

# Homework 1

## Question 1

Recall that  $N$  is the size of the data set and  $d$  is the dimensionality of the input space. The primal formulation of the linear soft-margin support vector machine problem, without going through the Lagrangian dual problem, is

- ☐ a quadratic programming problem with  $N + d + 1$  variables
- ☐ a quadratic programming problem with  $d + 1$  variables
- ☐ none of the other choices
- ☐ a quadratic programming problem with  $N$  variables
- ☐ a quadratic programming problem with  $2N$  variables

## Question 2

Consider the following training data set:

$$\begin{aligned} \mathbf{x}_1 &= (1, 0), y_1 = -1 & \mathbf{x}_2 &= (0, 1), y_2 = -1 & \mathbf{x}_3 &= (0, -1), y_3 = -1 \\ \mathbf{x}_4 &= (-1, 0), y_4 = +1 & \mathbf{x}_5 &= (0, 2), y_5 = +1 & \mathbf{x}_6 &= (0, -2), y_6 = +1 \\ \mathbf{x}_7 &= (-2, 0), y_7 = +1 \end{aligned}$$

Use following nonlinear transformation of the input vector  $\mathbf{x} = (x_1, x_2)$  to the transformed vector  $\mathbf{z} = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}))$ :

$$\phi_1(\mathbf{x}) = x_2^2 - 2x_1 + 3 \quad \phi_2(\mathbf{x}) = x_1^2 - 2x_2 - 3$$

What is the equation of the optimal separating "hyperplane" in the  $\mathcal{Z}$  space?

- ☐  $z_1 + z_2 = 4.5$
- ☐  $z_1 - z_2 = 4.5$
- ☐  $z_1 = 4.5$
- ☐  $z_2 = 4.5$
- ☐ none of the other choices

### Question 3

Consider the same training data set as Question 2, but instead of explicitly transforming the input space  $\mathcal{X}$  to  $\mathcal{Z}$ , apply the hard-margin support vector machine algorithm with the kernel function

$$K(\mathbf{x}, \mathbf{x}') = (1 + \mathbf{x}^T \mathbf{x}')^2,$$

which corresponds to a second-order polynomial transformation. Set up the optimization problem using  $(\alpha_1, \dots, \alpha_7)$  and numerically solve for them (you can use any package you want). Which of the following is true about the optimal  $\alpha$ ?

- ☐  $\min_{1 \leq n \leq 7} \alpha_n = \alpha_7$
- ☐  $\sum_{n=1}^7 \alpha_n = 5$
- ☐ there are 6 nonzero  $\alpha_n$
- ☐ none of the other choices
- ☐  $\max_{1 \leq n \leq 7} \alpha_n = \alpha_7$

### Question 4

Following Question 3, what is the corresponding nonlinear curve in the  $\mathcal{X}$  space?

- ☐  $\frac{1}{9} (8x_2^2 - 16x_2 + 6x_1^2 - 15) = 0$
- ☐  $\frac{1}{9} (8x_1^2 - 16x_1 + 6x_2^2 - 15) = 0$
- ☐  $\frac{1}{9} (8x_2^2 - 16x_2 + 6x_1^2 + 15) = 0$
- ☐ none of the other choices
- ☐  $\frac{1}{9} (8x_1^2 - 16x_1 + 6x_2^2 + 15) = 0$

### Question 5

Compare the two nonlinear curves found in Questions 2 and 4, which of the following is true?

- ☐ The curves should be different in the  $\mathcal{X}$  space, because they are learned with respect to different  $\mathcal{Z}$  spaces
- ☐ The curves should be the same in the  $\mathcal{X}$  space, because they are learned with respect to the same  $\mathcal{Z}$  space

☐ none of the other choices

☐

The curves should be the same in the  $\mathcal{X}$  space, because they are learned from the same raw data  $\{(\mathbf{x}_n, y_n)\}$

☐

The curves should be different in the  $\mathcal{X}$  space, because they are learned from different raw data  $\{(\mathbf{x}_n, y_n)\}$

## Question 6

Recall that for support vector machines,  $d_{vc}$  is upper bounded by  $\frac{R^2}{\rho^2}$ , where  $\rho$  is the margin and  $R$  is the radius of the minimum hypersphere that  $\mathcal{X}$  resides in. In general,  $R$  should come from our knowledge on the learning problem, but we can estimate it by looking at the minimum hypersphere that the training examples resides in. In particular, we want to seek for the optimal  $R$  that solves

$$(P) \quad \min_{R \in \mathbb{R}, \mathbf{c} \in \mathbb{R}^d} R^2 \quad \text{subject to } \|\mathbf{x}_n - \mathbf{c}\|^2 \leq R^2 \text{ for } n = 1, 2, \dots, N.$$

Let  $\lambda_n$  be the Lagrange multipliers for the  $n$ -th constraint above. Following the derivation of the dual support vector machine in class, write down  $(P)$  as an equivalent optimization problem

$$\min_{R \in \mathbb{R}, \mathbf{c} \in \mathbb{R}^d} \max_{\lambda_n \geq 0} L(R, \mathbf{c}, \lambda).$$

What is  $L(R, \mathbf{c}, \lambda)$ ?

