Homework #4

RELEASE DATE: 11/12/2013

DUE DATE: 11/28/2013, BEFORE NOON

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.

There are three 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.*
- *blank-filling question (BFQ): You should write down the answer that we ask you to fill.*

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

Overfitting and Deterministic Noise

1. (MCQ) 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 $\mathcal{H}$, how does deterministic noise behave?

[a] In general, deterministic noise will decrease.

[b] In general, deterministic noise will increase.

[c] In general, deterministic noise will be the same.

[d] There is no trend.

(+ explanation of 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. (MCQ) Consider the following hypothesis set defined by the constraint:

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

which of the following statements is correct:

[a] $\mathcal{H}(10, 0, 3) \cup \mathcal{H}(10, 0, 4) = \mathcal{H}_4$

[b] $\mathcal{H}(10, 0, 3) \cup \mathcal{H}(10, 1, 4) = \mathcal{H}_3$

[c] $\mathcal{H}(10, 0, 3) \cap \mathcal{H}(10, 0, 4) = \mathcal{H}_2$

[d] $\mathcal{H}(10, 1, 3) \cap \mathcal{H}(10, 1, 4) = \mathcal{H}_1$

[e] None of the above

(+ explanation of your choice)

Regularization and Weight Decay

Consider the augmented error

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

with some $\lambda > 0$.

3. (MRQ) If we want to minimize the augmented error $E_{\mathrm{aug}}(\mathbf{w})$ by gradient descent, with η as learning rate, which of the followings are the correct update rules:

[a] $\mathbf{w}(t+1) \longleftarrow \mathbf{w}(t) - \eta\nabla E_{\mathrm{aug}}(\mathbf{w})$.

[b] $\mathbf{w}(t+1) \longleftarrow \mathbf{w}(t) - \eta\nabla E_{\mathrm{in}}(\mathbf{w})$.

[c] $\mathbf{w}(t+1) \longleftarrow (1 - \frac{\eta\lambda}{N})\mathbf{w}(t) - \eta\nabla E_{\mathrm{in}}(\mathbf{w})$.

[d] $\mathbf{w}(t+1) \longleftarrow (1 - \frac{2\eta\lambda}{N})\mathbf{w}(t) - \eta\nabla E_{\mathrm{in}}(\mathbf{w})$.

[e] None of the above

(+ explanation of your choice)

4. (MRQ) Let $\mathbf{w}_{\mathrm{lin}}$ be the optimal solution for the plain-vanilla linear regression and $\mathbf{w}_{\mathrm{reg}}(\lambda)$ be the optimal solution for the formulation above. Select all the correct statements below:

[a] $\|\mathbf{w}_{\mathrm{reg}}(\lambda)\| \geq \|\mathbf{w}_{\mathrm{lin}}\|$ for any $\lambda > 0$

[b] $\|\mathbf{w}_{\mathrm{reg}}(\lambda)\| \leq \|\mathbf{w}_{\mathrm{lin}}\|$ for any $\lambda > 0$

[c] $\|\mathbf{w}_{\mathrm{reg}}(\lambda)\|$ is a non-increasing function of λ for $\lambda \geq 0$

[d] $\|\mathbf{w}_{\mathrm{reg}}(\lambda)\|$ is a non-decreasing function of λ for $\lambda \geq 0$

[e] None of the above

(+ proof of your choice)

Leave-One-Out Cross-Validation

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

[a] $\sqrt{\sqrt{3} + 4}$

[b] $\sqrt{\sqrt{3} - 1}$

[c] $\sqrt{9 + 4\sqrt{6}}$

[d] $\sqrt{9 - \sqrt{6}}$

[e] None of the above

(+ calculating step of your choice)

Learning Principles

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. 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. (MRQ) Which of the following statements are true?

[a] There are 32 win-lose predictions for 5 games.

[b] If the sender wants to make sure that at least one person receives correct predictions on all 5 games from him, the sender should target to begin with at least 5 people.

[c] After the first letter 'predicts' the outcome of the first game, the sender should target 16 people with the second letter.

[d] The number of letter sent at the end of the 5 weeks is 64.

[e] None of the above.

(+ explanation of your choice)

7. (MCQ) If the cost of printing and mailing out each letter is NTD 10, what is the maximum amount of money that the sender can make if one of the receipients does send him NTD 1,000 to receive the prediction of the 6-th game?

[a] NTD 340

[b] NTD 370

[c] NTD 400

[d] NTD 430

[e] None of the above.

(+ calculating step of your choice)

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. (MCQ) What is M, the size of your hypothesis set?

[a] 1

[b] N

[c] 2^N

[d] We have no idea about it.

