Midterm Practice Solutions

Introduction to Deep Learning (University of California Riverside)

EE 228 - Spring 2021 Midterm Practice Exam

Remarks:

- Any consultation with others will lead to a failing grade in the course and will be reported for disciplinary action. If you are caught copying someone else's solution, you will FAIL.
- If you have any questions, you can ask the instructor via zoom during the exam. For further questions, send an email to oymak@ece.ucr.edu AND ysatt001@ucr.edu.
- TOTAL: 27 points + 3 Bonus points.
- GOOD LUCK!

Problems

Problem 1. Pick the valid differences between deep learning and traditional machine learning.

1. ✓Deep learning aims to learn representations from data.
2. Deep learning works better for small datasets.
3. ✓Deep leaning makes use of neural networks.
4. Deep learning is typically more computationally efficient.

Problem 2. Pick the contributing factors to the rise of deep learning during the last decade.

1. Applications in financial sector such as algorithmic trading
2. ✓Availability of large datasets
3. Invention of convolutional networks
4. ✓State-of-the-art performance in computer vision and natural language processing

Problem 3. What are the advantages of stochastic gradient descent over using full gradient?

1. SGD reduces the randomness in the optimization process
2. ✓Full gradient may lead to redundant calculations due to similarities in the dataset.
3. ✓SGD is storage efficient.
4. SGD normalizes the data and speeds up the optimization.

Problem 4. Suppose you have a dataset $X = [x_1 \ \dots \ x_n]^T$ and $y = [y_1 \ \dots \ y_n]^T$. Consider the quadratic loss $L(\theta) = 10\|y - X\theta\|_{\ell_2}^2$. What is the gradient of $L(\theta)$?

1. $10X^T(X\theta - y)$
2. $\frac{10}{n}X^T(X\theta - y)$
3. ✓$20X^T(X\theta - y)$
4. $20X^T(y - X\theta)$

Problem 5. Let (x, y) be a training example where $x = [1 \ 1 \ 1]$ and $y = 1$. Consider the loss function $L(\theta) = (y - ReLU(x^T\theta))^2$. What is the gradient of $L(\theta)$ evaluated at $\theta = [-1 \ 1 \ -1]$?

1. $[1 \ 1 \ 1]$
2. $[-1 \ -1 \ -1]$
3. ✓$[0 \ 0 \ 0]$
4. $[-1 \ 1 \ -1]$

Problem 6. Let (x, y) be a training example where $x = [-1 \ 1 \ 1]$ and $y = -1$. Consider the loss function $L(\theta) = 0.5(y - ReLU(x^T\theta))^2$. What is the gradient of $L(\theta)$ evaluated at $\theta = [-1 \ 1 \ -1]$?

1. $[-1 \ 1 \ 1]$
2. $[0 \ 0 \ 0]$

3. ✓$[-2\ 2\ 2]$
4. $[-2\ 2\ -2]$

Problem 7. Let (x, y) be a training example. Consider the loss function $L(\theta) = \frac{1}{2}(y - \sigma(x^T\theta))^2$ where σ is the logistic function $\sigma(x) = \frac{1}{1+e^{-x}}$. What is the gradient of $L(\theta)$?

1. $(1 - \sigma(x^T\theta))(\sigma(x^T\theta) - y)x$
2. $\sigma(x^T\theta)(\sigma(x^T\theta) - y)x$
3. ✓$\sigma(x^T\theta)(1 - \sigma(x^T\theta))(\sigma(x^T\theta) - y)x$
4. $\sigma(x^T\theta)^2(\sigma(x^T\theta) - y)x$

Problem 8. Let (x, y) be a training example where $x = [-1\ 1\ 1]$ and $y = -1$. Consider the loss function $L(\theta) = (y - \sigma(x^T\theta))^2$ where σ is the logistic function. What is the gradient of $L(\theta)$ evaluated at $\theta = [-1\ 1\ -2]$?

