Homework #5

RELEASE DATE: 11/26/2015 (**THE LUCKY DAY**, good luck with your homework!)

DUE DATE: 12/9/2015, 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.

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. Recall that N is the size of the data set and d is the dimensionality of the input space. How many variables are within the primal formulation of the linear soft-margin support vector machine problem? How many constraints are within the problem? Please explain your answers.

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. 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? Explain your answer, mathematically or pictorially.

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 $(\alpha_1, \cdots, \alpha_7)$ and numerically solve for them (you can use any package you want). What is the optimal $\boldsymbol{\alpha}$? Based on those $\boldsymbol{\alpha}$, which are the support vectors?

4. Following Question 3, what is the corresponding nonlinear curve in the $\mathcal{X}$ space? Please provie calculation steps of your choice.

5. Compare the two nonlinear curves found in Questions 2 and 4, should they be the same? Why or why not?

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

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

6. 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

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

What is $L(R, \mathbf{c}, \boldsymbol{\lambda})$? Please provide explanation of your choice.

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

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

List all KKT conditions of (P) and (D). Then, prove or disprove the following for the optimal solution of (P) and (D):

$$\text{if } \textstyle\sum_{n=1}^N \lambda_n \neq 0, \text{ then } \mathbf{c} = \left(\sum_{n=1}^N \lambda_n \mathbf{x}_n\right) \Big/ \left(\sum_{n=1}^N \lambda_n\right)$$

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 λ_n. One form of the dual problem should look like

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

Derive the dual problem step by step.

9. Continue from Question 8 and 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. What is Objective$(\boldsymbol{\lambda})$ of (D') after applying the kernel trick? Please provide the derivation steps.

10. Continue from Question 9 and solve the (D') that involves the kernel K. How can the optimal R be calculated using the kernel trick based on some i with $\lambda_i > 0$? Please provide the derivation steps.

Dual Problem of ℓ_2 Loss Soft-Margin Support Vector Machines

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

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

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.

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

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.

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

11. 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). Design such an $\tilde{\mathbf{x}}_n$ and prove how to connect $\tilde{\mathbf{w}}$ back to $\mathbf{w}$? (*Hint: there is more than one way to design so*)

Operation of Kernels

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

(+ proof of your choices)

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 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 $0 < K_1(\mathbf{x},\mathbf{x}') < 1$

(+ proof of your 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. Prove or disprove that 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.

Experiments with Soft-Margin Support Vector Machine

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:

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

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. (*) 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. For the binary classification problem of "0" versus "not 0", plot $\|\mathbf{w}\|$ versus $\log_{10} C \in \{-6, -4, -2, 0, 2\}$. Describe your findings.

16. (*) 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 $Q = 2$, and the binary classification problem of "8" versus "not 8", plot E_{in} versus $\log_{10} C \in \{-6, -4, -2, 0, 2\}$. Describe your findings.

17. (*) Following Question 16, plot $\sum_{n=1}^{N} \alpha_n$ versus $\log_{10} C \in \{-6, -4, -2, 0, 2\}$ instead. Describe your findings.

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 $\log_{10} C$ within $\{-3, -2, -1, 0, 1\}$. Plot the the distance of any free support vector to the hyperplane in the (infinite-dimensional) $\mathcal{Z}$ space versus $\log_{10} C$. Describe your findings.

19. (*) Following Question 18, when fixing $C = 0.1$, plot E_{out} versus $\log_{10}\gamma \in \{0, 1, 2, 3, 4\}$. Describe your findings.

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 $\log_{10}\gamma \in \{0, 1, 2, 3, 4\}$ according to E_{val}. If there is a tie of E_{val}, choose the smallest γ. Repeat the procedure 100 times. Plot a histogram for the number of times each $\log_{10}\gamma$ is selected.

Bonus: Dual of Dual

21. (BBQ, 20 points) Derive a simplified Langrange dual problem of the hard-margin SVM dual. Is your Lagrange dual problem of the hard-margin SVM dual the same as the hard-margin SVM primal? Are they "similar" in any sense? Describe your findings.