^+, \forall \{x_i\}_{i=1}^n \in X^n, x_i \in \mathbb{R}^d, \forall \{a_i\}_{i=1}^n \in \mathbb{R}^n, \sum_i \sum_j a_i a_j K(x_i, x_j) \geq 0 \tag{1}$$

Now we plan to prove

$$K(x, z) = c_1 K_1(x, z) + c_2 K_2(x, z) \tag{2}$$

The right side of equality (2), marked as T, is like follows if you substitute it with the equality (1) as $K1, G2$ are kernels

$$\begin{aligned} T &= \sum_i \sum_j a_i a_j K_1(x_i, x_j) + \sum_i \sum_j a_i a_j K_2(x_i, x_j) \\ &= \sum_i \sum_j a_i a_j [K_1(x_i, x_j) + K_2(x_i, x_j)] \\ &\geq 0 \end{aligned} \tag{3}$$

So T is a kernel, which means $K(x, z)$ is a kernel, too.

2) Prove $K(x, z) = K_1(x, z) \cdot K_2(x, z)$ $K_1(x, z)$ is a kernel, which means $\exists$ feature maps, $\phi_1 : \mathbb{R} \to R^{D_1}, \phi_2 : \mathbb{R} \to \mathbb{R}^{D_2}$, let $K_1(x, z) = \phi_1(x, z)^T \phi_1(x, z), K_2(x, z) = \phi_2(x, z)^T \phi(x, z)$. We compose

$\phi : \mathbb{R} \to \mathbb{R}^{D_1 \cdot D_2}$, let $\phi(x) = \phi_{11}(x)\phi_{21}(x) + \phi_{11}(x)\phi_{22}(x) + \ldots + \phi_{1D_1}(x)\phi_{1D_2}(x)$, then

$$\begin{aligned}
\phi(x)^T \phi(z) &= \sum_{i}^{D1} \sum_{j}^{D2} \phi_{1i}(x)\phi_{2j}(x)\phi_{1i}(z)\phi_{2j}(z) \\
&= \sum^{D_1} \phi_{1i}(x)\phi_{1i}(z) \sum^{D_2} \phi_{2i}(x)\phi_{2j}(z) \\
&= \phi_1(x)^T \phi_1(z) \cdot \phi_2(z)^T \phi_2(z) \\
&= K_1(x, z) K_2(x, z) \\
&= K(x, z)
\end{aligned}$$

So, we find a feature map to obtain the kernel $K(x, z)$.

Answer to Problem 4 - Generalized Lagrangian Function

The objective we need to prove as follows:

$$\max_{\alpha \geq 0, \beta} \min_{w} \mathcal{L}(w, \alpha, \beta)) \leq \min_{w} \max_{\alpha \geq 0, \beta} \mathcal{L}(w, \alpha, \beta)$$

First, define $p(\alpha, \beta) = \min_{w} \mathcal{L}(w, \alpha, \beta)$, which to minimize the w by fixing the α, β. $p(\alpha, \beta)$ is a function of parameters of α, β. For any w, we have relation below,

$$p(\alpha, \beta) \leq \mathcal{L}(w, \alpha, \beta) \tag{4}$$

Then, define $q(w) = \max_{\alpha \geq 0, \beta} \mathcal{L}(w, \alpha, \beta)$ as a function of parameter w.
Next, apply $\max_{\alpha \geq 0, \beta}$ to both sides of inequality (1),

$$\max_{\alpha \geq 0, \beta} p(\alpha, \beta) \leq \max_{\alpha \geq 0, \beta} \mathcal{L}(w, \alpha, \beta) = q(w) \tag{5}$$

From inequality (2), the left side is a constant, while the right one is a function of parameter w If we apply $\min_{w}$ operation to both sides, the left side remains

$$\max_{\alpha \geq 0, \beta} p(\alpha, \beta) \leq \min_{w} q(w) \tag{6}$$

Finally, use $p(\alpha, \beta), q(w)$ to substitute inequality (3), The objective can be proved.

Answer to Problem 5 - PCA

1) The first principal component can be either $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^T$ or $(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})^T$. The intuition is that the first principal components lies in one line that distributes data most widely. Here, we just link origin points and the other two points and normalized them as the first principal component.

2) The transformed coordinates for three points are $-\sqrt{2}, 0, \sqrt{2}$. You can solve by geometry intuition or inner dot between $x^T v$, where x is the coordinates of points and v is the first principal vectors. The variance of the projected points is 2, which can be the farthest distance between points in the line or the average sum of difference between points and the their mean value.

3) The reconstruction error is 0 since all points lies in the direction of the first principal component.