Homework #3

RELEASE DATE: 10/29/2014

DUE DATE: **11/17/2014** (**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.

There are two 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.*

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 40 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. (MCQ) 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 $d = 8$, which among the following choices is the smallest number of examples N that will result in an expected E_{in} greater than 0.008?

[**a**] 10

[**b**] 25

[**c**] 100

[**d**] 500

[**e**] 1000

(+ explanation of your choice)

2. (MRQ) Recall that we have introduced the hat matrix $\mathrm{H} = \mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T$ in class, where $\mathrm{X} \in \mathbb{R}^{N\times(d+1)}$. That is, there are d features. Assume $\mathrm{X}^T\mathrm{X}$ is invertible, which statements of H are 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**] $\mathrm{H}^{1126} = \mathrm{H}$.

(+ explanation of your choice)

Questions 3-5 are about *error* and SGD

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

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

(+ explanation of your choice)

4. (MRQ) Which of the following are differentiable functions of $\mathbf{w}$ everywhere?

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

(+ explanation of your choice)

5. (MCQ) When using SGD on the following error functions and 'ignoring' some singular points that are not differentiable, which of the following error function results in PLA?

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

(+ explanation of your choice)

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

6. (MCQ) 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 $(u, v) = (0, 0)$?

[**a**] $(-3, 1)$

[b] $(3, -1)$

[c] $(0, -2)$

[d] $(-2, 0)$

[e] none of the other choices

(+ explanation of your derivations)

7. (MCQ) 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?

[a] 4.904

[b] 3.277

[c] 2.825

[d] 1.436

[e] 0.365

(+ explanation of your derivations)

8. (MCQ) 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 $(u, v) = (0, 0)$?

[a] $(3, 8, -1, -2, 0, 3)$

[b] $(1.5, 4, -1, -2, 0, 3)$

[c] $(3, 8, -0.5, -1, -2, 0)$

[d] $(1.5, 4, -0.5, -1, -2, 0)$

[e] none of the other choices

(+ explanation of your derivations)

9. (MCQ) 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*.

[a] $-\left(\nabla^2 E(u, v)\right)^{-1} \nabla E(u, v)$

[b] $+\left(\nabla^2 E(u, v)\right)^{-1} \nabla E(u, v)$

[c] $-\nabla^2 E(u, v) \nabla E(u, v)$

[d] $+\nabla^2 E(u, v) \nabla E(u, v)$

[e] none of the other choices

(+ explanation of your choice)

10. (MCQ) 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?

[a] 4.532

[b] 3.046

[c] 2.361

[d] 1.279

[e] 0.356

(optional: + explanation of your derivations)

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

11. (MCQ) 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 $\mathbf{x}$?

[a] $\mathbf{x}_1$, $\mathbf{x}_3$

[b] $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$

[c] $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$

[d] $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$, $\mathbf{x}_5$

[e] $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$, $\mathbf{x}_5$, $\mathbf{x}_6$

(+ explanation of your choice)

12. (MCQ) 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)?

[a] 1

[b] $d + 1$

[c] $N + 1$

[d] $N + d + 1$

[e] ∞

(+ explanation of your choice)

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. (MCQ, *) 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. What is the closest value to the classification (0/1) in-sample error (E_{in})? Run the experiment 1000 times and take the average E_{in} in order to reduce variation in your results.

[a] 0.1

[b] 0.3

[c] 0.5

[d] 0.7

[e] 0.9

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. (MCQ, *) Which of the following hypotheses is closest to the one you find using Linear Regression on the transformed input? Closest here means agrees the most with your hypothesis (has the most probability of agreeing on a randomly selected point).

[a] $g(x_1, x_2) = \text{sign}(-1 - 0.05x_1 + 0.08x_2 + 0.13x_1x_2 + 1.5x_1^2 + 1.5x_2^2)$

[b] $g(x_1, x_2) = \text{sign}(-1 - 0.05x_1 + 0.08x_2 + 0.13x_1x_2 + 1.5x_1^2 + 15x_2^2)$

[c] $g(x_1, x_2) = \text{sign}(-1 - 0.05x_1 + 0.08x_2 + 0.13x_1x_2 + 15x_1^2 + 1.5x_2^2)$

[d] $g(x_1, x_2) = \text{sign}(-1 - 1.5x_1 + 0.08x_2 + 0.13x_1x_2 + 0.05x_1^2 + 0.05x_2^2)$

[e] $g(x_1, x_2) = \text{sign}(-1 - 1.5x_1 + 0.08x_2 + 0.13x_1x_2 + 0.05x_1^2 + 1.5x_2^2)$

15. (MCQ, *) What is the closest value to the classification out-of-sample error E_{out} of your hypothesis? Estimate it by generating a new set of 1000 points and adding noise as before. Average over 1000 runs to reduce the variation in your results.

