Homework #8

RELEASE DATE: 1/1/2016 (Happy New Year)

DUE DATE: 1/20/2016, BEFORE NOON (IN R217)

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.

Neural Network and Deep Learning

1. A fully connected Neural Network has $L = 2$; $d^{(0)} = A$, $d^{(1)} = B$, $d^{(2)} = 1$. If only products of the form $w_{ij}^{(\ell)} x_i^{(\ell-1)}$, $w_{ij}^{(\ell+1)} \delta_j^{(\ell+1)}$, and $x_i^{(\ell-1)} \delta_j^{(\ell)}$ count as operations (even for $x_0^{(\ell-1)} = 1$), without counting anything else, what is the total number of operations required in a single iteration of backpropagation (using SGD on one data point), in terms of A and B? List your calculation steps.

2. Consider a Neural Network with $d^{(0)} + 1 = 10$ input units (the constant $x_0^{(0)}$ is counted here as a unit), one output unit, and 36 hidden units (each $x_0^{(\ell)}$ is also counted as a unit). The hidden units can be arranged in any number of layers $\ell = 1, \cdots, L-1$, and each layer is fully connected to the layer above it. What is the minimum possible number of weights that such a network can have? Explain your answer.

3. Following Question 2, what is the maximum possible number of weights that such a network can have? Explain your answer.

Autoencoder

4. Assume an autoencoder with $\tilde{d} = 1$. That is, the $d \times \tilde{d}$ weight matrix W becomes a $d \times 1$ weight vector $\mathbf{w}$, and the linear autoencoder tries to minimize

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T \mathbf{x}_n\|^2.$$

We can solve this problem with stochastic gradient decent by defining

$$\text{err}_n(\mathbf{w}) = \|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T \mathbf{x}_n\|^2$$

and calculate $\nabla_{\mathbf{w}} \text{err}_n(\mathbf{w})$. What is $\nabla_{\mathbf{w}} \text{err}_n(\mathbf{w})$? List your derivation steps.

5. Following Question 4, assume that noise vectors $\boldsymbol{\epsilon}_n$ are generated i.i.d. from a zero-mean, unit variance Gaussian distribution and added to $\mathbf{x}_n$ to make $\tilde{\mathbf{x}}_n = \mathbf{x}_n + \boldsymbol{\epsilon}_n$, a noisy version of $\mathbf{x}_n$. Then, the linear denoising autoencoder tries to minimize

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T(\mathbf{x}_n + \boldsymbol{\epsilon}_n)\|^2.$$

For any fixed $\mathbf{w}$, $\mathcal{E}\left(E_{\text{in}}(\mathbf{w})\right) = \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|^2 + \Omega(\mathbf{w})$. What is $\Omega(\mathbf{w})$? List your derivation steps.

Nearest Neighbor and RBF Network

6. Consider getting the 1 Nearest Neighbor hypothesis from a data set of two examples $(\mathbf{x}_+, +1)$ and $(\mathbf{x}_-, -1)$. The resulting hypothesis would actually be linear. What is the linear hypothesis $g_{\text{LIN}}(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$ (where $\mathbf{w}$ does not include $b = w_0$) in terms of $\mathbf{x}_+$ and $\mathbf{x}_-$? List your derivation steps.

7. Consider an RBF Network hypothesis for binary classification

$$g_{\text{RBFNET}}(\mathbf{x}) = \text{sign}\left(\beta_+ \exp(-\|\mathbf{x} - \boldsymbol{\mu}_+\|^2) + \beta_- \exp(-\|\mathbf{x} - \boldsymbol{\mu}_-\|^2)\right)$$

and assume that $\beta_+ > 0 > \beta_-$. The resulting hypothesis would actually be linear. What is the linear hypothesis $g_{\text{LIN}}(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$ (where $\mathbf{w}$ does not include $b = w_0$)? List your derivation steps.

8. Assume that a full RBF network (page 9 of class 214) using $\text{RBF}(\mathbf{x}, \boldsymbol{\mu}) = [\![\mathbf{x} = \boldsymbol{\mu}]\!]$ is solved for squared error regression on a data set where all inputs $\mathbf{x}_n$ are different. What are the optimal coefficients β_n for each $\text{RBF}(\mathbf{x}, \mathbf{x}_n)$ in terms of $(\mathbf{x}_n, y_n)$? List your derivation steps.