1. $(1 + \exp(1/2))[-1\ 1\ 1]$
2. ✓$0.75[-1\ 1\ 1]$
3. $0.25[-1\ 1\ 2]$
4. $0.5(1 + \exp(1/2))[-1\ 1\ -2]$

Problem 9. Suppose you have a dataset $X = [x_1\ \dots\ x_n]^T$ and $y = [y_1\ \dots\ y_n]^T$. Consider the ridge regularized problem $L(\theta) = \|y - X\theta\|_{\ell_2}^2 + \lambda\|\theta\|_{\ell_2}^2$. What is the gradient of $L(\theta)$?

1. ✓$2(X^TX + \lambda I)\theta - 2X^Ty$
2. $(X^TX + \lambda I)\theta - X^Ty$
3. $2(X^TX + \lambda I)\theta$
4. $(X^TX - \lambda I)\theta + X^Ty$

Problem 10. Suppose you have a dataset $X = [x_1\ \dots\ x_n]^T$ and $y = [y_1\ \dots\ y_n]^T$. Consider the ridge regularized problem $L(\theta) = \|y - X\theta\|_{\ell_2}^2 + \lambda\|\theta\|_{\ell_2}^2$. What is the minimizer of $L(\theta)$?

Answer: $(X^TX + \lambda I)^{-1}X^Ty$

Problem 11. Suppose you have a fully-connected network with three hidden layers. Input dimension is 200, and the dimensions of the first, second, and third hidden layers are 200, 100, and 50 respectively. The output layer is 10 dimensional. What is the total number of parameters in this network? (Please assume that there are no bias variables).

1. 76800
2. ✓65500
3. 65000
4. 25500

Problem 12. Suppose you have a fully-connected network with three hidden layers. Input dimension is 200, and the dimensions of the first, second, and third hidden layers are 200, 100, and 50 respectively. The output layer is 10 dimensional. Given 200 dimensional input vector, what is the total number of floating point multiplication operations carried out by this network to obtain the output? Hint: There are 4 matrix-vector multiplications. Ignore the operations involving activations.

1. 360
2. 560
3. ✓65500
4. 10255000

Problem 13. Suppose input tensor to a convolutional layer is 200x200x3 dimensional. The layer consists of 80 5x5 filters. The filters are applied with stride 1 and without zero padding. What is the output dimension?

1. 198x198x3
2. ✓196x196x80
3. 200x200x80
4. 200x200x5

Problem 14. Suppose input tensor to a convolutional layer is 200x200x3 dimensional. The layer consists of 80 5x5 filters. The filters are applied with stride 1 and without zero padding. What is the total number of weights in this layer (assume no bias terms)?

1. 400
2. 2000
3. ✓6000
4. 1200

Problem 15. Suppose input tensor to a convolutional layer is 100x100x50 dimensional. The convolutional layer consists of 7x7 filters and uses stride 1. We wish to ensure that output dimension is same as the input dimension. What are the number of filters and the amount of zero-padding respectively? Note: Zero-padding of 1 increases the dimension by two e.g. 4x4 input becomes 6x6.

1. 100, 6
2. ✓50, 3
3. 100, 0
4. 50, 2

Problem 16. Select the key advantages of convolutional networks over fully connected networks?

1. ✓weight sharing
2. simpler backpropagation formulas
3. ✓translation invariance
4. ✓fewer and local connections between neurons

Problem 17. Which of the following statements are correct about data augmentation?

1. Data augmentation typically speeds up the optimization process.
2. Data augmentation is only useful for supervised learning.
3. Data augmentation is a norm-based regularization strategy.

4. ✓Data augmentation typically increases the test accuracy.

Problem 18. If your goal is training a sparse model (i.e. model with few nonzero weights), which regularization technique you might want to use?

1. Weight decay
2. ✓ℓ_1 regularization
3. Data augmentation
4. Max-norm regularization

Problem 19. True or False: Early-stopping regularization chooses the model with the highest test accuracy.
Answer: False

Problem 20. Suppose you are running SGD with batch size 32. You train the model with 200 epochs which corresponds to 50000 minibatch SGD iterations. What is the dataset size?

