Homework #5

RELEASE DATE: 12/03/2014

DUE DATE: 12/18/2014, BEFORE NOON

QUESTIONS ABOUT HOMEWORK MATERIALS ARE WELCOMED ON THE COURSERA FORUM.

Unless granted by the instructor in advance, you must turn in a printed/written copy of your solutions (without the source code) for all problems. For problems marked with (*), please follow the guidelines on the course website and upload your source code to designated places. You are encouraged to (but not required to) include a README to help the TAs check your source code. Any programming language/platform is allowed.

Any form of cheating, lying, or plagiarism will not be tolerated. Students can get zero scores and/or fail the class and/or be kicked out of school and/or receive other punishments for those kinds of misconducts.

Discussions on course materials and homework solutions are encouraged. But you should write the final solutions alone and understand them fully. Books, notes, and Internet resources can be consulted, but not copied from.

Since everyone needs to write the final solutions alone, there is absolutely no need to lend your homework solutions and/or source codes to your classmates at any time. In order to maximize the level of fairness in this class, lending and borrowing homework solutions are both regarded as dishonest behaviors and will be punished according to the honesty policy.

You should write your solutions in English or Chinese with the common math notations introduced in class or in the problems. We do not accept solutions written in any other languages.

There are two kinds of regular problems.

- *multiple-choice question (MCQ): There are several choices and* ***only one of them is correct****. You should choose one and only one.*
- *multiple-response question (MRQ): There are several choices and* ***none, some, or all of them are correct****. You should write down every choice that you think to be correct.*

Some problems also come with ***(+ ...)*** *that contains additional todo items.*
If there are big bonus questions (BBQ), please simply follow the problem guideline to write down your solution, if you choose to tackle them.

This homework set comes with 200 points and 20 bonus points. In general, every homework set would come with a full credit of 200 points, with some possible bonus points.

Primal versus Dual Problem

1. (MCQ) 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**] a quadratic programming problem with N variables

[✓ [**b**] a quadratic programming problem with $N + d + 1$ variables]

[**c**] a quadratic programming problem with $d + 1$ variables

[**d**] a quadratic programming problem with $2N$ variables

[**e**] none of the other choices

Transforms: Explicit versus Implicit

Consider the following training data set:

$$\mathbf{x}_1 = (1, 0), y_1 = -1 \qquad \mathbf{x}_2 = (0, 1), y_2 = -1 \qquad \mathbf{x}_3 = (0, -1), y_3 = -1$$

$$\mathbf{x}_4 = (-1, 0), y_4 = +1 \qquad \mathbf{x}_5 = (0, 2), y_5 = +1 \qquad \mathbf{x}_6 = (0, -2), y_6 = +1$$

$$\mathbf{x}_7 = (-2, 0), y_7 = +1$$

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

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

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

[a] $z_1 + z_2 = 4.5$

[b] $z_1 - z_2 = 4.5$

[✓ [c] $z_1 = 4.5$]

[d] $z_2 = 4.5$

[e] none of the other choices

3. (MRQ) Consider the same training data set, but instead of explicitly transforming the input space $\mathcal{X}$, 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 $(\alpha_1, \cdots, \alpha_7)$ and numerically solve for them (you can use any package you want). Which of the followings are true about the optimal $\boldsymbol{\alpha}$?

[a] there are 6 nonzero α_n

[✓ [b] $\sum_{n=1}^{7} \alpha_n \approx 2.8148$]

[c] $\max_{1 \leq n \leq 7} \alpha_n = \alpha_7$

[✓ [d] $\min_{1 \leq n \leq 7} \alpha_n = \alpha_7$]

[e] none of the other choices

4. (MCQ) Following Question 3, what is the corresponding nonlinear curve in the $\mathcal{X}$ space?

