QUIZ4

1. **Neural Network and Deep Learning**
A fully connected Neural Network has $L = 2$; $d^{(0)} = 5$, $d^{(1)} = 3$, $d^{(2)} = 1$. 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, which of the following is the total number of operations required in a single iteration of backpropagation (using SGD on one data point)?
    - A. 47
    - B. 43
    - C. 53
    - D. 59
    - E. none of the other choices

2. Consider a Neural Network without any bias terms $x_0^{(\ell)}$. Assume that the network contains $d^{(0)} = 10$ input units, 1 output unit, and 36 hidden units. 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?
    - A. 46
    - B. 44
    - C. none of the other choices
    - D. 43
    - E. 45

3. Following Question 2, what is the maximum possible number of weights that such a network can have?
    - A. 510
    - B. 520
    - C. none of the other choices
    - D. 500
    - E. 490

4. **Autoencoder**
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

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

We can solve this problem with stochastic gradient descent 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 $\nabla_{\mathbf{w}} \text{err}_n(\mathbf{w})$?

A. $(4\mathbf{x}_n - 4)(\mathbf{w}^T\mathbf{w})$
B. none of the other choices
C. $(4\mathbf{w} - 4)(\mathbf{x}_n^T\mathbf{x}_n)$
D. $2(\mathbf{x}_n^T\mathbf{w})^2\mathbf{w} + 2(\mathbf{x}_n^T\mathbf{w})(\mathbf{w}^T\mathbf{w})$
E. $2(\mathbf{x}_n^T\mathbf{w})^2\mathbf{x}_n + 2(\mathbf{x}_n^T\mathbf{w})(\mathbf{w}^T\mathbf{w})\mathbf{w} - 4(\mathbf{x}_n^T\mathbf{w})\mathbf{w}$

5. Following Question 4, assume that noise vectors $\boldsymbol{\epsilon}_n$ are generated i.i.d. from a zero-mean, unit var iance 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

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

For any fixed $\mathbf{w}$, what is $\mathcal{E}\left(E_{in}(\mathbf{w})\right)$, where the expectation $\mathcal{E}$ is taken over the noise generation process?

A. $\frac{1}{N}\sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|^2$
B. $\frac{1}{N}\sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|^2 + d\mathbf{w}^T\mathbf{w}$
C. none of the other choices
D. $\frac{1}{N}\sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|^2 + \mathbf{w}^T\mathbf{w}$
E. $\frac{1}{N}\sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{w}\mathbf{w}^T\mathbf{x}_n\|^2 + \frac{1}{d}\mathbf{w}^T\mathbf{w}$

6. **Nearest Neighbor and RBF Network**
Consider getting the 1 Nearest Neighbor hypothesis from a data set of two examples $(\mathbf{x}_+, +1)$ and $(\mathbf{x}_-, -1)$. Which of the following linear hypothesis $g_{LIN}(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$ (where $\mathbf{w}$ does not include $b = w_0$) is equivalent to the hypothesis?

A. none of the other choices
B. $\mathbf{w} = 2(\mathbf{x}_+ - \mathbf{x}_-)$, $b = +\mathbf{x}_+^T\mathbf{x}_-$
C. $\mathbf{w} = 2(\mathbf{x}_- - \mathbf{x}_+)$, $b = +\|\mathbf{x}_+\|^2 - \|\mathbf{x}_-\|^2$
D. $\mathbf{w} = 2(\mathbf{x}_- - \mathbf{x}_+)$, $b = -\mathbf{x}_+^T\mathbf{x}_-$
E. $\mathbf{w} = 2(\mathbf{x}_+ - \mathbf{x}_-)$, $b = -\|\mathbf{x}_+\|^2 + \|\mathbf{x}_-\|^2$

7. Consider an RBF Network hypothesis for binary classification

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

and assume that $\beta_+ > 0 > \beta_-$. Which of the following linear hypothesis $g_{LIN}(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$ (where $\mathbf{w}$ does not include $b = w_0$) is equivalent to $g_{RBFNET}(\mathbf{x})$?

A. $\mathbf{w} = 2(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)$, $b = \ln\left|\frac{\beta_+}{\beta_-}\right| - \|\boldsymbol{\mu}_+\|^2 + \|\boldsymbol{\mu}_-\|^2$
B. $\mathbf{w} = 2(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)$, $b = \ln\left|\frac{\beta_-}{\beta_+}\right| + \|\boldsymbol{\mu}_+\|^2 - \|\boldsymbol{\mu}_-\|^2$
C. $\mathbf{w} = 2(\beta_+\boldsymbol{\mu}_+ + \beta_-\boldsymbol{\mu}_-)$, $b = -\beta_+\|\boldsymbol{\mu}_+\|^2 + \beta_-\|\boldsymbol{\mu}_-\|^2$
D. $\mathbf{w} = 2(\beta_+\boldsymbol{\mu}_+ + \beta_-\boldsymbol{\mu}_-)$, $b = +\beta_+\|\boldsymbol{\mu}_+\|^2 - \beta_-\|\boldsymbol{\mu}_-\|^2$
E. none of the other choices

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 $\text{RBF}(\mathbf{x}, \mathbf{x}_n)$?