☐  $R^2 - \sum_{n=1}^N \lambda_n (\|\mathbf{x}_n - \mathbf{c}\|^2 - R^2)$

☐ none of the other choices

☐  $R^2 + \sum_{n=1}^N \lambda_n (\|\mathbf{x}_n - \mathbf{c}\|^2 - R^2)$

☐  $R^2 + \sum_{n=1}^N \lambda_n (\|\mathbf{x}_n - \mathbf{c}\|^2 + R^2)$

☐  $R^2 - \sum_{n=1}^N \lambda_n (\|\mathbf{x}_n - \mathbf{c}\|^2 + R^2)$

## Question 7

Using (assuming) strong duality, the solution to  $(P)$  in Question 6 would be the same as the Lagrange dual problem

$$(D) \quad \max_{\lambda_n \geq 0} \min_{R \in \mathbb{R}, \mathbf{c} \in \mathbb{R}^d} L(R, \mathbf{c}, \lambda).$$

Which of the following can be derived from the KKT conditions of  $(P)$  and  $(D)$  at the optimal

$(R, \mathbf{c}, \lambda)$ ?

- ☐ if  $\lambda_n = 0$ , then  $\|\mathbf{x}_n - \mathbf{c}\|^2 - R^2 = 0$
- ☐ if  $\|\mathbf{x}_n - \mathbf{c}\|^2 - R^2 < 0$ , then  $\lambda_n < 0$
- ☐ if  $\sum_{n=1}^N \lambda_n = 0$ , then  $\mathbf{c} = \left( \sum_{n=1}^N \lambda_n \mathbf{x}_n \right) / \left( \sum_{n=1}^N \lambda_n \right)$
- ☐ none of the other choices
- ☐ if  $R \neq 0$ , then  $\sum_{n=1}^N \lambda_n = 1$

## Question 8

Continue from Question 7 and assume that all the  $\mathbf{x}_n$  are different, which implies that the optimal  $R > 0$ . Using the KKT conditions to simplify the Lagrange dual problem, and obtain a dual problem that involves only  $\lambda_n$ . One form of the dual problem should look like

$$(D') \quad \max_{\lambda_n \geq 0} \quad \text{Objective}(\lambda) \quad \text{subject to} \quad \sum_{n=1}^N \lambda_n = \text{constant}$$

Which of the following is  $\text{Objective}(\lambda)$ ?

- ☐  $\sum_{n=1}^N \lambda_n (\|\mathbf{x}_n - \sum_{m=1}^N \lambda_m \mathbf{x}_m\|^2) + 2(\sum_{n=1}^N \lambda_n \mathbf{x}_n)^2$
- ☐  $\sum_{n=1}^N \lambda_n (\|\mathbf{x}_n - \sum_{m=1}^N \lambda_m \mathbf{x}_m\|^2)$
- ☐ none of the other choices
- ☐  $\sum_{n=1}^N \lambda_n (\|\mathbf{x}_n + \sum_{m=1}^N \lambda_m \mathbf{x}_m\|^2)$
- ☐  $\sum_{n=1}^N \lambda_n (\|\mathbf{x}_n + \sum_{m=1}^N \lambda_m \mathbf{x}_m\|^2) + 2(\sum_{n=1}^N \lambda_n \mathbf{x}_n)^2$

## Question 9

Continue from Question 8 and consider using  $\mathbf{z}_n = \phi(\mathbf{x}_n)$  instead of  $\mathbf{x}_n$  while assuming that all the  $\mathbf{z}_n$  are different. Then, write down the optimization problem that uses  $K(\mathbf{x}_n, \mathbf{x}_m)$  to replace  $\mathbf{z}_n^T \mathbf{z}_m$ ---that is, the kernel trick. Which of the following is  $\text{Objective}(\lambda)$  of  $(D')$  after applying the kernel trick?

