Homework #5

RELEASE DATE: 12/05/2013

DUE DATE: 12/19/2013, BEFORE THE END OF CLASS

QUESTIONS ABOUT HOMEWORK MATERIALS ARE WELCOMED ON THE 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.

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 of ours would come with a full credit of 200 points, with some possible bonus points.

Transforms: Explicit versus Implicit

Consider the following training set:

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

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

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

(1) (10%) Use following nonlinear transformation of the input vector $\mathbf{x}$ to the transformed vector $\mathbf{z} = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}))$:

$$\phi_1(\mathbf{x}) = (\mathbf{x}[2])^2 - 2\mathbf{x}[1] - 2 \qquad \phi_2(\mathbf{x}) = (\mathbf{x}[1])^2 - 2\mathbf{x}[2] + 3$$

Write down the equation of the optimal separating "hyperplane" in the $\mathcal{Z}$ space. Then, plot the transformed training points in the $\mathcal{Z}$ space as well as the boundary between the +1 and −1 regions, and mark the on-the-fat-boundary vectors (the candidate support vectors).

(2) (10%) Following the previous problem, write down the equation of the corresponding nonlinear curve in the $\mathcal{X}$ space. Then, plot the original training points on the $\mathcal{X}$ plane as well as the boundary between the +1 and −1 regions, and mark the on-the-fat-boundary vectors (the candidate support vectors).

(3) (10%) Consider the same training set, but instead of explicitly transforming the input space $\mathcal{X}$, apply the (hard-margin) SVM algorithm with the kernel function

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

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 $\boldsymbol{\alpha}$?

(4) (10%) Following the previous problem, write down the equation of the corresponding nonlinear curve in the $\mathcal{X}$ space. Then, plot the original training points on the $\mathcal{X}$ plane as well as the boundary between the $+1$ and -1 regions, and mark the on-the-fat-boundary vectors (the candidate support vectors).

(5) (10%) Should the two nonlinear curves (and candidate support vectors) found in Questions 2 and 4 be the same? Why or why not? Make a comparison and briefly describe your findings.

Dual Problem of L2-Error Soft-Margin Support Vector Machines

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

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

The support vector machine penalizes the margin violation linearly. Another popular formulation penalizes the margin violation quadratically. In this problem, we derive the dual of such a formulation. The formulation is as follows.

$$\begin{aligned}(P_2')\min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\xi_n^2 \\ \text{s.t.} \quad & y_n\Big(\mathbf{w}^T\mathbf{x}_n + b\Big) \geq 1 - \xi_n, \text{ for } n = 1, 2, \cdots, N; \\ & \xi_n \geq 0, \text{ for } n = 1, 2, \cdots, N.\end{aligned}$$

(6) (10%) Argue 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.

$$\begin{aligned}(P_2)\min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\xi_n^2 \\ \text{s.t.} \quad & y_n\Big(\mathbf{w}^T\mathbf{x}_n + b\Big) \geq 1 - \xi_n, \text{ for } n = 1, 2, \cdots, N.\end{aligned}$$

(7) (10%) Let α_n be the Lagrange multipliers for the n-th constraint in (P_2). Following the derivation of the dual SVM in class, write down (P_2) as an equivalent optimization problem

$$\min_{(b,\mathbf{w},\boldsymbol{\xi})} \max_{\alpha_n \geq 0} \mathcal{L}((b,\mathbf{w},\boldsymbol{\xi}),\boldsymbol{\alpha}).$$

What is $\mathcal{L}((b,\mathbf{w},\boldsymbol{\xi}),\boldsymbol{\alpha})$?

(8) (10%) Using (assuming) strong duality, the solution to (P_2) would be the same as the Lagrange dual problem

$$\max_{\alpha_n \geq 0} \min_{(b,\mathbf{w},\boldsymbol{\xi})} \mathcal{L}((b,\mathbf{w},\boldsymbol{\xi}),\boldsymbol{\alpha}).$$

Use the KKT conditions to simplify the Lagrange dual problem, and obtain a dual problem that involves only α_n.

(9) (10%) Explain what would happen when we use $\mathbf{z}_n = \boldsymbol{\phi}(\mathbf{x}_n)$ instead of $\mathbf{x}_n$, and write down the optimization problem that uses $K(\mathbf{x}_n, \mathbf{x}_m)$ to replace $\boldsymbol{\phi}(\mathbf{x}_n)^T\boldsymbol{\phi}(\mathbf{x}_m)$—that is, the kernel trick.

Operation of Kernels

(10) (10%) For two valid kernels $K_1(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}_1(\mathbf{x})^T \boldsymbol{\phi}_1(\mathbf{x}')$ and $K_2(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}_2(\mathbf{x})^T \boldsymbol{\phi}_2(\mathbf{x}')$, consider a kernel function $K(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') \cdot K_2(\mathbf{x}, \mathbf{x}')$. Prove that K is a valid kernel by deriving a transform function $\boldsymbol{\phi}(\mathbf{x})$ such that

$$K(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\phi}(\mathbf{x}').$$

(The result shows that a multiplication of valid kernels is still a valid kernel.)