[a] $\frac{1}{9}(8x_1^2 - 16x_1 + 6x_2^2 + 15) = 0$

[✓ [b] $\frac{1}{9}(8x_1^2 - 16x_1 + 6x_2^2 - 15) = 0$]

[c] $\frac{1}{9}(8x_2^2 - 16x_2 + 6x_1^2 + 15) = 0$

[d] $\frac{1}{9}(8x_2^2 - 16x_2 + 6x_1^2 - 15) = 0$

[e] none of the other choices

5. (MCQ) Compare the two nonlinear curves found in Questions 2 and 4, which of the following is true?

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

[b] The curves should be the same in the $\mathcal{X}$ space, because they are learned with respect to the same $\mathcal{Z}$ space

[✓ [c]] The curves should be different in the $\mathcal{X}$ space, because they are learned with respect to different $\mathcal{Z}$ spaces

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

[e] none of the other choices

Radius of Transformed Vectors via the Kernel

Recall that for support vector machines, d_{VC} is upper bounded by $\frac{R^2}{\rho^2}$, where ρ 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

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

6. (MCQ) Let λ_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

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

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

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

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

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

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

[e] none of the other choices

7. (MRQ) Using (assuming) strong duality, the solution to (P) would be the same as the Lagrange dual problem

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

Which of the following can be derived from the KKT conditions of (P) and (D) at the optimal $(R, \mathbf{c}, \boldsymbol{\lambda})$?

[✓ [a]] if $R \ne 0$, then $\sum_{n=1}^{N} \lambda_n = 1$

[b] if $\lambda_n = 0$, then $\|\mathbf{x}_n - \mathbf{c}\|^2 - R^2 = 0$

[✓ [c]] if $\|\mathbf{x}_n - \mathbf{c}\|^2 - R^2 < 0$, then $\lambda_n = 0$

[✓ [d]] if $\sum_{n=1}^{N} \lambda_n \ne 0$, then $\mathbf{c} = \left(\sum_{n=1}^{N} \lambda_n \mathbf{x}_n\right) \Big/ \left(\sum_{n=1}^{N} \lambda_n\right)$

[e] none of the other choices

8. (MCQ) Assuming 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 λ_n. One form of the dual problem should look like

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

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

[a] $\sum_{n=1}^{N} \lambda_n (\|\mathbf{x}_n - \sum_{m=1}^{N} \lambda_m \mathbf{x}_m\|^2)$

[b] $\sum_{n=1}^{N} \lambda_n (\|\mathbf{x}_n + \sum_{m=1}^{N} \lambda_m \mathbf{x}_m\|^2)$

[c] $\sum_{n=1}^{N} \lambda_n (\|\mathbf{x}_n - \sum_{m=1}^{N} \lambda_m \mathbf{x}_m\|^2) + 2(\sum_{n=1}^{N} \lambda_n \mathbf{x}_n)^2$

[d] $\sum_{n=1}^{N} \lambda_n (\|\mathbf{x}_n + \sum_{m=1}^{N} \lambda_m \mathbf{x}_m\|^2) + 2(\sum_{n=1}^{N} \lambda_n \mathbf{x}_n)^2$

[✓ **[e]**] none of the other choices

9. (MCQ) Consider using $\mathbf{z}_n = \boldsymbol{\phi}(\mathbf{x}_n)$ 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 Objective($\boldsymbol{\lambda}$) of (D') after applying the kernel trick?

[a] $\sum_{n=1}^{N} \lambda_n K(\mathbf{x}_n, \mathbf{x}_n) + 3 \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m K(\mathbf{x}_n, \mathbf{x}_m)$

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

[c] $\sum_{n=1}^{N} \lambda_n K(\mathbf{x}_n, \mathbf{x}_n) - 1 \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m K(\mathbf{x}_n, \mathbf{x}_m)$

[d] $\sum_{n=1}^{N} \lambda_n K(\mathbf{x}_n, \mathbf{x}_n) - 3 \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m K(\mathbf{x}_n, \mathbf{x}_m)$

[✓ **[e]**] none of the other choices

10. (MCQ) After solving the (D') that involves the kernel K, which of the following formula evaluates the optimal R?

