Homework #4 solution key

Prayag
Neural networks and learning systems-I

April 23, 2019

Problem 6.1**.**

Solution. The given data $\mathbf{x}$ is linearly separable and the separating hyperplane is given by $\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$ where $\mathbf{w}$ denotes the weight vector and b denotes the bias. The hyperplane is said to correspond to a canonical pair $(\mathbf{w}, b)$ if for the set of input patters $\{\mathbf{x}_i\}_{i=1}^{N}$ satisfies

$$\min_{i=1,2,\ldots,N} |\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b| = 1. \tag{1}$$

Let $\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b = g(\mathbf{x}_i)$ where y_i gives the distance of the input data $\mathbf{x}_i$ from the separating hyperplane. We know that any point $\mathbf{x}_i$ can be decomposed into two components as given below:

$$\mathbf{x}_i = \mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}$$

where $\mathbf{x}_p$ is the normal projection of the point $\mathbf{x}$ on to the hyperplane and r is the distance of the data point from the hyperplane.

$$\begin{aligned} g(\mathbf{x}_i) &= \mathbf{w}^{\mathrm{T}}\left(\mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) + b \\ &= \mathbf{w}^{\mathrm{T}}\mathbf{x}_p + r\frac{\mathbf{w}^{\mathrm{T}}\mathbf{w}}{\|\mathbf{w}\|} + b \\ &= \mathbf{w}^{\mathrm{T}}\mathbf{x}_p + b + r\|\mathbf{w}\| \\ &= g(\mathbf{x}_p) + r\|\mathbf{w}\| \\ &= r\|\mathbf{w}\| \qquad \text{Since } g(\mathbf{x}_p) = 0. \\ \implies \quad r &= \frac{g(\mathbf{x}_i)}{\|\mathbf{w}\|} \end{aligned}$$

From (1), we know that there exists at least one $\mathbf{x}_i$ such that $\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b = 1$ or $\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b = -1$. Therefore $g(\mathbf{x}_i) = \pm 1$. Therefore

$$r = \begin{cases} \frac{1}{\|\mathbf{w}\|} & \text{if Class 1} \\ -\frac{1}{\|\mathbf{w}\|} & \text{if Class -1} \end{cases}$$

The optimal separation between the two classes is given by $2r = \frac{2}{\|\mathbf{w}\|}$. ■

Problem 6.3**.**

Solution. Given problem:

$$\begin{aligned} \min \quad & \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{N}\zeta_i \\ \text{subject to} \quad & \zeta_i \geq 0 \ \ \forall i = 1, 2, \ldots, N \\ & d_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \zeta_i \ \ \forall i = 1, 2, \ldots, N \end{aligned}$$

Writing this in the standard form to write the Lagrange, we get

$$\begin{aligned} \min \quad & \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{N}\zeta_i \\ \text{subject to} \quad & \zeta_i \geq 0 \ \ \forall i = 1, 2, \ldots, N \\ & d_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - 1 + \zeta_i \geq 0 \ \ \forall i = 1, 2, \ldots, N \end{aligned}$$

The Lagrange can now be written as follows using the Lagrange multipliers λ_i and α_i as

$$L = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{N}\zeta_i - \sum_{i=1}^{N}\lambda_i\zeta_i - \sum_{i=1}^{N}\alpha_i\left(d_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - 1 + \zeta_i\right)$$

$$L = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{N}\zeta_i - \sum_{i=1}^{N}\lambda_i\zeta_i - \sum_{i=1}^{N}\alpha_i d_i\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - \sum_{i=1}^{N}\alpha_i d_i b + \sum_{i=1}^{N}\alpha_i - \sum_{i=1}^{N}\alpha_i\zeta_i$$