- ☐  $\sum_{n=1}^N \lambda_n K(\mathbf{x}_n, \mathbf{x}_n) - 3 \sum_{n=1}^N \sum_{m=1}^N \lambda_n \lambda_m K(\mathbf{x}_n, \mathbf{x}_m)$
- ☐ none of the other choices
- ☐  $\sum_{n=1}^N \lambda_n K(\mathbf{x}_n, \mathbf{x}_n) + 3 \sum_{n=1}^N \sum_{m=1}^N \lambda_n \lambda_m K(\mathbf{x}_n, \mathbf{x}_m)$
- ☐  $\sum_{n=1}^N \lambda_n K(\mathbf{x}_n, \mathbf{x}_n) - 1 \sum_{n=1}^N \sum_{m=1}^N \lambda_n \lambda_m K(\mathbf{x}_n, \mathbf{x}_m)$

☐  $\sum_{n=1}^N \lambda_n K(\mathbf{x}_n, \mathbf{x}_n) + 1 \sum_{n=1}^N \sum_{m=1}^N \lambda_n \lambda_m K(\mathbf{x}_n, \mathbf{x}_m)$

## Question 10

Continue from Question 9 and solve the  $(D')$  that involves the kernel  $K$ , which of the following formula evaluates the optimal  $R$ ?



Pick some  $i$  with  $\lambda_i > 0$ , and

$$R = \sqrt{K(\mathbf{x}_i, \mathbf{x}_i) - 2 \sum_{m=1}^N \lambda_m K(\mathbf{x}_i, \mathbf{x}_m) + \sum_{n=1}^N \sum_{m=1}^N \lambda_n \lambda_m K(\mathbf{x}_n, \mathbf{x}_m)}$$



Pick some  $i$  with  $\lambda_i = 0$ , and

$$R = \sqrt{K(\mathbf{x}_i, \mathbf{x}_i) - 2 \sum_{m=1}^N \lambda_m K(\mathbf{x}_i, \mathbf{x}_m) + \sum_{n=1}^N \sum_{m=1}^N \lambda_n \lambda_m K(\mathbf{x}_n, \mathbf{x}_m)}$$



Pick some  $i$  with  $\lambda_i = 0$ , and

$$R = \sqrt{K(\mathbf{x}_i, \mathbf{x}_i) + 2 \sum_{m=1}^N \lambda_m K(\mathbf{x}_i, \mathbf{x}_m) + \sum_{n=1}^N \sum_{m=1}^N \lambda_n \lambda_m K(\mathbf{x}_n, \mathbf{x}_m)}$$



Pick some  $i$  with  $\lambda_i > 0$ , and

$$R = \sqrt{K(\mathbf{x}_i, \mathbf{x}_i) + 2 \sum_{m=1}^N \lambda_m K(\mathbf{x}_i, \mathbf{x}_m) + \sum_{n=1}^N \sum_{m=1}^N \lambda_n \lambda_m K(\mathbf{x}_n, \mathbf{x}_m)}$$



none of the other choices

## Question 11

In the class, we taught the soft-margin support vector machine as follows.

$$\begin{aligned} (P_1) \quad & \min_{\mathbf{w}, b, \xi} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{n=1}^N \xi_n \\ & \text{subject to} \quad y_n (\mathbf{w}^T \mathbf{x}_n + b) \geq 1 - \xi_n, \\ & \quad \quad \quad \xi_n \geq 0. \end{aligned}$$

The support vector machine (called  $\ell_1$  loss) penalizes the margin violation linearly. Another popular formulation (called  $\ell_2$  loss) penalizes the margin violation quadratically. In this problem, we show one simple approach for deriving the dual of such a formulation. The formulation is as follows.

$$\begin{aligned} (P'_2) \quad & \min_{\mathbf{w}, b, \xi} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{n=1}^N \xi_n^2 \\ & \text{subject to} \quad y_n (\mathbf{w}^T \mathbf{x}_n + b) \geq 1 - \xi_n, \\ & \quad \quad \quad \xi_n \geq 0. \end{aligned}$$

It is not hard to see that the constraints  $\xi_n \geq 0$  are not necessary for the new formulation. In other words, the formulation  $(P'_2)$  is equivalent to the following optimization problem.

$$(P_2) \quad \min_{\mathbf{w}, b, \xi} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{n=1}^N \xi_n^2$$

$$\text{subject to } y_n (\mathbf{w}^T \mathbf{x}_n + b) \geq 1 - \xi_n.$$

Problem  $(P_2)$  is equivalent to a linear hard-margin support vector machine (primal problem) that takes examples  $(\tilde{\mathbf{x}}_n, y_n)$  instead of  $(\mathbf{x}_n, y_n)$ . That is, the hard-margin dual problem that involves  $\tilde{\mathbf{x}}_n$  is simply the dual problem of  $(P_2)$ . Use  $[[\cdot]]$  to denote the boolean function which evaluates to 1 if and only if the inner condition is true. Which of the following is  $\tilde{\mathbf{x}}_n$ ? (Hint: let