[✓ **[a]**] Pick some i with $\lambda_i > 0$, and $R = \sqrt{K(\mathbf{x}_i, \mathbf{x}_i) - 2 \sum_{m=1}^{N} \lambda_m K(\mathbf{x}_i, \mathbf{x}_m) + \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m K(\mathbf{x}_n, \mathbf{x}_m)}$

[b] Pick some i with $\lambda_i = 0$, and $R = \sqrt{K(\mathbf{x}_i, \mathbf{x}_i) - 2 \sum_{m=1}^{N} \lambda_m K(\mathbf{x}_i, \mathbf{x}_m) + \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m K(\mathbf{x}_n, \mathbf{x}_m)}$

[c] Pick some i with $\lambda_i > 0$, and $R = \sqrt{K(\mathbf{x}_i, \mathbf{x}_i) + 2 \sum_{m=1}^{N} \lambda_m K(\mathbf{x}_i, \mathbf{x}_m) + \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m K(\mathbf{x}_n, \mathbf{x}_m)}$

[d] Pick some i with $\lambda_i = 0$, and $R = \sqrt{K(\mathbf{x}_i, \mathbf{x}_i) + 2 \sum_{m=1}^{N} \lambda_m K(\mathbf{x}_i, \mathbf{x}_m) + \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m K(\mathbf{x}_n, \mathbf{x}_m)}$

[e] none of the other choices

Dual Problem of ℓ_2 Loss Soft-Margin Support Vector Machines

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

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

The support vector machine (called ℓ_1 loss) penalizes the margin violation linearly. Another popular formulation (called ℓ_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.

$$\begin{aligned} (P_2') \quad \min_{\mathbf{w}, b, \boldsymbol{\xi}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N} \xi_n^2 \\ \text{subject to} \quad & y_n\Big(\mathbf{w}^T\mathbf{x}_n + b\Big) \ge 1 - \xi_n, \\ & \xi_n \ge 0. \end{aligned}$$

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.

$$
\begin{array}{lll}
(P_2) & \displaystyle\min_{\mathbf{w},b,\boldsymbol{\xi}} & \displaystyle\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\xi_n^2 \\
 & \text{subject to} & y_n\Big(\mathbf{w}^T\mathbf{x}_n + b\Big) \geq 1 - \xi_n.
\end{array}
$$

11. (MCQ) 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). Which of the following is $\tilde{\mathbf{x}}_n$? (*Hint: let* $\tilde{\mathbf{w}} = (\mathbf{w}, \textit{constant} \cdot \boldsymbol{\xi})$)

[✓ [a]] $\tilde{\mathbf{x}}_n = (\mathbf{x}_n, v_1, v_2, \cdots, v_N)$, where $v_i = \frac{1}{\sqrt{2C}} [\![i = n]\!]$

[b] $\tilde{\mathbf{x}}_n = (\mathbf{x}_n, v, v, \cdots, v)$, where there are N components of $v = \frac{1}{\sqrt{2C}}$

[c] $\tilde{\mathbf{x}}_n = (\mathbf{x}_n, v_1, v_2, \cdots, v_N)$, where $v_i = \frac{1}{\sqrt{C}} [\![i = n]\!]$

[d] $\tilde{\mathbf{x}}_n = (\mathbf{x}_n, v, v, \cdots, v)$, where there are N components of $v = \frac{1}{\sqrt{C}}$

[e] none of the other choices

Operation of Kernels

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.

12. (MRQ) Which of the followings are always valid kernels, assuming that $K_2(\mathbf{x}, \mathbf{x}') \neq 0$ for all $\mathbf{x}$ and $\mathbf{x}'$?

[✓ [a]] $K(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') + K_2(\mathbf{x}, \mathbf{x}')$

[b] $K(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') - K_2(\mathbf{x}, \mathbf{x}')$

[✓ [c]] $K(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') \cdot K_2(\mathbf{x}, \mathbf{x}')$

[d] $K(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}')/K_2(\mathbf{x}, \mathbf{x}')$

[e] none of the other choices

13. (MRQ) Which of the followings are always valid kernels?

