Homework #4

TA email: `ml2011ta@csie.ntu.edu.tw`

RELEASE DATE: 11/07/2011

DUE DATE: 11/21/2011, BEFORE THE END OF CLASS

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 and predictions 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 with the common math notations introduced in class or in the problems. We do not accept solutions written in any other languages.

4.1 VC Dimension: Union and Intersection

(1) (5%) Do Problem 2.14(a) of LFD.

(2) (10%) Do Problem 2.14(b) of LFD.

(3) (10%) Do Problem 2.14(c) of LFD for the case of $K = 2$ **and** the assumption that the intersection of $\mathcal{H}_1$ and $\mathcal{H}_2$ is non-empty.

(4) (5%) Under the assumption that the intersection of all $\mathcal{H}_i$ is non-empty, do Problem 2.15(a) of LFD.

(5) (10%) Under the assumption that the intersection of all $\mathcal{H}_i$ is non-empty, prove Problem 2.15(b) of LFD.

(6) (10%) Under the assumption that the intersection of all $\mathcal{H}_i$ is non-empty, do Problem 2.15(c) of LFD.

(7) (10%) The bound in Problem 2.15(c) of LFD gives you $d_{\mathsf{VC}} \le 3 \log_2 d + 2$ for the decision stump learning model (in Homework 3.4) in Exercise 3.15. Assuming that you know the VC dimension of the one dimensional decision stump to be 2 (proved in class), do the last part of Exercise 3.15 of LFD and improve the upper bound to $d_{\mathsf{VC}} \le 2 \log_2 d + 2$.

4.2 The Expected In-Sample Error of Linear Regression

(1) (5%) Do Exercise 3.3(b) of LFD.

(2) (5%) Do Exercise 3.3(c) of LFD.

(3) (5%) Do Exercise 3.3(d) of LFD.

(4) (5%) Do Exercise 3.4(a) of LFD.

(5) (5%) Do Exercise 3.4(b) of LFD. What is the "matrix" that the hint talks about?

(6) (5%) Do Exercise 3.4(c) of LFD. The simpler equation (in terms of H and $\boldsymbol{\epsilon}$ you write down), the better.

(7) (5%) Do Exercise 3.4(d) of LFD.

4.3 The Feature Transforms

(1) (5%) Do Exercise 3.12(a) of LFD.

(2) (5%) Do Exercise 3.12(b) of LFD.

(3) (5%) Do Exercise 3.12(c) of LFD.

(4) (4%) Do Exercise 3.13(a) of LFD.

(5) (4%) Do Exercise 3.13(c) of LFD.

(6) (4%) Do Exercise 3.13(d) of LFD.

(7) (4%) Do Exercise 3.13(e) of LFD.

(8) (4%) Do Exercise 3.13(f) of LFD.

4.4 Least-squared Linear Regression (*)

Implement the least-squared linear regression algorithm in Section 3.2 of LFD to compute the optimal $(d+1)$-dimensional $\mathbf{w}$ that solves

$$\min_{\mathbf{w}} \sum_{n=1}^{N} \Big(y_n - \big(\mathbf{w}^T \mathbf{x}_n\big) \Big)^2 .$$

(1) (15%) Generate a training data set of size 100 as directed by Exercise 3.2 of LFD. Generate a test set of size 1000 of the same nature. Run the pocket algorithm (Homework 3.5) on the training set for $T = 1000$ to get $\mathbf{w}_{\text{pocket}}$. Run the linear regression algorithm to get $\mathbf{w}_{\text{lin}}$. Estimate the performance of the two weight vectors with the test set to get $E_{\text{test}}(\mathbf{w}_{\text{pocket}})$ and $E_{\text{test}}(\mathbf{w}_{\text{lin}})$, in terms of the 0/1 loss (classification). Repeat the experiment (with fresh data sets) 100 times and plot $E_{\text{test}}(\mathbf{w}_{\text{pocket}})$ versus $E_{\text{test}}(\mathbf{w}_{\text{lin}})$ as a scatter plot.

(2) (5%) Based on your findings in the previous problem, which algorithm would you recommend to your boss for this data set? Why?

