# Homework #2

RELEASE DATE: 04/24/2018

DUE DATE: 05/29/2018, BEFORE 14:00

QUESTIONS ABOUT HOMEWORK MATERIALS ARE WELCOMED ON THE FACEBOOK 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 160 points and 40 bonus points. In general, every homework set would come with a full credit of 160 points, with some possible bonus points.

### Descent Methods for Probabilistic SVM

Recall that the probabilistic SVM is based on solving the following optimization problem:

$$\min_{A,B} \quad F(A,B) = \frac{1}{N}\sum_{n=1}^{N} \ln\left(1+\exp\left(-y_n\left(A\cdot\left(\mathbf{w}_{\text{SVM}}^T\boldsymbol{\phi}(\mathbf{x}_n)+b_{\text{SVM}}\right)+B\right)\right)\right).$$

**1.** When using the gradient descent for minimizing $F(A,B)$, we need to compute the gradient first. Let $z_n = \mathbf{w}_{\text{SVM}}^T\boldsymbol{\phi}(\mathbf{x}_n)+b_{\text{SVM}}$, and $p_n = \theta(-y_n(Az_n+B))$, where $\theta(s) = \frac{\exp(s)}{1+\exp(s)}$ is the usual logistic function. What is the gradient $\nabla F(A,B)$ in terms of only $y_n, p_n, z_n$ and $N$? Prove your answer.

**2.** When using the Newton method for minimizing $F(A,B)$ (see Homework 3 of Machine Learning Foundations), we need to compute $-(H(F))^{-1}\nabla F$ in each iteration, where $H(F)$ is the Hessian matrix of $F$ at $(A,B)$. Following the notations of Problem 1, what is $H(F)$ in terms of only $y_n, p_n, z_n$ and $N$? Prove your answer.

### Kernel Ridge Regression

**3.** Assume that all $\mathbf{x}_n$ are different. When using the Gaussian kernel with $\gamma \to \infty$, what does the kernel matrix K used in kernel ridge regression look like? What is the optimal $\boldsymbol{\beta}$? Prove your answer.

### Blending

**4.** Consider $T+1$ hypotheses $g_0, g_1, \cdots, g_T$. Let $g_0(\mathbf{x}) = 0$ for all $\mathbf{x}$. Assume that your boss holds a test set $\{(\tilde{\mathbf{x}}_m, \tilde{y}_m)\}_{m=1}^M$, where you know $\tilde{\mathbf{x}}_m$ but $\tilde{y}_m$ is hidden. Nevertheless, you are allowed

to know the squared test error $E_{\text{test}}(g_t) = \frac{1}{M}\sum_{m=1}^{M}(g_t(\tilde{\mathbf{x}}_m) - \tilde{y}_m)^2 = e_t$ for $t = 0, 1, 2, \cdots, T$. Also, assume that $\frac{1}{M}\sum_{m=1}^{M}(g_t(\tilde{\mathbf{x}}_m))^2 = s_t$. In terms of all $M$, $e_t$, and $s_t$, how do you calculate $\sum_{m=1}^{M} g_t(\tilde{\mathbf{x}}_m)\tilde{y}_m$? Prove your answer.

**5.** For the given $T + 1$ hypotheses in the previous problem, design an algorithm to solve

$$\min_{\alpha_0,\alpha_1,\cdots,\alpha_T} E_{\text{test}}(\sum_{t=0}^{T} \alpha_t g_t),$$

and obtain the optimal weights $\alpha_0, \cdots, \alpha_T$. The algorithm the key to the test set blending technique that the NTU team has used in KDDCup 2011.

**6.** Consider the case where the target function $f : [0, 1] \to \mathbb{R}$ is given by $f(x) = 2x - x^2$ and the input probability distribution is uniform on $[0, 1]$. Assume that the training set has only two examples generated independently from the input probability distribution and noiselessly by $f$, and the learning model is usual linear regression that minimizes the mean squared error within all hypotheses of the form $h(x) = w_1 x + w_0$. What is $\bar{g}(x)$, the expected value of the hypothesis, that the learning algorithm produces (see Page 10 of Lecture 207)? Prove your answer.