Differentiating the Lagrange and equating to zero, we get

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \implies \mathbf{w} - \sum_{i=1}^{N}\alpha_i d_i\mathbf{x}_i = 0 \implies \mathbf{w} = \sum_{i=1}^{N}\alpha_i d_i\mathbf{x}_i$$

$$\frac{\partial L}{\partial \mathbf{b}} = 0 \implies \sum_{i=1}^{N}\alpha_i d_i = 0$$

$$\frac{\partial L}{\partial \zeta_i} = 0 \implies C - \lambda_i - \alpha_i = 0 \implies C = \lambda_i + \alpha_i$$

Substituting the above in the Lagrange, we get

$$\begin{aligned} L = & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j d_i d_j\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j + \sum_{i=1}^{N}(\lambda_i + \alpha_i)\zeta_i - \sum_{i=1}^{N}\lambda_i\zeta_i - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j d_i d_j\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j - \sum_{i=1}^{N}\alpha_i d_i b + \sum_{i=1}^{N}\alpha_i \\ & - \sum_{i=1}^{N}\alpha_i\zeta_i \\ = & \left(\frac{1}{2} - 1\right)\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j d_i d_j\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j + \left[\sum_{i=1}^{N}(\lambda_i + \alpha_i) - \sum_{i=1}^{N}\lambda_i - \sum_{i=1}^{N}\alpha_i\right]\zeta_i - b\sum_{i=1}^{N}\alpha_i d_i + \sum_{i=1}^{N}\alpha_i \\ = & -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j d_i d_j\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j + \sum_{i=1}^{N}\alpha_i \end{aligned}$$

From this, the dual can be written as follows

$$\begin{aligned} \max \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j \\ \text{subject to} \quad & \left.\begin{array}{l} \sum_{i=1}^{N} \alpha_i d_i = 0 \\ C - \lambda_i - \alpha_i = 0 \\ \alpha_i \geq 0 \\ \lambda_i \geq 0 \end{array}\right\} \quad \forall i = 1, 2, \ldots, N \end{aligned}$$

We observe that the Lagrange multiplier λ_i appears only in the constraint $C - \lambda_i - \alpha_i \implies \lambda_i = C - \alpha_i$. For $\lambda_i \geq 0$ to be true, $C - \alpha_i \geq 0 \implies C \geq \alpha_i$. Combining this with $\alpha_i \geq 0$ constraint, we get $0 \leq \alpha_i \leq C$. The dual problem can now be written as

$$\begin{aligned} \max \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j \\ \text{subject to} \quad & \left.\begin{array}{l} \sum_{i=1}^{N} \alpha_i d_i = 0 \\ 0 \leq \alpha_i \leq C \end{array}\right\} \quad \forall i = 1, 2, \ldots, N \end{aligned}$$

■

Problem 6.11**.**

Solution. It is given that a joint probability density function $p_{X_1,X_2}(x_1, x_2)$ over an $\mathcal{H}$-by-$\mathcal{H}$ product space is said to be a *P-matrix* provided it satisfies finitely positive semidefinite property. The matrix P will be positive semidefinite if for every non-zero column vector $\mathbf{z}$, the value obtained from $\mathbf{z}^{\mathrm{T}} P \mathbf{z}$ is positive or zero.
Let us consider the simple case of two-element set $\mathbf{X} = [X_1, X_2]$ of random variables.
Case 1: Are all P-kernels joint distributions?
From a given $P-$kernel $P(x, y)$, we can generate an identical kernel the $\hat{P}$-kernel if it satisfies $\sum_{x \in X} \sum_{y \in X} P(x, y) = C$, where C is some constant such that $C < \infty$. We can define the $\hat{P}$-kernels as $\hat{P}(x, y) = \frac{1}{C} P(x, y)$. This definition satisfies the properties of a *P-matrix* since we have only scaled the elements. Since $\hat{P}(x, y)$ is also a joint distribution, we can say that all P-kernels are joint distributions.
Case 2: Are all joint distributions P-kernels?
Considering the two element case, let us create a joint distribution and verify if it satisfies the properties of a P-kernel. The joint probability matrix for a two element case would be given by

$$\mathbf{P}_{X,Y} = \begin{bmatrix} p(x_1, x_1) & p(x_1, x_2) \\ p(x_2, x_1) & p(x_2, x_2) \end{bmatrix}$$