(11) (10%) Prove that the function $K(\mathbf{x}, \mathbf{x}') = -2(\mathbf{x}^T\mathbf{x}')^2 + 3\mathbf{x}^T\mathbf{x}'$ is NOT a valid kernel function.

(Hint: Check Mercer's condition.)

(12) (10%) Let $K_1(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}_1(\mathbf{x})^T \boldsymbol{\phi}_1(\mathbf{x}')$ be a kernel function. Consider a kernel function $K(\mathbf{x}, \mathbf{x}') = \exp(K_1(\mathbf{x}, \mathbf{x}'))$. Prove that K is a valid kernel by deriving a transform function $\boldsymbol{\phi}(\mathbf{x})$ such that

$$K(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\phi}(\mathbf{x}').$$

Large-Margin Perceptron Classification

(13) (20%, *) Implement the large-margin perceptron (linear hard-margin SVM) formulation below:

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{subject to} \quad & y_n\left(\mathbf{w}^T\mathbf{x}_n + b\right) \geq 1 \text{ for } n = 1, 2, \ldots, N. \end{aligned}$$

Use the following data set for training:

`http://www.csie.ntu.edu.tw/~htlin/course/ml13fall/hw5/hw5_13_train.dat`

and the following set for testing

`http://www.csie.ntu.edu.tw/~htlin/course/ml13fall/hw5/hw5_13_test.dat`

Run 100 experiments. In each experiment, randomly sampling 80% of the training set above, and obtain $g(\mathbf{x}) = \text{sign}\left(\mathbf{w}^T\mathbf{x} + b\right)$ with $(b, \mathbf{w})$ by applying SVM on the 80%-subset. Then, record the following two items:

- the margin of the hyperplane
- the out-of-sample error E_{out} of g

Make a scatter plot of the 100 pairs of (margin, E_{out}). Briefly state your findings.

(*Note: You can use any general-purpose packages for quadratic programming to solve this problem, but you **cannot** use any SVM-specific packages.*)

Experiments with Three SVMs

(14) (20%, *) Write a program to implement the nonlinear soft-margin Support Vector Machine by solving

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^{N}\alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^{N} y_i\alpha_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned}$$

Use the following set for training:

```
http://www.csie.ntu.edu.tw/~htlin/course/ml13fall/hw5/hw5_14_train.dat
```

Consider the Gaussian-RBF kernel $\exp\left(\frac{-\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right)$ with $\sigma = 0.125, 0.5, 2$ and $C = 0.001, 1, 1000$. For each (σ, C) combination, show E_{in}, E_{cv} with 5-fold cross validation, and $\frac{\#SV}{N}$ (an upper-bound of leave-one-out cross validation error). Briefly describe your findings. (*Note: For this problem, you CAN use any package you want. A recommended choice is LIBSVM developed by Prof. Chih-Jen Lin in our department*)

(15) (20%, *) Write a program to implement the nonlinear SVR from Lecture 22, and use the SVR for classification. Use the following set for training:

```
http://www.csie.ntu.edu.tw/~htlin/course/ml13fall/hw5/hw5_14_train.dat
```

Consider the Gaussian-RBF kernel $\exp\left(\frac{-\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right)$ with $\sigma = 0.125, 0.5, 2$ and $C = 0.001, 1, 1000$. For each (σ, C) combination, show E_{in} (by 0/1 error), E_{cv} (by 0/1 error) with 5-fold cross validation, and $\frac{\#SV}{N}$. Briefly describe your findings. (*Note: For this problem, you CAN use any package you want. A recommended choice is LIBSVM developed by Prof. Chih-Jen Lin in our department*)

(16) (20%, *) Write a program to implement the nonlinear LSSVR from Lecture 22, and use it for classification. Use the following set for training:

```
http://www.csie.ntu.edu.tw/~htlin/course/ml13fall/hw5/hw5_14_train.dat
```

Consider the Gaussian-RBF kernel $\exp\left(\frac{-\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right)$ with $\sigma = 0.125, 0.5, 2$ and $\lambda = 0.001, 1, 1000$. For each (σ, λ) combination, show E_{in} (by 0/1 error), E_{cv} (by 0/1 error) with 5-fold cross validation, and $\frac{\#SV}{N}$. Briefly describe your findings.

Bonus: L2-Loss and Hard-Margin

(17) (Bonus 10%) The L2-Loss soft-margin SVM (P_2) is actually equivalent to a hard-margin SVM that takes examples $(\tilde{\mathbf{x}}_n, \tilde{y}_n)$ instead of $(\mathbf{x}_n, y_n)$. Write down $(\tilde{\mathbf{x}}_n, \tilde{y}_n)$ and prove the equivalence.

(*Note: You can actually use the equivalence to make the derivations in Questions 6-9 simpler.*)

Bonus: Kernel Scaling

For a valid kernel K, consider a new kernel

$$\tilde{K}(\mathbf{x}, \mathbf{x}') = pK(\mathbf{x}, \mathbf{x}')$$

for some positive p. It is not difficult to see that $\tilde{K}$ is also a valid kernel.

(18) (Bonus 10%) Argue that for the dual of soft-margin SVM, using $\tilde{K}$ along with some new $\tilde{C}$ instead of K with some original C leads to an equivalent solution. What is the relation between $\tilde{C}$ and C?