[a] $K(\mathbf{x}, \mathbf{x}') = (1 - K_1(\mathbf{x}, \mathbf{x}'))^2$

[✓ [b]] $K(\mathbf{x}, \mathbf{x}') = 1126 \cdot K_1(\mathbf{x}, \mathbf{x}')$

[c] $K(\mathbf{x}, \mathbf{x}') = \exp(-K_1(\mathbf{x}, \mathbf{x}'))$

[d] $K(\mathbf{x}, \mathbf{x}') = (1 - K_1(\mathbf{x}, \mathbf{x}'))^{-1}$, assuming that $0 < K_1(\mathbf{x}, \mathbf{x}') < 1$

[e] none of the other choices

Kernel Scaling and Shifting

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 $q > 0$.

14. (MCQ) Which of the following statement is true?

[a] 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.

[b] 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.

[✓ [c]] 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.

[d] 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.

[e] none of the other choices

Experiments with Soft-Margin Support Vector Machine

Next, 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 symmetry and intensity 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 symmetry intensity`

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:

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

In the following questions, please use the 0/1 error for evaulating E_{in}, E_{val} and E_{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.

15. (MCQ, *) 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 $K(\mathbf{x}_n,\mathbf{x}_m) = \mathbf{x}_n^{\mathrm{T}}\mathbf{x}_m$ 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?

[a] 0.2

[✓ [b] 0.6]

[c] 1.0

[d] 1.4

[e] 1.8

16. (MCQ, *) Consider the polynomial kernel $K(\mathbf{x}_n,\mathbf{x}_m) = (1+\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_m)^Q$, 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_{in}?

[a] "0" versus "not 0"

[b] "2" versus "not 2"

[c] "4" versus "not 4"

[d] "6" versus "not 6"

[✓ [e] "8" versus "not 8"]

17. (MCQ, *) 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?

[a] 5.0

[b] 10.0

[c] 15.0

[✓ [d] 20.0]

[e] 25.0

18. (MRQ, *) Consider the Gaussian kernel $K(\mathbf{x}_n, \mathbf{x}_m) = \exp\left(-\gamma||\mathbf{x}_n - \mathbf{x}_m||^2\right)$. 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 C?

[✓ [a]] the distance of any unbounded support vector to the hyperplane in the (infinite-dimensional) $\mathcal{Z}$ space

[✓ [b]] $\sum_{n=1}^{N} \xi_n$

[c] number of support vectors

[d] E_{out}

[✓ [e]] the objective value of the dual problem

19. (MCQ, *) Following Question 18, when fixing $C = 0.1$, which of the following values of γ results in the lowest E_{out}?

[a] 1

[✓ [b] 10]

[c] 100

[d] 1000

[e] 10000

20. (MCQ, *) 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 γ among $\{1, 10, 100, 1000, 10000\}$ according to E_{val}. If there is a tie of E_{val}, choose the smallest γ. Repeat the procedure 100 times. Which of the following values of γ is selected the most number of times?

[a] 1

[✓ [b] 10]

[c] 100

[d] 1000

[e] 10000

Bonus: Properties of Soft-Margin SVM

21. (BBQ, 10 points) For the linear soft-margin SVM, if there is no free support vector after training, can we conclude that the data is not linearly separable?

22. (BBQ, 10 points) For the linear soft-margin SVM, if there is no free support vector after training, and all the bounded support vectors satisfy $\xi_n > 1$, can we conclude that the data is not linearly separable?

Answer guidelines. First, please write down your name and school ID number.

Name: School ID:

Then, fill in your answers for MCQ, MRQ and BFQ in the table below.

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16
17	18	19	20

Lastly, please write down your solution to those (+ ...) parts and bonus problems, using as many additional pages as you want.

Each problem is of 10 points.

- For Problem with (+ ...), the answer in the table is of 3 score points, and the (+ ...) part is of 7 score points. If your solution to the (+ ...) part is clearly different from your answer in the table, it is regarded as a suspicious violation of the class policy (plagiarism) and the TAs can deduct some more points based on the violation.
- For Problem without (+ ...), the problem is of 10 points by itself and the TAs can decide to give you partial credit or not as long as it is fair to the whole class.