Considering a particular case where $p(x_1, x_1) = 0,\ p(x_1, x_2) = 0.5,\ p(x_2, x_1) = 0.5,\ p(x_2, x_2) = 0$, we get the

$$\mathbf{P}_{X,Y} = \begin{bmatrix} 0 & 0.5 \\ 0.5 & 0 \end{bmatrix}$$

Solving for the eigenvalues, we get $\lambda = \pm 0.5$. From the given definition, the *P-matrix* must be positive semidefinite, but an eigenvalue in the above case is negative. Therefore not all joint distributions are P-kernels.

■

Problem 6.21**.**

Solution. Given $k(\mathbf{x}_i, .)$ and $k(\mathbf{x}_j, .)$ denote a pain of kernels, where $i, j = 1, 2, \ldots, N$ and the vectors have the same dimensionality. We need to show that

$$\langle k(\mathbf{x}_i, .), k(\mathbf{x}_j, .) \rangle = k(\mathbf{x}_i, \mathbf{x}_j). \tag{2}$$

Let $f(.)$ and $g(.)$ be two functions defined over a vector space $\mathcal{F}$ such that

$$f(.) = \sum_{i=1}^{N} a_i k(\mathbf{x}_i, .), \tag{3}$$

$$g(.) = \sum_{j=1}^{N} b_j k(\mathbf{x}_j, .) \tag{4}$$

where $k(\mathbf{x}, .)$ is a mercer kernel. We can write

$$f(\mathbf{x}_j) = \sum_{i=1}^{N} a_i k(\mathbf{x}_i, \mathbf{x}_j), \tag{5}$$

$$g(\mathbf{x}_i) = \sum_{j=1}^{N} b_j k(\mathbf{x}_j, \mathbf{x}_i). \tag{6}$$

We know that

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi^{\mathrm{T}}(\mathbf{x}_i)\phi(\mathbf{x}_j) \tag{7}$$

Using (7) in (5) and (6), we get

$$f(\mathbf{x}_j) = \sum_{i=1}^{N} a_i \phi^{\mathrm{T}}(\mathbf{x}_i)\phi(\mathbf{x}_j), \tag{8}$$

$$g(\mathbf{x}_i) = \sum_{j=1}^{N} b_j \phi^{\mathrm{T}}(\mathbf{x}_i)\phi(\mathbf{x}_j). \tag{9}$$

From the above, we obtain the functions below

$$f(.) = \sum_{i=1}^{N} a_i \phi(\mathbf{x}_i), \tag{10}$$

$$g(.) = \sum_{j=1}^{N} b_j \phi(\mathbf{x}_j) \tag{11}$$

Taking the inner product using (10) and (11), we get

$$\langle f, g \rangle = \left(\sum_{i=1}^{N} a_i \phi(\mathbf{x}_i) \right)^{\mathrm{T}} \left(\sum_{j=1}^{N} b_j \phi(\mathbf{x}_j) \right)$$

$$\langle f, g \rangle = \sum_{i=1}^{N} \sum_{j=1}^{N} a_i b_j \phi^{\mathrm{T}}(\mathbf{x}_i) \phi(\mathbf{x}_j)$$

$$\langle f, g \rangle = \sum_{i=1}^{N} \sum_{j=1}^{N} a_i b_j k(\mathbf{x}_i, \mathbf{x}_j) \tag{12}$$

