Homework #2

TA in charge: Ming-Feng Tsai

RELEASE DATE: 10/09/2008

DUE DATE: 10/16/2008, 4:00 pm IN CLASS

TA SESSION: 10/15/2008, noon to 2:00 pm IN R106

Unless granted by the instructor in advance, you must turn in a hard 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.

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.

2.1 Probably Approximately Correct

Consider a learning model G of a finite size $L < \infty$. We showed in class that

$$\Pr\left(\exists\, g \in G : \left|\pi(g) - \nu(g)\right| \geq \varepsilon\right) \leq 2Le^{-2\varepsilon^2 N}.$$

(1) (20%) Let $\delta = 2Le^{-2\varepsilon^2 N}$. We can rewrite the inequality above as follows:

With probability $\geq 1-\delta$, for all $g \in G$,

$$\nu(g) - \epsilon(L, N, \delta) \leq \pi(g) \leq \nu(g) + \epsilon(L, N, \delta).$$

Here ϵ is a function of L, N, and δ. Please derive $\epsilon(L, N, \delta)$.

(2) (10%) Take $\delta = 0.01$ and $L = 1$, how many examples do we need to make $\epsilon(L, N, \delta) \leq 0.05$?

(3) (10%) Take $\delta = 0.01$ and $L = 100$, how many examples do we need to make $\epsilon(L, N, \delta) \leq 0.05$?

(4) (10%) Take $\delta = 0.01$ and $L = 10000$, how many examples do we need to make $\epsilon(L, N, \delta) \leq 0.05$?

The title of this problem, *Probably Approximately Correct*, states what we can interpret from the inequalities above. "Probably" means the statement is true with a high probability ($\geq 1-\delta$). "Approximately" means that every $\pi(g)$ is close to $\nu(g)$ (within ϵ). "Correct" means that we can guarantee $\pi(g)$ to be small (by getting some decision function g with small $\nu(g)$).

2.2 Gradient and Newton Directions

Consider a function

$$E(u, v) = e^u + e^{2v} + e^{uv} + u^2 - 3uv + 4v^2 - 3u - 5v,$$

(1) (10%) Approximate $E(u + \Delta u, v + \Delta v)$ by $\hat{E}_1(\Delta u, \Delta v)$, where $\hat{E}_1$ is the first-order Taylor's expansion of E around $(u, v) = (0, 0)$. Suppose $\hat{E}_1(\Delta u, \Delta v) = a_u \Delta u + a_v \Delta v + a$. What are the values of a_u, a_v, and a?

(2) (10%) Minimize $\hat{E}_1$ over all possible $(\Delta u, \Delta v)$ such that $\|(\Delta u, \Delta v)\| = 0.5$. Prove that the optimal column vector $\begin{bmatrix}\Delta u \\ \Delta v\end{bmatrix}$ is parallel to the column vector $-\nabla E(u, v)$, which is called the *negative gradient direction*.

(3) (10%) Approximate $E(u + \Delta u, v + \Delta v)$ by $\hat{E}_2(\Delta u, \Delta v)$, where $\hat{E}_2$ is the second-order Taylor's expansion of E around $(u, v) = (0, 0)$. 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, and b?

(4) (10%) Minimize $\hat{E}_2$ over all possible $(\Delta u, \Delta v)$ (regardless of length). Use the fact that $\nabla^2 E(u, v)$ (the Hessian matrix) is positive definite to prove that the optimal column vector

$$\begin{bmatrix}\Delta u \\ \Delta v\end{bmatrix} = -\left(\nabla^2 E(u, v)\right)^{-1} \nabla E(u, v),$$

which is called the *Newton direction*.

(5) (10%) Numerically compute the following values:

(a) the vector $(\Delta u, \Delta v)$ of length 0.5 along the negative gradient direction, and the resulting $E(u + \Delta u, v + \Delta v)$.

(b) the vector $(\Delta u, \Delta v)$ of length 0.5 along the Newton direction, and the resulting $E(u + \Delta u, v + \Delta v)$.

(c) the vector $(\Delta u, \Delta v)$ of length 0.5 that minimizes $E(u + \Delta u, v + \Delta v)$, and the resulting $E(u + \Delta u, v + \Delta v)$. (*Hint: let* $\Delta u = 0.5 \sin\theta$.)

Compare the values of $E(u + \Delta u, v + \Delta v)$ in (5a), (5b), and (5c). Briefly state your findings.

The negative gradient direction and the Newton direction are quite fundamental for designing optimization algorithms. We mentioned in class that ML algorithms can often be decomposed to an optimization algorithm and some ML-related criteria. Thus, it is important to understand these directions and put them in your toolbox.

2.3 Least-squares Linear Regression (*)

(1) (10%) Implement the least-squares linear regression algorithm taught in class to compute the optimal (w_0, θ_0) that solves

$$\min_{w,\theta} \sum_{n=1}^{N} \Big(y_n - \big(\langle w, x_n\rangle - \theta\big)\Big)^2.$$

Run the algorithm on the following set for training:

`http://www.csie.ntu.edu.tw/~htlin/course/ml08fall/data/hw2_3_1_train.dat`

and the following set for testing:

`http://www.csie.ntu.edu.tw/~htlin/course/ml08fall/data/hw2_3_1_test.dat`

Report the (w_0, θ_0) you find. Let $g_0(x) = \text{sign}\big(\langle w_0, x\rangle - \theta_0\big)$. What is $\nu(g_0)$? How about $\hat{\pi}(g_0)$?