1. 960
2. 640
3. ✓8000
4. 64000

Problem 21. Consider a multiclass classification problem with 3 classes (labels are 1, 2 or 3). Given an input, suppose the softmax output of the network is $\hat{y} = [0.9\ 0.05\ 0.05]$ and the label is $y = 1$. What is the cross-entropy loss (up to three decimal digits)?
Answer: $-\ln(0.9) = 0.105$

Problem 22. Suppose input tensor to a convolutional layer is 100x100x3 dimensional. The convolutional layer consists of 50 7x7 filters and uses stride 5 and zero-padding 1. What is the output dimension? Note: Zero-padding of 1 increases the dimension by two e.g. 4x4 becomes 6x6.

1. 20x20x7
2. ✓20x20x50
3. 14x14x50
4. 19x19x50

Problem 23. Suppose input tensor to a two layer CNN is 102x102x3 dimensional. The first convolutional layer consists of 50 7x7 filters and uses stride 5. The second layer uses 50 3x3 filters with stride 1. What is the total number of parameters (in these two layers)?

1. 7350
2. 25000
3. ✓29850
4. 16750

Problem 24. Consider the fully connected network $f_{v,W}(x) = v^T \mathrm{ReLU}(Wx)$. Suppose

$$W = \begin{bmatrix} 1 & 3 & 1 \\ 1 & -1 & 2 \end{bmatrix} \quad \text{and} \quad v = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Define the loss function $L(v, W) = (y - f_{v,W}(x))^2$ on the training example $(x = [1\ -1\ 1],\ y = -3)$. What is the partial derivative $\partial L(v, W)/\partial v$?

1. $[0, 8]$

2. $[2, \ -2, 2]$

3. $[0, -6]$

4. ✓$[0, -8]$

Answer is from: 2*(-4+3)*[0 4]=[0 -8]

Problem 25. Consider the following chain of relations

$$
\begin{aligned}
f &= x_1 - x_2 \\
x_1 &= 2x_3 - 3x_4 \\
x_2 &= x_3 - 2x_5 \\
x_3 &= x_4 - x_5
\end{aligned}
$$

What is $\partial f / \partial x_5$?

1. -1

2. 0

3. ✓1

4. 2

Answer is: 1=-2+1+2

Problem 26. Consider the following chain of relations

$$
\begin{aligned}
f &= x_1 + x_2 \\
x_1 &= \text{ReLU}(2x_3 + 2x_4) \\
x_2 &= \text{ReLU}(x_3 - 2x_4) \\
x_3 &= x_4 - 1/2.
\end{aligned}
$$

What is $\partial f / \partial x_4$ evaluated at $x_4 = 1$?

1. 0

2. 2

3. ✓4

4. 6

Answer is: 4. Reason: $x_2 = 0$ and $f = x_1 = 4x_4 - 1$.

Problem 27. Consider the fully connected network $f(x) = v^T \text{ReLU}(W_2 \text{ReLU}(W_1 x))$. Suppose

$$W_1 = \begin{bmatrix} 2 & 1 \\ 1 & -2 \end{bmatrix} \quad W_2 = \begin{bmatrix} 2 & 1 \\ 1 & -2 \end{bmatrix} \quad \text{and} \quad v = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

What is the output evaluated at $x = [1, 1]$?

1. -3

2. 0

3. ✓3

4. 6

Problem 28. Consider a network $f(x) = v^T \text{max-pool}(x)$. Here x is a 7x1 vector, max-pool operation has receptive field 3 and stride 1 and v is a 5x1 vector. What is the output of the network at $x = [1, 2, 3, 4, 3, 2, 1]$ and $v = [1, 1, 0, 1, 1]$?

1. 6
2. 10
3. 18
4. ✓14

Answer is: 3+4+0+4+3=14.
Problem 29. Consider a network $f(x) = v^T \text{convolve}(x, f)$. Here x is a 5x1 vector, f is a 3x1 filter and v is a 3x1 vector. What is the output of the network at $x = [1, 2, 3, 2, 1]$ and $v = [1, 2, -1], f = [0, 1, 2]$? Hint: Don't forget to flip f.

1. 6
2. 8
3. ✓10
4. 12

Answer is: 4+2x7-8=10.
Problem 30. True or False: I will work hard on my EE-260 project!