Taking the inner product using (3) and (4), we get

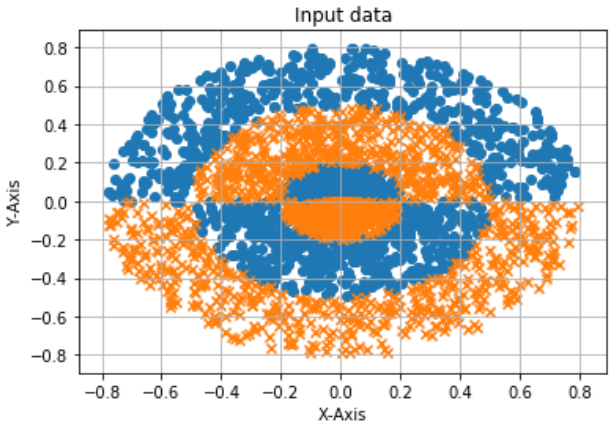
$$\langle f, g \rangle = \langle \sum_{i=1}^{N} a_i k(\mathbf{x}_i, .), \sum_{j=1}^{N} b_j k(\mathbf{x}_j, .) \rangle$$

$$\langle f, g \rangle = \sum_{i=1}^{N} \sum_{j=1}^{N} a_i b_j \langle k(\mathbf{x}_i, .) k(\mathbf{x}_j, .) \rangle \tag{13}$$

Comparing (12) and (13), we get $\langle k(\mathbf{x}_i, .), k(\mathbf{x}_j, .) \rangle = k(\mathbf{x}_i, \mathbf{x}_j)$. ■