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

(2) (15%) Run your algorithm in (1) on the following data set for training:

`http://www.csie.ntu.edu.tw/~htlin/course/ml08fall/data/hw2_3_2_train.dat`

and the following set for testing

`http://www.csie.ntu.edu.tw/~htlin/course/ml08fall/data/hw2_3_2_test.dat`

Again, report (w_0, θ_0), $\nu(g_0)$, and $\hat{\pi}(g_0)$. Also, plot the training examples (x_n, y_n) and the decision boundary $\langle w_0, x\rangle - \theta_0 = 0$ in the same figure. Use different symbols to distinguish examples with different y_n. Briefly state your findings.

(3) (10%) Consider an alternative formulation:

$$\min_{w,\theta} \frac{\lambda}{2}\langle w, w\rangle + \frac{\lambda}{2}\theta^2 + \frac{1}{2}\sum_{n=1}^{N}\Big(y_n - \big(\langle w, x_n\rangle - \theta\big)\Big)^2$$

Prove that the optimal solution for the problem above is

$$\big(X^T X + \lambda \mathcal{I}\big)^{-1} X^T Y,$$

where matrices X and Y are defined in class, and $\mathcal{I}$ is an identity matrix. The formulation is called *regularized least-squares linear regression*.

(4) (15%) Split the given training examples in

`http://www.csie.ntu.edu.tw/~htlin/course/ml08fall/data/hw2_3_2_train.dat`

to 120 "base" examples (the first 120) and 80 "validation" ones (the last 80).

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.*

Implement an algorithm that solves the formulation in (3). Run the algorithm on the 120 base examples using $\log_{10}\lambda = \{2, 1, 0, -1, \ldots, -8, -9, -10\}$. Let g_λ be the decision function returned from the algorithm when using parameter λ. Validate g_λ with the 80 validation examples and test it with the test examples in

`http://www.csie.ntu.edu.tw/~htlin/course/ml08fall/data/hw2_3_2_test.dat`

Plot $\nu_b(g_\lambda)$, $\nu_v(g_\lambda)$, $\hat{\pi}(g_\lambda)$ on the same figure as a function of $\log_{10}\lambda$, where the base training error

$$\nu_b(g) = \frac{1}{120}\sum_{(x_n,y_n)\in \text{ base}} I\Big[y_n \neq g(x_n)\Big]$$

and the validation error

$$\nu_v(g) = \frac{1}{80}\sum_{(x_n,y_n)\in \text{ validation}} I\Big[y_n \neq g(x_n)\Big].$$

Briefly state your findings.

2.4 Gradient Descent for Logistic Regression (*)

Consider the formulation (so-called *logistic regression*)

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

$$\text{where} \quad E(w,\theta) = \frac{1}{N}\sum_{n=1}^{N} E_n(w,\theta), \text{ and } E_n(w,\theta) = \log\left(1 + \exp\Big(-y_n\big(\langle w, x_n\rangle - \theta\big)\Big)\right).$$

(1) (10%) For a given (x_n, y_n), derive $\nabla E_n(w,\theta)$.

(2) (20%) Implement the (fixed-step) stochastic 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 $t = 1, 2, \ldots, T$

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

- update

$$\mathbf{w}^{(t)} \longleftarrow \mathbf{w}^{(t-1)} - \alpha \cdot \nabla E_n(w^{(t-1)}, \theta^{(t-1)}) \ ;$$

where $\mathbf{w}^{(t)}$ represents $\left(\theta^{(t)}, (w)_1^{(t)}, \ldots, w_d^{(t)}\right)$.

Assume that

$$g_1^{(t)}(x) = \text{sign}\left(\left\langle w^{(t)}, x \right\rangle - \theta^{(t)}\right),$$

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

`http://www.csie.ntu.edu.tw/~htlin/course/ml08fall/data/hw2_4_train.dat`

and the following set for testing:

`http://www.csie.ntu.edu.tw/~htlin/course/ml08fall/data/hw2_4_test.dat`

Plot $\nu\left(g_1^{(t)}\right)$ and $\hat{\pi}\left(g_1^{(t)}\right)$ as a function of t and briefly state your findings.

(3) (20%) Implement the (fixed-step) 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

$$\mathbf{w}^{(t)} \longleftarrow \mathbf{w}^{(t-1)} - \alpha \cdot \nabla E(w^{(t-1)}, \theta^{(t-1)}) \ ;$$

where $\mathbf{w}^{(t)}$ represents $\left(\theta^{(t)}, (w)_1^{(t)}, \ldots, w_d^{(t)}\right)$.

Assume that

$$g_2^{(t)}(x) = \text{sign}\left(\left\langle w^{(t)}, x \right\rangle - \theta^{(t)}\right),$$

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

`http://www.csie.ntu.edu.tw/~htlin/course/ml08fall/data/hw2_4_train.dat`

and the following set for testing:

`http://www.csie.ntu.edu.tw/~htlin/course/ml08fall/data/hw2_4_test.dat`

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