[a] 0.1

[b] 0.3

[c] 0.5

[d] 0.7

[e] 0.9

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. (MCQ) 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.

[a] $\frac{1}{N}\sum_{n=1}^N \left(\sum_{i=1}^K \mathbf{w}_i^T\mathbf{x}_n - \mathbf{w}_{y_n}^T\mathbf{x}_n\right)$

[b] $\frac{1}{N}\sum_{n=1}^N \left(\sum_{i=1}^K \left(\mathbf{w}_i^T\mathbf{x}_n - \mathbf{w}_{y_n}^T\mathbf{x}_n\right)\right)$

[c] $\frac{1}{N}\sum_{n=1}^N \left(\ln\left(\sum_{i=1}^K \exp(\mathbf{w}_i^T\mathbf{x}_n) - \exp(\mathbf{w}_{y_n}^T\mathbf{x}_n)\right)\right)$

[d] $\frac{1}{N}\sum_{n=1}^N \left(\ln\left(\sum_{i=1}^K \exp(\mathbf{w}_i^T\mathbf{x}_n)\right) - \mathbf{w}_{y_n}^T\mathbf{x}_n\right)$

[e] none of the other choices

(+ derivation steps)

17. (MCQ) 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 $\frac{\partial E_{\text{in}}}{\partial \mathbf{w}_i}$.

[a] $\frac{1}{N}\sum_{n=1}^{N}\left(\sum_{i=1}^{K}\left(\exp(\mathbf{w}_i^T\mathbf{x}_n) - [\![y_n = i]\!]\right)\mathbf{x}_n\right)$

[b] $\frac{1}{N}\sum_{n=1}^{N}\left(\sum_{i=1}^{K}\left(\exp(\mathbf{w}_i^T\mathbf{x}_n) - 1\right)\mathbf{x}_n\right)$

[c] $\frac{1}{N}\sum_{n=1}^{N}\left(\left(h_i(\mathbf{x}_n) - [\![y_n = i]\!]\right)\mathbf{x}_n\right)$

[d] $\frac{1}{N}\sum_{n=1}^{N}\left(\left(h_i(\mathbf{x}_n) - 1\right)\mathbf{x}_n\right)$

[e] none of the other choices

(+ derivation steps)

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

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

`http://www.csie.ntu.edu.tw/~htlin/course/ml14fall/hw3/hw3_train.dat`

and the following set for testing:

`http://www.csie.ntu.edu.tw/~htlin/course/ml14fall/hw3/hw3_test.dat`

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

[a] 0.475

[b] 0.412

[c] 0.322

[d] 0.220

[e] 0.103

19. (MCQ, *) Implement the fixed learning rate gradient descent algorithm below for logistic regression. Run the algorithm with $\eta = 0.01$ and $T = 2000$ on the following set for training:

`http://www.csie.ntu.edu.tw/~htlin/course/ml14fall/hw3/hw3_train.dat`

and the following set for testing:

`http://www.csie.ntu.edu.tw/~htlin/course/ml14fall/hw3/hw3_test.dat`

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

[a] 0.475

[b] 0.412

[c] 0.322

[d] 0.220

[e] 0.103

20. (MCQ, *) Implement the fixed learning rate stochastic gradient descent algorithm below for logistic regression. 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/ml14fall/hw3/hw3_train.dat`

and the following set for testing:

`http://www.csie.ntu.edu.tw/~htlin/course/ml14fall/hw3/hw3_test.dat`

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

[a] 0.475

[b] 0.412

[c] 0.322

[d] 0.220

[e] 0.103

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

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

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. To construct a hypothesis g with RMSE(g) = 0, what is the least number of queries?

22. (BBQ, 10 points) Continued from Question 21, 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?

23. (BBQ, 20 points) Continued from Question 22, 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?

Answer guidelines. First, please write down your name and school ID number.

Name: School ID:

Then, fill in your answers for MCQ and MRQ 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.