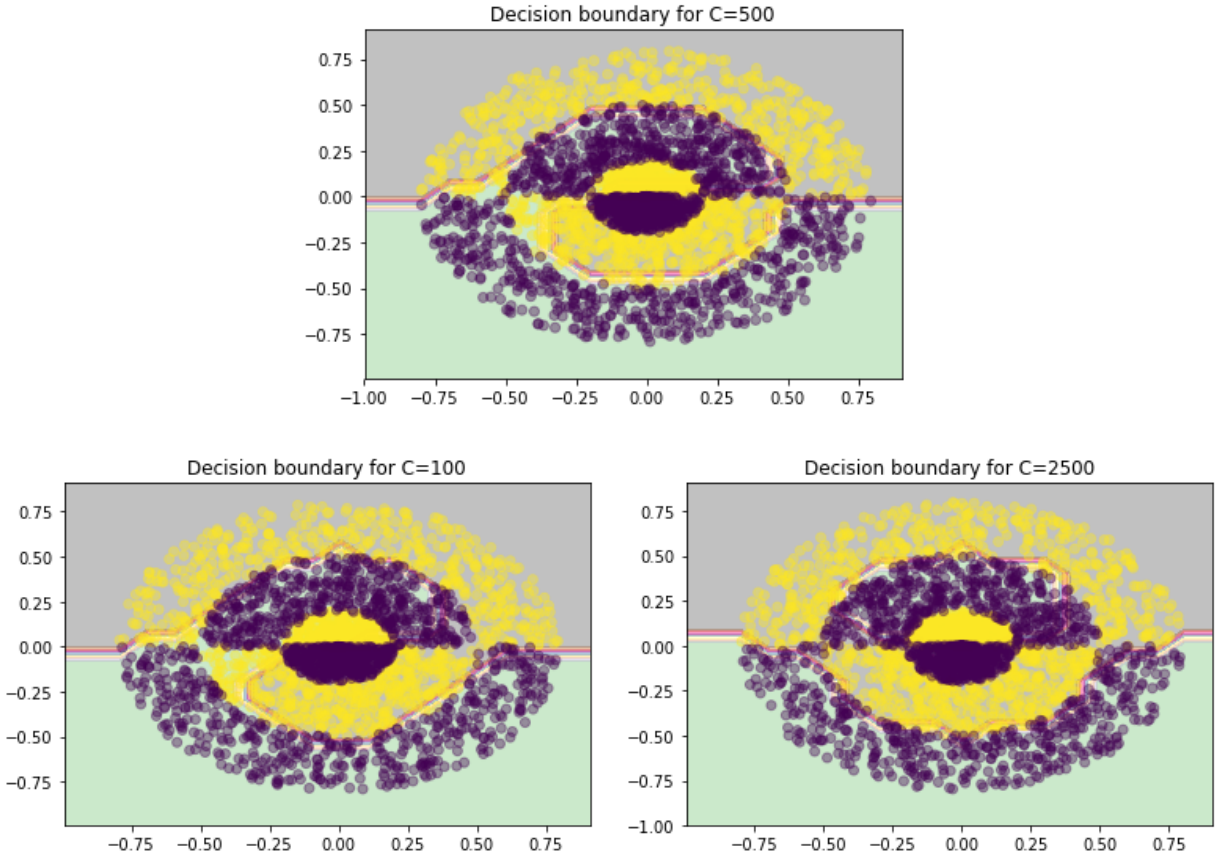
Problem 6.25**.**

Solution. (a) Generation of data set of three concentric circles with the radii as mentioned in the question. The data generated is as below.



(b) The support machine was trained with $C = 500$ and the decision boundary obtained is as given below.

(c) The network was tested and an accuracy of 68% was obtained. We could argue that the value of C might play a role in the accuracy of the SVM.



(d) The network was trained with $C = 100$ and $C = 2500$. The decision boundaries obtained are as given below.

It is observed that for the case of $C = 100$, the network accuracy was 62% and for $C = 2500$, the network accuracy was 64%.

■

Problem 2.

Solution. Given the kernel $K(\mathbf{x}, .) = \tanh(\beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x} + \beta_1)$ for $\mathbf{x} \in \mathbb{R}^{\mathrm{d}}$. Mercer's theorem is satisfied only for some choices of β_0 and β_1.
Let us first eliminate the conditions where it fails to be a Mercer kernel.

- Consider $\beta_0 < 0$ and $\beta_1 < 0$.
  Since $\mathbf{x}^{\mathrm{T}}\mathbf{x}$ is an inner product, it is always positive. Therefore, for the above case, $\beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x} + \beta_1 < 0$
  $$K(\mathbf{x}, .) = \tanh(\beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x} + \beta_1) < 0.$$
  But we know that $K(\mathbf{x}, .)$ cannot take negative values. Therefore for the condition $\beta_0 < 0$ and $\beta_1 < 0$, the kernel is not valid.

- Consider $\beta_0 < 0$ and $\beta_1 > 0$.
  For the kernel to be valid, $\beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x} + \beta_1 > 0$. It is given that $\beta_0 < 0 \implies \beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x} < 0$.

Therefore, for the kernel to be valid, the condition to be satisfied is $\beta_1 > -\beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x}$. However, we cannot comment on whether it satisfies Mercer's theorem or not.

We know that Mercer's theorem for polynomial type $(\mathbf{x}^{\mathrm{T}}\mathbf{x}+1)^p$ always satisfy Mercer's theorem. Let us consider the Maclaurin series for tanh function.

$$\tanh(x) = x - \frac{x^3}{3} + \frac{2x^5}{15} - \frac{17x^7}{315} + \cdots$$

Using the above, we get

$$\tanh(\beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x}+\beta_1) = (\beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x}+\beta_1) - \frac{(\beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x}+\beta_1)^3}{3} + \frac{2(\beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x}+\beta_1)^5}{15} - \cdots$$

Assuming $\beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x}+\beta_1$ is a small value, we take the 1^{st} order approximation of the function to get

$$\tanh(\beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x}+\beta_1) \approx (\beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x}+\beta_1)$$

Comparing this with the polynomial kernel, we see that for values $\beta_0 = 1,\ \beta_1 = 1$ and $p = 1$, the kernel satisfies the Mercer's theorem.

$$\tanh(\mathbf{x}^{\mathrm{T}}\mathbf{x}+1) \approx (\mathbf{x}^{\mathrm{T}}\mathbf{x}+1)$$

Using the above idea, for positive values of β_0, we can define a new variable such that $\tilde{\mathbf{x}} = \sqrt{\beta_0}\mathbf{x}$. Taking the inner product, we see that

$$\tilde{\mathbf{x}}^{\mathrm{T}}\tilde{\mathbf{x}} = (\sqrt{\beta_0}\mathbf{x})^{\mathrm{T}}(\sqrt{\beta_0}\mathbf{x}) = \sqrt{\beta_0}\sqrt{\beta_0}\mathbf{x}^{\mathrm{T}}\mathbf{x} = \beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x}$$

