Homework #4

RELEASE DATE: 11/12/2015

DUE DATE: 11/25/2015, 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.

Overfitting and Deterministic Noise

1. Deterministic noise depends on $\mathcal{H}$, as some models approximate f better than others. Assume $\mathcal{H}' \subset \mathcal{H}$ and that f is fixed. **In general** (but not necessarily in all cases), if we use $\mathcal{H}'$ instead of $\mathcal{H}$, argue whether deterministic noise will decrease, increase, or be the same. Explain and defend your choice.

Regularization With Polynomials

Polynomial models can be viewed as linear models in a space $\mathcal{Z}$, under a nonlinear transform $\Phi : \mathcal{X} \to \mathcal{Z}$. Here, Φ transforms the scalar x into a vector $\mathbf{z}$ of Legendre polynomials, $\mathbf{z} = (1, L_1(x), L_2(x), ..., L_Q(x))$. Our hypothesis set will be expressed as a linear combination of these polynomials,

$$\mathcal{H}_Q = \left\{ h \mid h(x) = \mathbf{w}^{\mathrm{T}}\mathbf{z} = \sum_{q=0}^{Q} w_q L_q(x) \right\},$$

where $L_0(x) = 1$.

2. Consider the following hypothesis set defined by the constraint:

$$\mathcal{H}(Q, c, Q_o) = \{h \mid h(x) = \mathbf{w}^{\mathrm{T}}\mathbf{z} \in \mathcal{H}_Q; w_q = c \text{ for } q \geq Q_o\}$$

. What Q satisfies $\mathcal{H}(10, 0, 3) \cap \mathcal{H}(10, 0, 4) = \mathcal{H}_Q$? Prove your answer.

Regularization and Weight Decay

Consider the augmented error

$$E_{\text{aug}}(\mathbf{w}) = E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N}\mathbf{w}^T\mathbf{w}$$

with some $\lambda > 0$.

3. If we want to minimize the augmented error $E_{\text{aug}}(\mathbf{w})$ by gradient descent, with η as learning rate, the resulting update rule should be

$$\mathbf{w}(t+1) \longleftarrow \alpha\mathbf{w}(t) + \beta\nabla E_{\text{in}}(\mathbf{w}(t))$$

what are α and β? Prove your answer.

4. Let $\mathbf{w}_{\text{reg}}(\lambda)$ be the optimal solution for the formulation above. Prove that $\|\mathbf{w}_{\text{reg}}(\lambda)\|$ is a non-increasing function of λ for $\lambda \geq 0$.

Leave-One-Out Cross-Validation

5. You are given the data points: $(-1, 0), (\rho, 1), (1, 0)$, $\rho \geq 0$, and a choice between two models: constant $[\, h_0(x) = b_0 \,]$ and linear $[\, h_1(x) = a_1x + b_1 \,]$. For which value of ρ would the two models be tied using leave-one-out cross-validation with the squared error measure? Provide your derivation steps.

Learning Principles

In Problems 6-7, suppose that for 5 weeks in a row, a letter arrives in the mail that predicts the outcome of the upcoming Monday night baseball game.(Assume there are no tie.) You keenly watch each Monday and to your surprise, the prediction is correct each time. On the day after the fifth game, a letter arrives, stating that if you wish to see next week's prediction, a payment of NTD 1,000 is required.

6. To make sure that at least one person receives correct predictions on all 5 games from him, how many letters should be sent before the fifth game? Provide derivation steps for your answer, and describe the corresponding sending scheme.

7. If the cost of printing and mailing out each letter is NTD 10. If the sender sends the minimum number of letters out, how much money can be made for the above 'fraud' to succeed once? That is, one of the recipients does send him NTD 1,000 to receive the prediction of the 6-th game? Provide your derivation steps.

For Problems 8-10, we consider the following. In our credit card example, the bank starts with some vague idea of what constitutes a good credit risk. So, as customers $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ arrive, the bank applies its vague idea to approve credit cards for some of these customers based on a formula $a(\mathbf{x})$. Then, only those who get credit cards are monitored to see if they default or not. For simplicity, suppose that the first $N = 10,000$ customers were given credit cards by the credit approval function $a(\mathbf{x})$. Now that the bank knows the behavior of these customers, it comes to you to improve their algorithm for approving credit. The bank gives you the data $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)$. Before you look at the data, you do mathematical derivations and come up with a credit approval function. You now test it on the data and, to your delight, obtain perfect prediction.

8. What is M, the size of your hypothesis set? Defend your answer.

9. With such an M, what does the Hoeffding bound say about the probability that the true average error rate of g is worse than 1% for $N = 10,000$? Provide your calculation steps.

10. You assure the bank that you have a got a system g for approving credit cards for new customers, which is nearly error-free. Your confidence is given by your answer to the previous question. The bank is thrilled and uses your g to approve credit for new customers. To their dismay, more than half their credit cards are being defaulted on. Assume that the customers that were sent to the old credit approval function and the customers that were sent to your g are indeed i.i.d. from the same distribution, and the bank is lucky enough (so the 'bad luck' that "the true error of g is worse than 1%" does not happen). If the old credit approval function was $a(\mathbf{x})$. Describe a scheme that uses both a and g to improve the performance of a, and defend your scheme.