Please check the course policy carefully and do not use sophisticated packages in your solution. You ***can*** *use standard matrix multiplication and inversion routines.*

4.5 Gradient Descent for Logistic Regression (*)

Consider the formulation

$$\min_{\mathbf{w}} \quad E(\mathbf{w}), \tag{A1}$$

$$\text{where} \quad E(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} E^{(n)}(\mathbf{w}), \text{ and } E^{(n)}(\mathbf{w}) = \ln\Bigg(1 + \exp\Big(-y_n\big(2\mathbf{w}^T\mathbf{x}_n\big)\Big)\Bigg).$$

(1) (10%) Do Problem 3.17(b) of LFD. See Exercise 3.6(c) for the definition of the cross-entropy error function. This shows that the formulation (A1) is equivalent to cross-entropy-minimizing logistic regression.

(2) (5%) Do Problem 3.17(c) of LFD.

(3) (5%) For a given $(\mathbf{x}_n, y_n)$, derive its gradient $\nabla E^{(n)}(\mathbf{w})$.

(4) (15%) Implement the fixed learning rate stochastic gradient descent algorithm for (A1).

(a) initialize a $(d+1)$-dimensional vector $\mathbf{w}^{(0)}$, say, $\mathbf{w}^{(0)} \longleftarrow (0, 0, \ldots, 0)$.

(b) for $t = 1, 2, \ldots, T$

- randomly pick one n from $\{1, 2, \ldots, N\}$.
- update

$$\mathbf{w}^{(t)} \longleftarrow \mathbf{w}^{(t-1)} - \eta \cdot \nabla E^{(n)}(\mathbf{w}^{(t-1)}).$$

Assume that

$$g_1^{(t)}(\mathbf{x}) = \text{sign}\left((\mathbf{w}^{(t)})^T \mathbf{x}\right),$$

where $\mathbf{w}^{(t)}$ are generated from stochastic gradient descent algorithm above. Run the algorithm with $\eta = 0.001$ and $T = 2000$ on the following set for training:

`http://www.csie.ntu.edu.tw/~htlin/course/ml11fall/data/hw4_train.dat`

and the following set for testing:

`http://www.csie.ntu.edu.tw/~htlin/course/ml11fall/data/hw4_test.dat`

Plot $E_{\text{in}}\left(g_1^{(t)}\right)$ and $E_{\text{test}}\left(g_1^{(t)}\right)$ as a function of t and briefly state your findings.

(5) (15%) Implement the fixed-learning-rate gradient descent algorithm below for (A1).

(a) initialize a $(d{+}1)$-dimensional vector $\mathbf{w}^{(0)}$, say, $\mathbf{w}^{(0)} \longleftarrow (0, 0, \ldots, 0)$.

(b) for $t = 1, 2, \ldots, T$

- update

$$\mathbf{w}^{(t)} \longleftarrow \mathbf{w}^{(t-1)} - \eta \cdot \nabla E(\mathbf{w}^{(t-1)}) \ .$$

Assume that

$$g_2^{(t)}(\mathbf{x}) = \text{sign}\left((\mathbf{w}^{(t)})^T \mathbf{x}\right),$$

where $\mathbf{w}^{(t)}$ are generated from gradient descent algorithm above. Run the algorithm with $\eta = 0.001$ and $T = 2000$ on the following set for training:

`http://www.csie.ntu.edu.tw/~htlin/course/ml11fall/data/hw4_train.dat`

and the following set for testing:

`http://www.csie.ntu.edu.tw/~htlin/course/ml11fall/data/hw4_test.dat`

Plot $E_{\text{in}}\left(g_2^{(t)}\right)$ and $E_{\text{test}}\left(g_2^{(t)}\right)$ as a function of t, compare it to your plot for $g_1^{(t)}$, and briefly state your findings.

4.6 Upper Bound of Union without Assumption

(1) (Bonus 10%) If the intersection of all $\mathcal{H}_i$ may be empty, show a counter-example of Problem 2.15(a) of LFD. Then, think of a way to fix the bound (while keeping it tight) and prove it.