Therefore, we can approximate any positive β_0 using the above method. To see how this works, let us consider an example of $\mathbf{x}$ to be a two element vector.

$$\beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x} = \beta_0 [x_1\ x_2] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \beta_0 (x_1^2 + x_2^2)$$

$$\tilde{\mathbf{x}}^{\mathrm{T}}\tilde{\mathbf{x}} = \begin{bmatrix} \sqrt{\beta_0}x_1 & \sqrt{\beta_0}x_2 \end{bmatrix} \begin{bmatrix} \sqrt{\beta_0}x_1 \\ \sqrt{\beta_0}x_2 \end{bmatrix} = \beta_0 x_1^2 + \beta_0 x_2^2 = \beta_0 (x_1^2 + x_2^2)$$

Therefore, for any $\beta_0 > 0$ and $\beta_1 = 1$, we can approximate

$$\tanh(\beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x}+1) \approx (\tilde{\mathbf{x}}^{\mathrm{T}}\tilde{\mathbf{x}}+1)$$

The above solution relies on the assumption that $\beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x}+\beta_1$ is small. Therefore, for $\beta_0 > 0$ and $\beta_1 < 0$, it would be a better approximation of the kernel as compared to the case of $\beta_0 > 0$ and $\beta_1 > 0$. The observations have been summarized in the table below.

These observations are in line with the theoretic proofs obtained in the paper "A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods." by Lin, H.T. and Lin, C.J. Their results are as given below.

Therefore, we see that the tanh(.) kernel satisfies the Mercer's theorem better when $\beta_0 > 0$ and $\beta_1 < 0$. ■

β_0	β_1	**Observations**
+	-	A good approximation of the Mercer kernel as $\beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x} + \beta_1$ is small
+	+	Not as good approximation as $\beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x} + \beta_1$ is larger than above
-	+	Valid kernel only when $\beta_1 > -\beta_0 \mathbf{x}^{\mathrm{T}}\mathbf{x}$, otherwise invalid
-	-	Not a valid kernel

β_0	β_1	**Results**
+	-	Kernel is conditionally positive semidefinite for small β_1, and is similar to RBF for small β_0
+	+	In general not as good as the $(+,-)$ case
-	+	Objective value of a function becomes $-\infty$ after β_1 is large
-	-	Easily the objective value of the function becomes $-\infty$

Problem 3.

Solution. Given problem:

$$L_\epsilon(d, y) = \begin{cases} |d-y| - \epsilon, & |d-y| \geq \epsilon \\ 0, & \text{else} \end{cases}$$

$$\begin{aligned} \min \quad & \frac{1}{N}\sum_{i=0}^{N-1} L_\epsilon(d_i, y_i) \\ \text{subject to} \quad & \left.\begin{array}{ll} \|\mathbf{w}\|^2 & \leq c_0 \\ d_i - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i) & \leq \epsilon + \zeta_i \\ \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i) - d_i & \leq \epsilon + \zeta_i' \\ \zeta_i & \geq 0 \\ \zeta_i' & \geq 0 \end{array}\right\} \quad \forall i = 0, 1, 2, \ldots, N-1. \end{aligned}$$

Substituting $y_i = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i)$ and writing the above in the standard form to write the Lagrange, we get

$$\begin{aligned} \min \quad & \frac{1}{N}\sum_{i=0}^{N-1} L_\epsilon(d_i, y_i) \\ \text{subject to} \quad & \left.\begin{array}{ll} c_0 - \|\mathbf{w}\|^2 & \geq 0 \\ \epsilon + \zeta_i - d_i + \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i) & \geq 0 \\ \epsilon + \zeta_i' - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i) + d_i & \geq 0 \\ \zeta_i & \geq 0 \\ \zeta_i' & \geq 0 \end{array}\right\} \quad \forall i = 0, 1, 2, \ldots, N-1. \end{aligned}$$