Virtual Examples and Regularization

Consider linear regression with virtual examples. That is, we add K virtual examples $(\tilde{\mathbf{x}}_1, \tilde{y}_1), (\tilde{\mathbf{x}}_2, \tilde{y}_2), \ldots, (\tilde{\mathbf{x}}_K, \tilde{y}_K)$ to the training data set, and solve

$$\min_{\mathbf{w}} \frac{1}{N+K} \left(\sum_{n=1}^{N} (y_n - \mathbf{w}^T \mathbf{x}_n)^2 + \sum_{k=1}^{K} (\tilde{y}_k - \mathbf{w}^T \tilde{\mathbf{x}}_k)^2 \right).$$

We will show that using some 'special' virtual examples, which were claimed to be a possible way to combat overfitting in Lecture 9, is related to regularization, another possible way to combat overfitting discussed in Lecture 10. Let $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1 \tilde{\mathbf{x}}_2 \ldots \tilde{\mathbf{x}}_K]^T$, and $\tilde{\mathbf{y}} = [\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_K]^T$.

11. What is the optimal $\mathbf{w}$ to the optimization problem above, assuming that all the inversions exist? Provide an analytic solution and prove its correctness.

12. For what $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ will the solution of this linear regression equal to

$$\mathbf{w}_{\text{reg}} = \arg\min_{\mathbf{w}} \frac{\lambda}{N} \|\mathbf{w}\|^2 + \frac{1}{N} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2?$$

Prove your answer.

Experiment with Regularized Linear Regression and Validation

Consider regularized linear regression (also called ridge regression) for classification.

$$\mathbf{w}_{\text{reg}} = \arg\min_{\mathbf{w}} \frac{\lambda}{N} \|\mathbf{w}\|^2 + \frac{1}{N} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2,$$

Run the algorithm on the following data set as $\mathcal{D}$:

`http://www.csie.ntu.edu.tw/~htlin/course/ml15fall/hw4/hw4_train.dat`

and the following set for evaulating E_{out}:

`http://www.csie.ntu.edu.tw/~htlin/course/ml15fall/hw4/hw4_test.dat`

Because the data sets are for classification, please consider only the 0/1 error for all the problems below.

13. (*) Let $\lambda = 11.26$, what is the corresponding E_{in} and E_{out}?

14. (*) Plot the curve of E_{in} with respect to $\log_{10} \lambda = \{2, 1, 0, -1, \ldots, -8, -9, -10\}$. What is the λ with the minimum E_{in}? What is $E_{\text{out}}(g_\lambda)$ on such λ? Break the tie by selecting the largest λ.

15. (*) Plot the curve of E_{out} with respect to $\log_{10} \lambda = \{2, 1, 0, -1, \ldots, -8, -9, -10\}$. What is the λ with the minimum E_{out}? Break the tie by selecting the largest λ.

Now split the given training examples in $\mathcal{D}$ to the first 120 examples for $\mathcal{D}_{\text{train}}$ and 80 for $\mathcal{D}_{\text{val}}$.

Ideally, you should randomly do the 120/80 split. Because the given examples are already randomly permuted, however, we would use a fixed split for the purpose of this problem.

Run the algorithm on $\mathcal{D}_{\text{train}}$ to get g_λ^-, and validate g_λ^- with $\mathcal{D}_{\text{val}}$.

16. (*) Plot $E_{\text{train}}(g_\lambda^-)$ with respect to $\log_{10} \lambda = \{2, 1, 0, -1, \ldots, -8, -9, -10\}$. What is the λ with the minimum $E_{\text{train}}(g_\lambda^-)$? What is $E_{\text{out}}(g_\lambda^-)$ on such λ? Break the tie by selecting the largest λ.

17. (*) Plot $E_{\text{val}}(g_\lambda^-)$ with respect to $\log_{10} \lambda = \{2, 1, 0, -1, \ldots, -8, -9, -10\}$. What is the λ with the minimum $E_{\text{val}}(g_\lambda^-)$? What is $E_{\text{out}}(g_\lambda^-)$ on such λ? Break the tie by selecting the largest λ.

18. (*) Run the algorithm with the optimal λ of the previous problem on the whole $\mathcal{D}$ to get g_λ. What is $E_{\text{in}}(g_\lambda)$ and $E_{\text{out}}(g_\lambda)$?

Now split the given training examples in $\mathcal{D}$ to five folds, the first 40 being fold 1, the next 40 being fold 2, and so on. Again, we take a fixed split because the given examples are already randomly permuted.

19. (*) Plot E_{cv} with respect to $\log_{10}\lambda = \{2, 1, 0, -1, \ldots, -8, -9, -10\}$. What is the λ with the minimum E_{cv}, where E_{cv} comes from the five folds defined above? Break the tie by selecting the largest λ.

20. (*) Run the algorithm with the optimal λ of the previous problem on the whole $\mathcal{D}$ to get g_λ. What is $E_{\text{in}}(g_\lambda)$ and $E_{\text{out}}(g_\lambda)$?

Bonus: More on Virtual Examples

21. (BBQ, 10 points) Continue from Problem 12. Assume that we take the more general

$$\mathbf{w}^T\mathbf{\Gamma}^T\mathbf{\Gamma}\mathbf{w}$$

as the regularizer instead of the squared $\mathbf{w}^T\mathbf{w}$. This is commonly called Tikhonov regularization. What virtual examples should we equivalently add to the original data set?

22. (BBQ, 10 points) Continue from Problem 12. Assume that we have some known hints $\mathbf{w}_{\text{hint}}$ about the rough value of $\mathbf{w}$ and hence want to use

$$\|\mathbf{w} - \mathbf{w}_{\text{hint}}\|^2$$

as the regularizer instead of the squared $\mathbf{w}^T\mathbf{w}$. What virtual examples should we equivalently add to the original data set?