$\tilde{\mathbf{w}} = (\mathbf{w}, \text{constant} \cdot \xi)$ )

- ☐  $\tilde{\mathbf{x}}_n = (\mathbf{x}_n, v_1, v_2, \dots, v_N)$ , where  $v_i = \frac{1}{\sqrt{2C}} [[i = n]]$
- ☐  $\tilde{\mathbf{x}}_n = (\mathbf{x}_n, v, v, \dots, v)$ , where there are  $N$  components of  $v = \frac{1}{\sqrt{2C}}$
- ☐  $\tilde{\mathbf{x}}_n = (\mathbf{x}_n, v_1, v_2, \dots, v_N)$ , where  $v_i = \frac{1}{\sqrt{C}} [[i = n]]$
- ☐  $\tilde{\mathbf{x}}_n = (\mathbf{x}_n, v, v, \dots, v)$ , where there are  $N$  components of  $v = \frac{1}{\sqrt{C}}$
- ☐ none of the other choices

## Question 12

Let  $K_1(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}_1(\mathbf{x})^T \boldsymbol{\phi}_1(\mathbf{x}')$  and  $K_2(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}_2(\mathbf{x})^T \boldsymbol{\phi}_2(\mathbf{x}')$  be two valid kernels. Which of the following is always a valid kernel, assuming that  $K_2(\mathbf{x}, \mathbf{x}') \neq 0$  for all  $\mathbf{x}$  and  $\mathbf{x}'$ ?

- ☐  $K(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') - K_2(\mathbf{x}, \mathbf{x}')$
- ☐  $K(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') + K_2(\mathbf{x}, \mathbf{x}')$
- ☐ none of the other choices
- ☐  $K(\mathbf{x}, \mathbf{x}') = -K_1(\mathbf{x}, \mathbf{x}') - K_2(\mathbf{x}, \mathbf{x}')$
- ☐  $K(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}')/K_2(\mathbf{x}, \mathbf{x}')$

## Question 13

Let  $K_1(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}_1(\mathbf{x})^T \boldsymbol{\phi}_1(\mathbf{x}')$  be a valid kernel. Which of the following is always a valid kernel?

- ☐  $K(\mathbf{x}, \mathbf{x}') = 1126/K_1(\mathbf{x}, \mathbf{x}')$
- ☐ none of the other choices

- ☐  $K(\mathbf{x}, \mathbf{x}') = (1 - K_1(\mathbf{x}, \mathbf{x}'))^2$
- ☐  $K(\mathbf{x}, \mathbf{x}') = (1 - K_1(\mathbf{x}, \mathbf{x}'))^{-1}$ , assuming that  $0 < K_1(\mathbf{x}, \mathbf{x}') < 1$
- ☐  $K(\mathbf{x}, \mathbf{x}') = \exp(-K_1(\mathbf{x}, \mathbf{x}'))$

## Question 14

For a given valid kernel  $K$ , consider a new kernel  $\tilde{K}(\mathbf{x}, \mathbf{x}') = pK(\mathbf{x}, \mathbf{x}') + q$  for some  $p > 0$  and  $q > 0$ . Which of the following statement is true?

- ☐ For the dual of soft-margin support vector machine, using  $\tilde{K}$  along with a new  $\tilde{C} = \frac{C}{p} + q$  instead of  $K$  with the original  $C$  leads to an equivalent  $g_{SVM}$  classifier.
- ☐ For the dual of soft-margin support vector machine, using  $\tilde{K}$  along with a new  $\tilde{C} = \frac{C}{p}$  instead of  $K$  with the original  $C$  leads to an equivalent  $g_{SVM}$  classifier.
- ☐ none of the other choices
- ☐ For the dual of soft-margin support vector machine, using  $\tilde{K}$  along with a new  $\tilde{C} = pC$  instead of  $K$  with the original  $C$  leads to an equivalent  $g_{SVM}$  classifier.
- ☐ For the dual of soft-margin support vector machine, using  $\tilde{K}$  along with a new  $\tilde{C} = pC + q$  instead of  $K$  with the original  $C$  leads to an equivalent  $g_{SVM}$  classifier.

## Question 15

For Questions 15 to 20, we are going to experiment with a real-world data set. Download the processed US Postal Service Zip Code data set with extracted features of intensity and symmetry for training and testing:

<http://www.amlbook.com/data/zip/features.train>

<http://www.amlbook.com/data/zip/features.test>

The format of each row is

digit intensity symmetry

We will consider binary classification problems of the form "one of the digits" (as the positive class) versus "other digits" (as the negative class).