The primal can be set up using the Lagrange as

$$L = \frac{1}{N}\sum_{i=0}^{N-1}\left(\zeta_i + \zeta_i^{'}\right) - \alpha\left(c_0 - \|\mathbf{w}\|^2\right) - \sum_{i=0}^{N-1}\beta_i\left(\epsilon + \zeta_i - d_i + \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i)\right) - \sum_{i=0}^{N-1}\beta_i^{'}\left(\epsilon + \zeta_i^{'} - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i) + d_i\right) - \sum_{i=0}^{N-1}\gamma_i\zeta_i - \sum_{i=0}^{N-1}\gamma_i^{'}\zeta_i^{'} \tag{14}$$

where $\alpha, \beta_i, \beta_i^{'}, \gamma_i$, and $\gamma_i^{'}$ are the Lagrangian multipliers. Differentiating the Lagrange and equating to zero, we get

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \implies 2\alpha\mathbf{w} - \sum_{i=0}^{N-1}\beta_i\phi(\mathbf{x}_i) + \sum_{i=0}^{N-1}\beta_i^{'}\phi(\mathbf{x}_i) = 0 \implies \mathbf{w} = \frac{1}{2\alpha}\sum_{i=0}^{N-1}\left(\beta_i - \beta_i^{'}\right)\phi(\mathbf{x}_i)$$

$$\frac{\partial L}{\partial \zeta_i} = 0 \implies \frac{1}{N} - \beta_i - \gamma_i = 0 \implies \beta_i + \gamma_i = \frac{1}{N}$$

$$\frac{\partial L^{'}}{\partial \zeta_i} = 0 \implies \frac{1}{N} - \beta_i^{'} - \gamma_i^{'} = 0 \implies \beta_i^{'} + \gamma_i^{'} = \frac{1}{N}$$

Grouping similar terms in (14), we get

$$L = \sum_{i=0}^{N-1}\left(\frac{1}{N} - \beta_i - \gamma_i\right)\zeta_i + \sum_{i=0}^{N-1}\left(\frac{1}{N} - \beta_i^{'} - \gamma_i^{'}\right)\zeta_i^{'} - \alpha c_0 + \alpha\|\mathbf{w}\|^2 - \sum_{i=0}^{N-1}\left(\beta_i - \beta_i^{'}\right)\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i) - \sum_{i=0}^{N-1}(\beta_i + \beta_i^{'})\epsilon + \sum_{i=0}^{N-1}(\beta_i + \beta_i^{'})d_i \tag{15}$$

Substituting the above values, we get

$$
\begin{aligned}
L = & \sum_{i=0}^{N-1} \left(\beta_i + \gamma_i - \beta_i - \gamma_i\right) \zeta_i + \sum_{i=0}^{N-1} \left(\beta_i^{'} + \gamma_i^{'} - \beta_i^{'} - \gamma_i^{'}\right) \zeta_i^{'} - \alpha c_0 \\
& + \frac{\alpha}{4\alpha^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \left(\beta_i - \beta_i^{'}\right) \left(\beta_j - \beta_j^{'}\right) \phi(\mathbf{x}_i)^{\mathrm{T}} \phi(\mathbf{x}_j) \\
& - \frac{1}{2\alpha} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \left(\beta_i - \beta_i^{'}\right) \left(\beta_j - \beta_j^{'}\right) \phi(\mathbf{x}_i)^{\mathrm{T}} \phi(\mathbf{x}_j) \\
& - \sum_{i=0}^{N-1} (\beta_i + \beta_i^{'}) \epsilon + \sum_{i=0}^{N-1} (\beta_i + \beta_i^{'}) d_i \\
= & \sum_{i=0}^{N-1} \cancelto{0}{\left(\beta_i + \gamma_i - \beta_i - \gamma_i\right)} \zeta_i + \sum_{i=0}^{N-1} \cancelto{0}{\left(\beta_i^{'} + \gamma_i^{'} - \beta_i^{'} - \gamma_i^{'}\right)} \zeta_i^{'} - \alpha c_0 \\
& + \left(\frac{1}{4\alpha} - \frac{1}{2\alpha}\right) \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \left(\beta_i - \beta_i^{'}\right) \left(\beta_j - \beta_j^{'}\right) \phi(\mathbf{x}_i)^{\mathrm{T}} \phi(\mathbf{x}_j) \\
& - \sum_{i=0}^{N-1} (\beta_i + \beta_i^{'}) \epsilon + \sum_{i=0}^{N-1} (\beta_i + \beta_i^{'}) d_i \\
= & -\alpha c_0 - \frac{1}{4\alpha} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \left(\beta_i - \beta_i^{'}\right) \left(\beta_j - \beta_j^{'}\right) \phi(\mathbf{x}_i)^{\mathrm{T}} \phi(\mathbf{x}_j) \\
& - \sum_{i=0}^{N-1} (\beta_i + \beta_i^{'}) \epsilon + \sum_{i=0}^{N-1} (\beta_i + \beta_i^{'}) d_i
\end{aligned}
$$