**Boosting**

**7.** Consider applying the AdaBoost algorithm on a binary classification data set where 87% of the examples are positive. Because there are so many positive examples, the base algorithm within AdaBoost returns a constant classifier $g_1(\mathbf{x}) = +1$ in the first iteration. Let $u_+^{(2)}$ be the individual example weight of each positive example in the second iteration, and $u_-^{(2)}$ be the individual example weight of each negative example in the second iteration. What is $u_+^{(2)}/u_-^{(2)}$? Prove your answer.

**Kernel for Decision Stumps**

When talking about non-uniform voting in aggregation, we mentioned that $\boldsymbol{\alpha}$ can be viewed as a weight vector learned from any linear algorithm coupled with the following transform:

$$\boldsymbol{\phi}(\mathbf{x}) = \Big(g_1(\mathbf{x}), g_2(\mathbf{x}), \cdots, g_T(\mathbf{x})\Big).$$

When studying kernel methods, we mentioned that the kernel is simply a computational short-cut for the inner product $(\boldsymbol{\phi}(\mathbf{x}))^T(\boldsymbol{\phi}(\mathbf{x}'))$. In this problem, we mix the two topics together using the decision stumps as our $g_t(\mathbf{x})$.

**8.** Assume that the input vectors contain only integers between (including) 0 and $M$.

$$\begin{aligned} & g_{s,i,\theta}(\mathbf{x}) = s \cdot \text{sign}\Big(x_i - \theta\Big), \\ \text{where} \quad & i \in \{1, 2, \cdots, d\}, d \text{ is the finite dimensionality of the input space,} \\ & s \in \{-1, +1\}, \theta \in \mathbb{R}, \text{ and } \text{sign}(0) = +1 \end{aligned}$$

Two decision stumps $g$ and $\hat{g}$ are defined as the *same* if $g(\mathbf{x}) = \hat{g}(\mathbf{x})$ for every $\mathbf{x} \in \mathcal{X}$. Two decision stumps are different if they are not the same. How many different decision stumps are there for the case of $d = 2$ and $M = 5$? Explain your answer.

**9.** Continuing from the previous problem, let $\mathcal{G} = \{$ all different decision stumps for $\mathcal{X}$ $\}$ and enumerate each hypothesis $g \in \mathcal{G}$ by some index $t$. Define

$$\boldsymbol{\phi}_{ds}(\mathbf{x}) = \left(g_1(\mathbf{x}), g_2(\mathbf{x}), \cdots, g_t(\mathbf{x}), \cdots, g_{|\mathcal{G}|}(\mathbf{x})\right).$$

For any given $(d, M)$, derive a simple equation that evaluates $K_{ds}(\mathbf{x}, \mathbf{x}') = (\boldsymbol{\phi}_{ds}(\mathbf{x}))^T(\boldsymbol{\phi}_{ds}(\mathbf{x}'))$ efficiently and prove your answer.

**10.** Assume that those integers between 0 and $M$ represents counts in histograms. A famous kernel called histogram intersection kernel is of the form

$$K_{hi}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{d} \min(x_i, x_i').$$

Consider some special decision stumps $h_t(\mathbf{x}) = \frac{g_t(\mathbf{x})+1}{2}$, where $g_t$'s are defined in the previous question. That is, $h_t(\mathbf{x})$ outputs $(0, 1)$-bits instead of $\pm 1$. Argue that for some $q_1, q_2, \ldots, q_t, \ldots$, where each $q_t \in \{0, 1\}$

$$\phi_{hi}(\mathbf{x}) = \left( q_1 h_1(\mathbf{x}), q_2 h_2(\mathbf{x}), \cdots, q_t h_t(\mathbf{x}), \cdots, q_{|\mathcal{G}|} h_{|\mathcal{G}|}(\mathbf{x}) \right).$$

Derive $q_t$'s and prove your answer. *Hint: the binary $q_t$'s physically mean selecting some $h_t$ to the transform. So the problem actually asks you to prove that some of those $h_t$'s can be used to form a transform that leads to the histogram intersection kernel.*

**Experiment with Kernel Ridge Regression**. Write a program to implement the kernel ridge regression algorithm from Lecture 206, and use it for classification (i.e. implement LSSVM). Consider the following data set

`hw2_lssvm_all.dat`

Use the first 400 examples for training and the remaining for testing. Calculate $E_{\text{in}}$ and $E_{\text{out}}$ with the 0/1 error. Consider the Gaussian-RBF kernel $\exp\left(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2\right)$. Try all combinations of parameters $\gamma \in \{32, 2, 0.125\}$ and $\lambda \in \{0.001, 1, 1000\}$.