A. y_n

B. $\|\mathbf{x}_n\|^2 y_n^2$

C. none of the other choices

D. $\|\mathbf{x}_n\| y_n$

E. y_n^2

9. **Matrix Factorization**
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), what is the optimal w_m, the $\tilde{d} \times 1$ movie 'vector' for the m-th movie?

A. the average rating of the m-th movie

B. the total rating of the m-th movie

C. the maximum rating of the m-th movie

D. the minimum rating of the m-th movie

E. none of the other choices

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 $\mathbf{v}_{N+1}^T \mathbf{w}_m$. What would the movie be?

A. the movie with the largest maximum rating

B. none of the other choices

C. the movie with the smallest rating variance

D. the movie with the largest minimum rating

E. the movie with the largest average rating

11. **Experiment with Backprop neural Network**

Implement the backpropagation algorithm (page 16 of lecture 212) for d-M-1 neural network with tanh-type neurons, **including the output neuron**. Use the squared error measure between the output $g_{NNET}(\mathbf{x}_n)$ and the desired y_n and backprop to calculate the per-example gradient. Because of the different output neuron, your $\delta_1^{(L)}$ would be different from the course slides! Run the algorithm on the following set for training (each row represents a pair of $(\mathbf{x}_n, y_n)$; the first column is $(\mathbf{x}_n)_1$; the second one is $(\mathbf{x}_n)_2$; the third one is y_n):
hw4_nnet_train.dat
and the following set for testing:
hw4_nnet_test.dat

Fix $T = 50000$ and consider the combinations of the following parameters:

- the number of hidden neurons M
- the elements of $w_{ij}^{(\ell)}$ chosen independently and uniformly from the range $(-r, r)$
- the learning rate η

Fix $\eta = 0.1$ and $r = 0.1$. Then, consider $M \in \{1, 6, 11, 16, 21\}$ and repeat the experiment for 500 times. Which M results in the lowest average E_{out} over 500 experiments?

A. 11

B. 16

C. 1

D. 21

E. 6

12. Following Question 11, fix $\eta = 0.1$ and $M = 3$. Then, consider $r \in \{0, 0.001, 0.1, 10, 1000\}$ and repeat the experiment for 500 times. Which r results in the lowest average E_{out} over 500 experiments?

    A. 0

    B. 0.1

    C. 0.001

    D. 10

    E. 1000

13. Following Question 11, fix $r = 0.1$ and $M = 3$. Then, consider $\eta \in \{0.001, 0.01, 0.1, 1, 10\}$ and repeat the experiment for 500 times. Which η results in the lowest average E_{out} over 500 experiments?

    A. 0.01

    B. 0.001

    C. 10

    D. 0.1

    E. 1

14. Following Question 11, deepen your algorithm by making it capable of training a d-8-3-1 neural network with tanh-type neurons. Do not use any pre-training. Let $r = 0.1$ and $\eta = 0.01$ and repeat the experiment for 500 times. Which of the following is true about E_{out} over 500 experiments?

    A. $0.02 \leq E_{out} < 0.04$

    B. none of the other choices

    C. $0.04 \leq E_{out} < 0.06$

    D. $0.06 \leq E_{out} < 0.08$

    E. $0.00 \leq E_{out} < 0.02$

15. **Experiment with 1 Nearest Neighbor**
    Implement any algorithm that 'returns' the 1 Nearest Neighbor hypothesis discussed in page 8 of lecture 214.

    $$g_{\text{nbor}}(\mathbf{x}) = y_m \text{ such that } \mathbf{x} \text{ closest to } \mathbf{x}_m$$

    Run the algorithm on the following set for training:
    hw4_knn_train.dat
    and the following set for testing:
    hw4_knn_test.dat
    Which of the following is closest to $E_{in}(g_{\text{nbor}})$?

    A. 0.2

    B. 0.3

    C. 0.0

    D. 0.1

    E. 0.4

16. Following Question 15, which of the following is closest to $E_{out}(g_{\text{nbor}})$?

    A. 0.30

    B. 0.28

    C. 0.34

D. 0.32

E. 0.26

17. Now, implement any algorithm for the k Nearest Neighbor with $k = 5$ to get $g_{\text{5-nbor}}(\mathbf{x})$. Run the algorithm on the same sets in Question 15 for training/testing. Which of the following is closest to $E_{in}(g_{\text{5-nbor}})$?

A. 0.1

B. 0.2

C. 0.3

D. 0.4

E. 0.0

18. Following Question 17, Which of the following is closest to $E_{out}(g_{\text{5-nbor}})$

A. 0.28

B. 0.26

C. 0.34

D. 0.32

E. 0.30

19. **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:
hw4_kmeans_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 $M = k$.
For $k = 2$, which of the following is closest to the average E_{in} of k-Means over 500 experiments?

A. 0.5

B. 1.0

C. 2.5

D. 1.5

E. 2.0

20. For $k = 10$, which of the following is closest to the average E_{in} of k-Means over 500 experiments?

A. 1.0

B. 1.5

C. 2.0

D. 0.5

E. 2.5