[e] None of the above.

(+ explanation of your choice)

9. (MCQ) 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 $N = 10,000$?

[a] ≤ 0.171

[b] ≤ 0.221

[c] ≤ 0.271

[d] ≤ 0.321

[e] None of the above.

(+ calculating step of your choice)

10. (MRQ) 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' in the previous problem does not happen). Select all the true claims for this situation.

[a] If the old credit approval function was $a(\mathbf{x}) = +1$ (approve all customers), the situation should not happen.

[b] By applying $a(\mathbf{x})$ OR $g(\mathbf{x})$ to approve credit for new customers, the performance of the overall credit approval system can be improved with guarantee provided by the previous problem.

[c] By applying $a(\mathbf{x})$ AND $g(\mathbf{x})$ to approve credit for new customers, the performance of the overall credit approval system can be improved with guarantee provided by the previous problem.

[d] By applying $a(\mathbf{x})$ XOR $g(\mathbf{x})$ to approve credit for new customers, the performance of the overall credit approval system can be improved with guarantee provided by the previous problem.

[e] None of the above.

(+ explanation of your choice)

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. (MCQ) What is the optimal $\mathbf{w}$ to the optimization problem above, assuming that all the inversions exist?

[a] $(\mathbf{X}^T\mathbf{X})^{-1}(\tilde{\mathbf{X}}^T\tilde{\mathbf{y}})$
[b] $(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y} + \tilde{\mathbf{X}}^T\tilde{\mathbf{y}})$
[c] $(\mathbf{X}^T\mathbf{X} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}(\tilde{\mathbf{X}}^T\tilde{\mathbf{y}})$
[d] $(\mathbf{X}^T\mathbf{X} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}(\mathbf{X}^T\mathbf{y} + \tilde{\mathbf{X}}^T\tilde{\mathbf{y}})$
[e] None of the above.

(+ proof of your choice)

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

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

[a] $\tilde{\mathbf{X}} = \mathbf{I}$, $\tilde{\mathbf{y}} = \mathbf{0}$
[b] $\tilde{\mathbf{X}} = \sqrt{\lambda}\mathbf{I}$, $\tilde{\mathbf{y}} = \mathbf{0}$
[c] $\tilde{\mathbf{X}} = \lambda\mathbf{I}$, $\tilde{\mathbf{y}} = \mathbf{1}$
[d] $\tilde{\mathbf{X}} = \sqrt{\lambda}\mathbf{X}$, $\tilde{\mathbf{y}} = \mathbf{y}$
[e] None of the above.

(+ proof of your choice)

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/ml13fall/doc/hw4_train.dat`

and the following set for evaulating E_{out}:

`http://www.csie.ntu.edu.tw/~htlin/course/ml13fall/doc/hw4_test.dat`

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

13. (BFQ, *) Let $\lambda = 10$, report E_{in} and E_{out}.

14. (BFQ, *) Among $\log_{10}\lambda = \{2, 1, 0, -1, \ldots, -8, -9, -10\}$. What is the λ with the minimum E_{in}? Report λ and the corresponding E_{in} and E_{out}.

15. (BFQ, *) Among $\log_{10}\lambda = \{2, 1, 0, -1, \ldots, -8, -9, -10\}$. What is the λ with the minimum E_{out}? Report λ and the corresponding E_{in} and E_{out}.

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. (BFQ, *) Among $\log_{10}\lambda = \{2, 1, 0, -1, \ldots, -8, -9, -10\}$. What is the λ with the minimum $E_{\text{train}}(g_\lambda^-)$? Report λ and the corresponding $E_{\text{train}}(g_\lambda^-)$, $E_{\text{val}}(g_\lambda^-)$ and $E_{\text{out}}(g_\lambda^-)$.

17. (BFQ, *) Among $\log_{10}\lambda = \{2, 1, 0, -1, \ldots, -8, -9, -10\}$. What is the λ with the minimum $E_{\text{val}}(g_\lambda^-)$? Report λ and the corresponding $E_{\text{train}}(g_\lambda^-)$, $E_{\text{val}}(g_\lambda^-)$ and $E_{\text{out}}(g_\lambda^-)$.

18. (BFQ, *) Run the algorithm with the optimal λ of the previous problem on the whole $\mathcal{D}$ to get g_λ. Report $E_{\text{in}}(g_\lambda)$ 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. (BFQ, *) Among $\log_{10} \lambda = \{2, 1, 0, -1, \ldots, -8, -9, -10\}$. What is the λ with the minimum E_{cv}, where E_{cv} comes from the five folds defined above? Report λ and the corresponding E_{cv}.

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

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.