From the above, the dual can be written as

$$
\begin{aligned}
\max \quad & -\alpha c_0 - \frac{1}{4\alpha} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \left(\beta_i - \beta_i^{'}\right) \left(\beta_j - \beta_j^{'}\right) \phi(\mathbf{x}_i)^{\mathrm{T}} \phi(\mathbf{x}_j) \\
& - \sum_{i=0}^{N-1} (\beta_i + \beta_i^{'}) \epsilon + \sum_{i=0}^{N-1} (\beta_i + \beta_i^{'}) d_i \\
\text{subject to} \quad & \left.\begin{array}{l} \beta_i + \gamma_i = \frac{1}{N} \\ \beta_i^{'} + \gamma_i^{'} = \frac{1}{N} \\ \alpha \geq 0 \\ \beta_i \geq 0 \\ \beta_i^{'} \geq 0 \\ \gamma_i \geq 0 \\ \gamma_i^{'} \geq 0 \end{array}\right\} \quad \forall i = 0, 1, 2, \ldots, N-1
\end{aligned}
$$

We observe that the Lagrange multiplier γ_i and $\gamma_i^{'}$ appear only in the constraint $\beta_i + \gamma_i = \frac{1}{N}$ and $\beta_i^{'} + \gamma_i^{'} = \frac{1}{N}$ respectively. For $\gamma_i \geq 0$ to be true, $\frac{1}{N} - \beta_i \geq 0 \implies \frac{1}{N} \geq \beta_i$ and for $\gamma_i^{'} \geq 0$ to be true $\frac{1}{N} - \beta_i^{'} \geq 0 \implies \frac{1}{N} \geq \beta_i^{'}$. Combining this with $\beta_i \geq 0$ and $\beta_i^{'} \geq 0$ constraint, we get $0 \leq \beta_i \leq \frac{1}{N}$ and $0 \leq \beta_i^{'} \leq \frac{1}{N}$ respectively. The dual problem can now be written as

$$
\begin{aligned}
\max \quad & -\alpha c_0 - \frac{1}{4\alpha} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \left(\beta_i - \beta_i^{'}\right) \left(\beta_j - \beta_j^{'}\right) \phi(\mathbf{x}_i)^{\mathrm{T}} \phi(\mathbf{x}_j) \\
& - \sum_{i=0}^{N-1} (\beta_i + \beta_i^{'}) \epsilon + \sum_{i=0}^{N-1} (\beta_i + \beta_i^{'}) d_i \\
\text{subject to} \quad & \left.\begin{array}{l} \alpha \geq 0 \\ 0 \leq \beta_i \leq \frac{1}{N} \\ 0 \leq \beta_i^{'} \leq \frac{1}{N} \end{array}\right\} \quad \forall i = 0, 1, 2, \ldots, N-1.
\end{aligned}
$$

■