**11.** (*) Among all parameter combinations, which combination results in the minimum $E_{\text{in}}(g)$? What is the corresponding $E_{\text{in}}(g)$?

**12.** (*) Among all parameter combinations, which combination results in the minimum $E_{\text{out}}(g)$? What is the corresponding $E_{\text{out}}(g)$?

.

**Experiment with Bagging Ridge Regression**.

First, write a program to implement linear LSSVM (i.e. linear ridge regression for classification). You can reuse the code in the previous problem if you want. Again consider the following data set

`hw2_lssvm_all.dat`

Please do add $x_0 = 1$ to your data. Use the first 400 examples for training to get $g$ and the remaining for testing. Calculate $E_{\text{in}}$ and $E_{\text{out}}$ with the 0/1 error. Consider $\lambda \in \{0.01, 0.1, 1, 10, 100\}$.

**13.** (*) Among all $\lambda$, which $\lambda$ results in the minimum $E_{\text{in}}(g)$? What is the corresponding $E_{\text{in}}(g)$?

**14.** (*) Among all $\lambda$, which $\lambda$ results in the minimum $E_{\text{out}}(g)$? What is the corresponding $E_{\text{out}}(g)$?

Next, write a program to implement bagging on top of linear LSSVM. Again consider the following data set

`hw2_lssvm_all.dat`

Please do add $x_0 = 1$ to your data. Use the first 400 examples for training and the remaining for testing. Calculate $E_{\text{in}}$ and $E_{\text{out}}$ with the 0/1 error. Note that each linear LSSVM should take the sign operation before uniform aggregation (with voting). Consider $\lambda \in \{0.01, 0.1, 1, 10, 100\}$. Use 400 bootstrapped examples in bagging and take 250 iterations of bagging (e.g. 250 $g_t$'s) to get $G$.

**15.** (*) Among all $\lambda$, which $\lambda$ results in the minimum $E_{\text{in}}(G)$? What is the corresponding $E_{\text{in}}(G)$? Compare your results with the one in Problem 13 and describe your findings.

**16.** (*) Among all $\lambda$, which $\lambda$ results in the minimum $E_{\text{out}}(G)$? What is the corresponding $E_{\text{out}}(G)$? Compare your results with the one in Problem 14 and describe your findings.

## Bonus: Equivalent "Kernels" for Soft-Margin SVM

**17.** (Bonus 20%) Argue that for soft-margin SVM with some given data set, solving the dual problem with a valid kernel $K_1(\mathbf{x}, \mathbf{x}')$ or another function $K_2(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') + \kappa$ for any constant $\kappa \in \mathbb{R}$ yields exactly the same optimal solution $\boldsymbol{\alpha}$ and exactly the same optimal $g_{\text{svm}}$. Note that $K_2$ may not need to be a valid kernel. For instance, if $K_1$ is the Gaussian kernel and $\kappa = -1126$, then $K_2$ won't be a valid kernel.

*(instructor's words: The results could help you simplify your kernel derived in Problem 9.)*

**18.** (Bonus 20%) Argue that for soft-margin SVM with some given data set, solving the dual problem with a valid kernel $K_1(\mathbf{x}, \mathbf{x}')$ or $K_3(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') + r(\mathbf{x}) + r(\mathbf{x}')$ for any function $r$ yields exactly the same optimal solution $\boldsymbol{\alpha}$ and exactly the same optimal $g_{\text{svm}}$.

*(instructor's words: The previous problem is a special case of this problem with $r(\mathbf{x}) = r(\mathbf{x}') = \frac{\kappa}{2}$. Also, consider the kernel in Problem 9 as $K_1$ and the kernel in Problem 10 as $K_3$. Now you might be able to see their connections even more easily (up to a constant scale, which can be addressed by Problem 10 of Homework 1))*