Homework #3.5

RELEASE DATE: 12/29/2016

DUE DATE: **1/17/2017** (**TUESDAY!!!**), 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 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.

Linear Regression

1. Consider a noisy target $y = \mathbf{w}_f^T \mathbf{x} + \epsilon$, where $\mathbf{x} \in \mathbb{R}^d$ (with the added coordinate $x_0 = 1$), $y \in \mathbb{R}$, $\mathbf{w}_f$ is an unknown vector, and ϵ is a noise term with zero mean and σ^2 variance. Assume ϵ is independent of $\mathbf{x}$ and of all other ϵ's. If linear regression is carried out using a training data set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, and outputs the parameter vector $\mathbf{w}_{\text{lin}}$, it can be shown that the expected in-sample error E_{in} with respect to $\mathcal{D}$ is given by:

$$\mathbb{E}_{\mathcal{D}}[E_{\text{in}}(\mathbf{w}_{\text{lin}})] = \sigma^2 \left(1 - \frac{d+1}{N}\right)$$

For $\sigma = 0.1$ and $d = 8$, what is the smallest number of examples N that will result in an expected E_{in} greater than 0.008? Please provide calculation steps of your answer.

Error and SGD

2. Which of the following are upper bounds of $[\![\text{sign}(\mathbf{w}^T\mathbf{x}) \neq y]\!]$ for $y \in \{-1, +1\}$? Explain your choices.

[a] $err(\mathbf{w}) = \max(0, 1 - y\mathbf{w}^T\mathbf{x})$

[b] $err(\mathbf{w}) = \big(\max(0, 1 - y\mathbf{w}^T\mathbf{x})\big)^2$

graph

[c] $err(\mathbf{w}) = \max(0, -y\mathbf{w}^T\mathbf{x})$

[d] $err(\mathbf{w}) = \theta(-y\mathbf{w}^T\mathbf{x})$

[e] $err(\mathbf{w}) = \exp(-y\mathbf{w}^T\mathbf{x})$

3. When using SGD on the following error functions and 'ignoring' some singular points that are not differentiable, prove or disprove that $err(\mathbf{w}) = \max(0, -y\mathbf{w}^T\mathbf{x})$ results in PLA.

Gradient Descent and Beyond

4. Consider a function

program

$$E(u, v) = e^u + e^{2v} + e^{uv} + u^2 - 2uv + 2v^2 - 3u - 2v.$$

In class, we have taught that the update rule of the gradient descent algorithm is

$$(u_{t+1}, v_{t+1}) = (u_t, v_t) - \eta \nabla E(u_t, v_t)$$

Please start from $(u_0, v_0) = (0, 0)$, and fix $\eta = 0.01$, what is $E(u_5, v_5)$ after five updates? Please provide derivation steps.

5. Continued from the previous question, if we approximate the $E(u + \Delta u, v + \Delta v)$ by $\hat{E}_2(\Delta u, \Delta v)$, where $\hat{E}_2$ is the second-order Taylor's expansion of E around (u, v). Suppose

$$\hat{E}_2(\Delta u, \Delta v) = b_{uu}(\Delta u)^2 + b_{vv}(\Delta v)^2 + b_{uv}(\Delta u)(\Delta v) + b_u \Delta u + b_v \Delta v + b.$$

What are the values of $(b_{uu},\ b_{vv},\ b_{uv},\ b_u,\ b_v,\ b)$ when $(u, v) = (0, 0)$? Please provide derivation steps.

6. Continued from the previous question and denote the Hessian matrix to be $\nabla^2 E(u, v)$, and assume that the Hessian matrix is positive definite. What is the optimal $(\Delta u, \Delta v)$ to minimize $\hat{E}_2(\Delta u, \Delta v)$? The direction is called the *Newton Direction.* Please provide derivation steps.

7. Using the Newton direction (without η) to update, please start from $(u_0, v_0) = (0, 0)$, what is $E(u_5, v_5)$ after five updates? Please provide derivation steps. use 6. solve 4.

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 $\lambda > 0$.

8. 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.

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 $\tilde{\mathbf{y}} = [\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_K]^T$.

9. 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.

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

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

Prove your answer.

Experiments with Logistic Regression

11. (*) Implement the fixed learning rate gradient descent algorithm below for logistic regression, initialized with $\mathbf{0}$. Run the algorithm with $\eta = 0.001$ and $T = 2000$ on the following set for training:

`http://www.csie.ntu.edu.tw/~htlin/course/ml15fall/hw3/hw3_train.dat`

and the following set for testing:

`http://www.csie.ntu.edu.tw/~htlin/course/ml15fall/hw3/hw3_test.dat`

What is the weight vector within your g? What is the $E_{out}(g)$ from your algorithm, evaluated using the 0/1 error on the test set?

12. (*) Implement the fixed learning rate stochastic gradient descent algorithm below for logistic regression, initialized with $\mathbf{0}$. Instead of randomly choosing n in each iteration, please simply pick the example with the cyclic order $n = 1, 2, \ldots, N, 1, 2, \ldots$. Run the algorithm with $\eta = 0.001$ and $T = 2000$ on the following set for training:

`http://www.csie.ntu.edu.tw/~htlin/course/ml15fall/hw3/hw3_train.dat`

and the following set for testing:

`http://www.csie.ntu.edu.tw/~htlin/course/ml15fall/hw3/hw3_test.dat`

What is the weight vector within your g? What is the $E_{out}(g)$ from your algorithm, evaluated using the 0/1 error on the test set?

Experiment with Regularized Linear Regression and Validation

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

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

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 $\lambda = 1.126$, 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. (10 points) Continue from Question 9. Assume that we take the more general

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

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. (10 points) Continue from Question 9. 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?