Homework #4

TA in charge: Ming-Feng Tsai

RELEASE DATE: 11/06/2008

DUE DATE: 11/20/2008, 4:00 pm IN CLASS

TA SESSION: 11/19/2008, 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.

4.1 More about VC Dimension

In this problem, $VCD(G)$ stands for the VC dimension of G.

(1) (10%) Assume $L = |G| < \infty$. Prove that $VCD(G) \le \log_2 L$.

(2) (20%) Assume $VCD(G_1) = D_1$ and $VCD(G_2) = D_2$. Prove that

$$0 \le VCD(G_1 \cap G_2) \le \min(D_1, D_2).$$

(3) (20%) Assume $VCD(G_1) = D_1$ and $VCD(G_2) = D_2$. Prove that

$$\max(D_1, D_2) \le VCD(G_1 \cup G_2) \le D_1 + D_2 + 1.$$

4.2 Curse of Dimensionality

(1) (20%) Consider a hypercube $[-0.5, 0.5]^d$ and a hypersphere of radius 0.1 in $\mathbb{R}^d$. Compute the ratio between the volume of the sphere and the volume of the cube for $d = 1, 2, \cdots, 10$.
(*Note: be careful about numerical accuracy.*)

(2) (15%) If the training examples are generated uniformly at random within $[-0.5, 0.5]^d$, how many examples must be included in a training set to ensure that with probability $\ge \frac{1}{2}$, the origin's nearest neighbor in the training set is within distance 0.1? Compute this number for $d = 1, 2, \ldots, 10$.

(3) (15%) Repeat Problem 4.2 (2) using $(0.5, 0.5, \cdots, 0.5)$ instead of the origin as the reference (test) point. Compare the numbers between Problem 4.2 (2) and Problem 4.2 (3) and briefly describe your findings.

4.3 Experiment with Radial Basis Function Network (*)

Implement the RBFN algorithm taught in class to obtain a decision function of the form

$$g(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m \cdot \phi(\mu_m, x) \right).$$

Given a positive integer M, a positive number σ, and the training set $\Big\{(x_n, y_n)\Big\}_{n=1}^{N}$, the algorithm returns g by the following steps:

- Apply K-means clustering (Lloyd's algorithm) with $K = M$ to find the centers μ_m.
- Compute the influences using Gaussian basis functions $\phi(\mu_m, x) = \exp\left(\frac{-\|x-\mu_m\|^2}{2\sigma^2}\right)$ with the given σ.
- Calculate the "optimal" α_m by generalized linear regression.

Run the algorithm on the following set for training:

`http://www.csie.ntu.edu.tw/~htlin/course/ml08fall/data/hw4_train.dat`

and the following set for testing:

`http://www.csie.ntu.edu.tw/~htlin/course/ml08fall/data/hw4_test.dat`

Assume that $g_{M,\sigma}$ is the decision function obtained using M and σ as the parameters. Try $M = 3, 6, 9$ and $\sigma = 0.1, 0.5, 2.5$.

(1) (30%) For each (M, σ) combination, plot the training points on the plane. Then, mark the centers μ_m found and plot the boundary between the $+1$ and -1 regions indicated by $g_{M,\sigma}$. Compare the figures and briefly describe your findings.

(2) (20%) For each (M, σ) combination, report $\nu(g_{M,\sigma})$ and $\hat{\pi}(g_{M,\sigma})$. Briefly describe your findings.

(*Note: You need to implement your own K-means algorithm and your own generalized linear regression algorithm instead of using sophisticated packages*)

4.4 Experiment with Backprop Neural Network (*)

Implement the backpropagation algorithm for d-M-1 neural network with tanh-type neurons. That is, each neuron implements the function

$$x^{\text{out}} = \tanh\left(\sum_{i=0}^{d^{\text{in}}} w_i \cdot x_i^{\text{in}}\right).$$

Given two integers (M, T) and the training set $\Big\{(x_n, y_n)\Big\}_{n=1}^{N}$, the algorithm returns a trained d-M-1 neural network g, which can be represented by a long vector $\mathbf{w}$ of size $\Big(M \cdot (d+1) + 1 \cdot (M+1)\Big)$. In particular, the algorithm performs stochastic gradient descent for T iterations on

$$\begin{aligned} \min_{\mathbf{w}} \quad & E(\mathbf{w}) \\ \text{where} \quad & E(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} E_n(\mathbf{w}), \\ & E_n(\mathbf{w}) = \big(y_n - \text{NN}(x_n, \mathbf{w})\big)^2. \end{aligned}$$

Here $\text{NN}(\cdot, \mathbf{w})$ "simulates" the d-M-1 neural network represented by $\mathbf{w}$. The pseudo-code of the algorithm is as follows:

```
• randomly initialize the weights $\mathbf{w}^{(0)}$
• for $t = 1, 2, \ldots, T$
    – randomly pick one $n$ from $\{1, 2, \ldots, N\}$.
    – update
          $\mathbf{w}^{(t)} \longleftarrow \mathbf{w}^{(t-1)} - \eta \cdot \nabla E_n\big(\mathbf{w}^{(t-1)}\big)$.
```

Run the algorithm on the following set for training:

`http://www.csie.ntu.edu.tw/~htlin/course/ml08fall/data/hw4_train.dat`

and the following set for testing:

`http://www.csie.ntu.edu.tw/~htlin/course/ml08fall/data/hw4_test.dat`

Try $T = 50000$ and the combinations of the following parameters:

- the number of hidden neurons $M = 1$, 3, or 5
- the elements of $\mathbf{w}^{(0)}$ chosen independently and uniformly from the range $(-0.02, 0.02)$, the range $(-0.2, 0.2)$, or the range $(-2, 2)$
- the learning rate $\eta = 0.01$, 0.1, or 1

(1) (50%) Let

$$g^{(t)}(x) \quad = \quad \text{sign}\Big(\text{NN}\big(x_n, \mathbf{w}^{(t)}\big)\Big).$$

For each of the 27 combinations above, plot $E\big(\mathbf{w}^{(t)}\big)$, $\nu\big(g^{(t)}\big)$, and $\hat{\pi}\big(g^{(t)}\big)$ as functions of t on the same figure. Compare the curve of $\hat{\pi}$ to the curves of E and ν for each combination, and briefly state your findings.

(*Note: You should have* 27 *separate plots with scales, legends, labels, and captions clearly marked. You need to implement your own backpropagation algorithm instead of using sophisticated packages*)