The training set contains thousands of examples, and some quadratic programming packages

cannot handle this size. We recommend that you consider the LIBSVM package

<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

Regardless of the package that you choose to use, please read the manual of the package carefully to make sure that you are indeed solving the soft-margin support vector machine taught in class like the dual formulation below:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N \alpha_n \alpha_m y_n y_m K(\mathbf{x}_n, \mathbf{x}_m) - \sum_{n=1}^N \alpha_n \\ \text{s. t.} \quad & \sum_{n=1}^N y_n \alpha_n = 0 \\ & 0 \leq \alpha_n \leq C \quad n = 1, \dots, N \end{aligned}$$

In the following questions, please use the 0/1 error for evaluating  $E_{\text{in}}$ ,  $E_{\text{val}}$  and  $E_{\text{out}}$  (through the test set). Some practical remarks include

(i) Please tell your chosen package to **not** automatically scale the data for you, lest you should change the effective kernel and get different results.

(ii) It is your responsibility to check whether your chosen package solves the designated formulation with enough numerical precision. Please read the manual of your chosen package for software parameters whose values affect the outcome---any ML practitioner needs to deal with this kind of added uncertainty.

Consider the linear soft-margin SVM. That is, either solve the primal formulation of soft-margin SVM with the given  $\mathbf{x}_n$ , or take the linear kernel  $K(\mathbf{x}_n, \mathbf{x}_m) = \mathbf{x}_n^T \mathbf{x}_m$  in the dual formulation. With  $C = 0.01$ , and the binary classification problem of "0" versus "not 0", which of the following numbers is closest to  $\|\mathbf{w}\|$  after solving the linear soft-margin SVM?

- ☐ 1.4
- ☐ 1.8
- ☐ 0.2
- ☐ 1.0
- ☐ 0.6

## Question 16

Consider the polynomial kernel  $K(\mathbf{x}_n, \mathbf{x}_m) = (1 + \mathbf{x}_n^T \mathbf{x}_m)^Q$ , where  $Q$  is the degree of the polynomial. With  $C = 0.01$ ,  $Q = 2$ , which of the following soft-margin SVM classifiers reaches the lowest  $E_{\text{in}}$ ?

- ☐ "0" versus "not 0"



- ☐ "2" versus "not 2"
- ☐ "4" versus "not 4"
- ☐ "6" versus "not 6"
- ☐ "8" versus "not 8"

## Question 17

Following Question 16, which of the following numbers is closest to the maximum  $\sum_{n=1}^N \alpha_n$  within those five soft-margin SVM classifiers?

- ☐ 15.0
- ☐ 25.0
- ☐ 5.0
- ☐ 10.0
- ☐ 20.0

## Question 18

Consider the Gaussian kernel  $K(\mathbf{x}_n, \mathbf{x}_m) = \exp(-\gamma \|\mathbf{x}_n - \mathbf{x}_m\|^2)$ . With  $\gamma = 100$ , and the binary classification problem of "0" versus "not 0". Consider values of  $C$  within  $\{0.001, 0.01, 0.1, 1, 10\}$ . Which of the following properties of the soft-margin SVM classifier strictly decreases with those five  $C$ ?

- ☐  $\sum_{n=1}^N \sum_{m=1}^N K(\mathbf{x}_n, \mathbf{x}_m)$
- ☐ number of support vectors
- ☐ none of the other choices
- ☐  $E_{\text{out}}$
- ☐ the distance of any free support vector to the hyperplane in the (infinite-dimensional)  $\mathcal{Z}$  space

## Question 19

Following Question 18, when fixing  $C = 0.1$ , which of the following values of  $\gamma$  results in the lowest  $E_{\text{out}}$ ?

- ☐ 1000
- ☐ 1
- ☐ 10
- ☐ 100
- ☐ 10000

## Question 20

Following Question 18 and consider a validation procedure that randomly samples 1000 examples from the training set for validation and leaves the other examples for training  $g_{SVM}^-$ . Fix  $C = 0.1$  and use the validation procedure to choose the best  $\gamma$  among  $\{1, 10, 100, 1000, 10000\}$  according to  $E_{val}$ . If there is a tie of  $E_{val}$ , choose the smallest  $\gamma$ . Repeat the procedure 100 times. Which of the following values of  $\gamma$  is selected the most number of times?

- ☐ 1
- ☐ 1000
- ☐ 10000
- ☐ 100
- ☐ 10