Matrix Factorization

9. Consider matrix factorization of $\tilde{d} = 1$ with alternating least squares. Assume that the $\tilde{d} \times N$ user factor matrix V is initialized to a constant matrix of 1. After step 2.1 of alternating least squares (page 10 of lecture 215), prove that the optimal w_m, the $\tilde{d} \times 1$ movie 'vector' for the m-th movie, is the average rating of the m-th movie.

10. Assume that for a full rating matrix R, we have obtained a perfect matrix factorization $\mathrm{R} = \mathrm{V}^T\mathrm{W}$. That is, $r_{nm} = \mathbf{v}_n^T\mathbf{w}_m$ for all n, m. Then, a new user $(N + 1)$ comes. Because we do not have any information for the type of the movie she likes, we initialize her feature vector $\mathbf{v}_{N+1}$ to $\frac{1}{N}\sum_{n=1}^{N}\mathbf{v}_n$, the average user feature vector. Now, our system decides to recommend her a movie m with the maximum predicted score $\mathbf{v}_{N+1}^T\mathbf{w}_m$. Prove that the movie would be the one with the largest average rating.

Experiment with k Nearest Neighbor

Implement any algorithm that 'returns' the k Nearest Neighbor hypothesis discussed in page 8 of lecture 214.

$$g_{k\text{-nbor}}(\mathbf{x}) = \text{sign}\left(\sum_{m:\ k \text{ closest examples to } \mathbf{x}} y_m\right)$$

Evaluate with the 0/1 error. Run the algorithm on the following set for training:

`http://www.csie.ntu.edu.tw/~htlin/course/ml15fall/hw8/hw8_train.dat`

and the following set for testing:

`http://www.csie.ntu.edu.tw/~htlin/course/ml15fall/hw8/hw8_test.dat`

11. (see next question :-))

12. (*) Plot $E_{\text{in}}(g_{k\text{-nbor}})$ for $k = 1, 3, 5, 7, 9$. Briefly describe your findings.

13. (see next question :-))

14. (*) Plot $E_{\text{out}}(g_{k\text{-nbor}})$ for $k = 1, 3, 5, 7, 9$. Briefly describe your findings.

15. (see next question :-))

16. (*) Implement g_{uniform} on page 8 of lecture 214. Plot $E_{\text{in}}(g_{\text{uniform}})$ for $\gamma = 0.001, 0.1, 1, 10, 100$. Briefly describe your findings.

17. (see next question :-))

18. (*) Plot $E_{\text{out}}(g_{\text{uniform}})$ for $\gamma = 0.001, 0.1, 1, 10, 100$. Briefly describe your findings.

Experiment with k-Means

Implement the k-Means algorithm (page 16 of lecture 214). Randomly select k instances from $\{\mathbf{x}_n\}$ to initialize your $\boldsymbol{\mu}_m$. Run the algorithm on the following set for training:

`http://www.csie.ntu.edu.tw/~htlin/course/ml15fall/hw8/hw8_nolabel_train.dat`

and repeat the experiment for 500 times. Calculate the clustering E_{in} by

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M}[[\mathbf{x}_n \in S_m]]\|\mathbf{x}_n - \boldsymbol{\mu}_m\|^2$$

as described on page 13 of lecture 214 for $M = k$.

19. (see next question :-))

20. (*) Plot the average E_{in} over 500 experiments for $k = 2, 4, 6, 8, 10$. Briefly describe your findings.

Bonus: VC Dimension of Neural Networks

21. (10%) Prove that for $\Delta \geq 2$, if $N \geq 3\Delta \log_2 \Delta$, $N^{\Delta} + 1 < 2^N$.

22. (10%) Consider a hypothesis set $\mathcal{H}_{3A}$ that consists of all d-3-1 neural networks with $\text{sign}(\cdot)$ as all the transformation functions, and with $(w_0, w_1, w_2, w_3) = (-2.5, +1, +1, +1)$ for the output neuron ***only***. Use the facts above (or not) to prove that the VC dimension of $\mathcal{H}_{3A}$ is less than

$$3 \cdot \big(3(d+1)+1\big)\log_2\big(3(d+1)+1\big).$$