Homework #3

RELEASE DATE: 11/02/2015

DUE DATE: **11/16/2015** (**MONDAY!!!**), 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.

Questions 1-2 are about *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.

2. Recall that we have introduced the hat matrix $\text{H} = \text{X}(\text{X}^T\text{X})^{-1}\text{X}^T$ in class, where $\text{X} \in \mathbb{R}^{N \times (d+1)}$. That is, there are d features. Assume X^TX is invertible, which statements of H are true? Please prove all the statements that you believe to be true.

[a] H is positive semi-definite.

[b] H is always invertible.

[c] Some eigenvalues of H are bigger than 1.

[d] $d + 1$ eigenvalues of H are 1.

[e] $\text{H}^{1126} = \text{H}$.

Questions 3-5 are about *error* and SGD

3. 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] $err(\mathbf{w}) = \big(\max(0, 1 - y\mathbf{w}^T\mathbf{x})\big)^2$
[c] $err(\mathbf{w}) = \max(0, -y\mathbf{w}^T\mathbf{x})$
[d] $err(\mathbf{w}) = \theta(-y\mathbf{w}^T\mathbf{x})$
[e] $err(\mathbf{w}) = \exp(-y\mathbf{w}^T\mathbf{x})$

4. Which of the following are differentiable functions of $\mathbf{w}$ everywhere? Explain your choices.

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

5. 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.

For Questions 6-10, you will play with gradient descent algorithm and variants

6. Consider a function

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

What is the gradient $\nabla E(u, v)$ around $(u, v) = (0, 0)$? Please provide derivation steps.

7. 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.

8. Continued from Question 7, 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.

9. Continued from Question 8 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.

10. 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.

For Questions 11-12, you will play with feature transforms

11. Consider six inputs $\mathbf{x}_1 = (1, 1)$, $\mathbf{x}_2 = (1, -1)$, $\mathbf{x}_3 = (-1, -1)$, $\mathbf{x}_4 = (-1, 1)$, $\mathbf{x}_5 = (0, 0)$, $\mathbf{x}_6 = (1, 0)$. What is the biggest subset of those input vectors that can be shattered by the union of quadratic, linear, or constant hypotheses of $\mathbf{x}$? Please show how the vectors can be shattered.

12. Assume that a transformer peeks the data and decides the following transform $\mathbf{\Phi}$ "intelligently" from the data of size N. The transform maps $\mathbf{x} \in \mathbb{R}^d$ to $\mathbf{z} \in \mathbb{R}^N$, where

$$(\mathbf{\Phi}(\mathbf{x}))_n = z_n = [\![\mathbf{x} = \mathbf{x}_n]\!]$$

Consider a learning algorithm that performs linear classification after the feature transform. That is, the algorithm effectively works on an $\mathcal{H}_{\mathbf{\Phi}}$ that includes *all* possible $\mathbf{\Phi}$. What is $d_{\text{VC}}(\mathcal{H}_{\mathbf{\Phi}})$ (i.e. the maximum number of points that can be shattered by the process above)? Please provide explanation of your answer.

For Questions 13-15, you will play with linear regression and feature transforms.

Consider the target function:

$$f(x_1, x_2) = \text{sign}(x_1^2 + x_2^2 - 0.6)$$

Generate a training set of $N = 1000$ points on $\mathcal{X} = [-1, 1] \times [-1, 1]$ with uniform probability of picking each $\mathbf{x} \in \mathcal{X}$. Generate simulated noise by flipping the sign of the output in a random 10% subset of the generated training set.

13. (*) Carry out Linear Regression without transformation, i.e., with feature vector:

$$(1, x_1, x_2),$$

to find $\mathbf{w}_{\text{lin}}$, and use $\mathbf{w}_{\text{lin}}$ directly for classification. Run the experiments for 1000 times and plot a histogram on the **classification** (0/1) in-sample error (E_{in}). What is the average E_{in} over 1000 experiments?

Now, transform the training data into the following nonlinear feature vector:

$$(1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$$

Find the vector $\tilde{\mathbf{w}}$ that corresponds to the solution of Linear Regression, and take it for classification.

14. (*) Run the experiment for 1000 times, and plot a histogram on $\tilde{w}_3$, the weight associated with x_1x_2. What is the average $\tilde{w}_3$?

15. (*) Continue from Question 14, and plot a histogram on the **classification** E_{out} instead. You can estimate it by generating a new set of 1000 points and adding noise as before. What is the average E_{out}?

For Questions 16-17, you will derive an algorithm for multinomial (multiclass) logistic regression.

For a K-class classification problem, we will denote the output space $\mathcal{Y} = \{1, 2, \cdots, K\}$. The hypotheses considered by MLR are indexed by a list of weight vectors $(\mathbf{w}_1, \cdots, \mathbf{w}_K)$, each weight vector of length $d + 1$. Each list represents a hypothesis

$$h_y(\mathbf{x}) = \frac{\exp(\mathbf{w}_y^T \mathbf{x})}{\sum_{i=1}^{K} \exp(\mathbf{w}_i^T \mathbf{x})}$$

that can be used to approximate the target distribution $P(y|\mathbf{x})$. MLR then seeks for the maximum likelihood solution over all such hypotheses.

16. For general K, derive an $E_{\text{in}}(\mathbf{w}_1, \cdots, \mathbf{w}_K)$ like page 11 of Lecture 10 slides by minimizing the negative log likelihood. Please provide derivation steps.

17. For the E_{in} derived above, its gradient ∇E_{in} can be represented by $\left(\frac{\partial E_{\text{in}}}{\partial \mathbf{w}_1}, \frac{\partial E_{\text{in}}}{\partial \mathbf{w}_2}, \cdots, \frac{\partial E_{\text{in}}}{\partial \mathbf{w}_K}\right)$, write down the derivation step and the final $\frac{\partial E_{\text{in}}}{\partial \mathbf{w}_i}$.

For Questions 18-20, you will play with logistic regression.

18. (*) 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?

19. (*) Implement the fixed learning rate gradient descent algorithm below for logistic regression, initialized with $\mathbf{0}$. Run the algorithm with $\eta = 0.01$ 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?

20. (*) 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?

Bonus: Smart 'Cheating'

21. (BBQ, 10 points) For a regression problem, the root-mean-square-error (RMSE) of a hypothesis h on a test set $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$ is defined as

$$\text{RMSE}(h) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - h(\mathbf{x}_n))^2}.$$

Please consider a case of knowing all the $\mathbf{x}_n$, none of the y_n, but allowed to query RMSE(h) for some h.

For any given hypothesis h, let

$$\begin{aligned}\mathbf{h} &= (h(x_1), h(x_2), \cdots, h(x_N)) \\ \mathbf{y} &= (y_1, y_2, \cdots, y_N).\end{aligned}$$

To compute $\mathbf{h}^T\mathbf{y}$, what is the least number of queries?

22. (BBQ, 10 points) Continuing from Question 21, for any given set of hypotheses $\{h_1, h_2, \cdots, h_K\}$. Let $H(\mathbf{x}) = \sum_{k=1}^K w_k h_k(\mathbf{x})$. To solve

$$\min_{w_1, w_2, \cdots, w_K} \text{RMSE}\left(H\right),$$

what is the least number of queries?