

Student details: Please check that this PDF is yours.

- Student name: Papernot, Nicolas
- Student ID: 000000000

Honor code. Due to the circumstances, this exam is being conducted remotely. The goal of this midterm is for you to evaluate how much of the course content you understand. For this reason, I ask that you:

1. Work on the exam alone
2. Not use your lecture notes (i.e., this is a closed book exam)
3. Not use external resources (e.g., the Internet, books, etc.)

Instructions. At 4pm Eastern time, you should stop working on the midterm and immediately send your file(s) to the course instructor at nicolas.papernot@utoronto.ca. No extensions will be granted: we will not grade midterms received in my mailbox after 4.10pm (the 10mins are here to give you time to send the file and for me to receive it). Any format will be accepted (e.g., PNG, JPEG, PDF). Any filename will be accepted. To hand in your exam, you can do any of the following:



- Print this PDF, write your solutions in the boxes, scan or take photos of them
- Write your solutions on blank sheets of paper, scan or take photos of them

Exam structure. Each question in the exam is independent: if you can't find the solution to one question, skip it and come back to it later.

Questions This midterm contains 27 questions worth a total of 53 points.

1. (1 point) What is the difference between supervised and unsupervised learning?

2. (2 points) Give the name of two activation functions used in neural networks, their analytical expression as a function of an input x , and their output value if they are applied to the input $x = 0.3$.

3. (2 points) Assume that we are training a regression model with the mean squared error loss. Compute the loss value of a model predicting $(0.7, 0.6, 0.4)$ on a batch of 3 examples whose labels are $(0.8, 0.5, 0.2)$. Justify your answer by writing down the expression of the loss.

4. (2 points) Let f be a multivariate function taking two scalar inputs x_0 and x_1 and outputting a single scalar, i.e., $f : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$. We would like to compute the partial derivative of f with respect to x_0 . We are given the numerical value of $f(x_0, x_1)$ for two pairs of inputs: $f(0.11, 0.0) = 0.8$ and $f(0.09, 0.0) = 0.9$. Write down the analytical expression for an approximation of the partial derivative of f with respect to x_0 and compute its numerical value for $x_0 = 0.1$.

5. (2 points) We are training a model with vanilla gradient descent (i.e., the most basic version without refinements we saw first in the lectures). Write down analytically the update rule for a weight w as a function of (1) the learning rate α and (2) a partial derivative (1 point). Introduce any notation you use (e.g., to write down the partial derivative). Assume the value of the weight was $w = 0.2$ and the value of the partial derivative was 0.3. Pick a value for the learning rate (you don't need to justify it) and report the value of the weight after one step of gradient descent (1 point).

6. (1 point) We trained a polynomial regression model using a linear regression and a feature mapping $\phi : x_0, x_1 \mapsto 0.3 \cdot x_0^2 + 0.7 \cdot x_1$. Given that the weight of our model is $w = 0.6$, what is the prediction of our model on $(x_0, x_1) = (0.9, 0.6)^T$? Justify your answer.



7. (1 point) What is the difference between a model parameter and a hyperparameter?

8. (3 points) Describe briefly what each set should be used for: training set, validation set, test set.

9. (3 points) Consider the boolean tasks “AND”, “OR”, and “XOR”. Pick one task that can be learned with a linear model, and one task that cannot. For each of the two tasks you picked, draw a visualization of its input domain that illustrates why it can or cannot be learned with a linear model.


10. (3 points) We are using the perceptron algorithm, on data either labeled as $t = -1$ (negative class) or $t = 1$ (positive class). Write down the update rule for a training example $(x^{(i)}, t^{(i)})$. Assume the perceptron’s weight is $w = 0.4$ and the perceptron is given a training example $(0.3, 1)$. What is the value of w after one update?

11. (1 point) What condition does the training data need to satisfy for the perceptron algorithm to converge in a finite number of steps?

12. (1 point) Why is the 0-1 loss difficult to train with gradient descent?

13. (2 points) The mean squared error loss penalizes a model for being overconfident on a correct prediction. Why is using a logistic activation function in conjunction with the mean squared error loss also not a good approach to training a model by gradient descent? Give an example of a loss function that would be more appropriate to use with a logistic activation function.

14. (4 points) We are given a classifier for a task with 3 classes. When given an input, the model predicts the following scores: (0.1, 9.1, 2.2). What would be the probability assigned by the model on this input, for each class, if we used a softmax to normalize the scores into probabilities? Describe a practical problem that can arise when training a model with a softmax and a cross-entropy function. How did we address this problem in one of the assignments?



15. (1 point) What is the value of a dual variable associated with a training point that is not a support vector in the Lagrangian function corresponding to a hard-margin SVM?

16. (1 point) What component of a neural network gives it more expressive power than a linear model?

17. (2 points) How do we name a point corresponding to a value of w that is not a minimum but still satisfies $\frac{\partial L}{\partial w} = 0$ where w is the weight vector of our neural network and L is the loss function we use during training? What is a simple countermeasure to adopt when encountering such a value of w during training?

18. (1 point) What could cause us to encounter a plateau during training?

19. (2 points) What happens when the learning rate is too large? When the learning is too small?

20. (2 points) Describe briefly the difference between batch gradient descent, stochastic gradient descent, and minibatch stochastic gradient descent.

21. (2 points) If a dataset contains 46610 training examples, how many steps of minibatch stochastic gradient descent will be needed to complete 4 epochs with minibatches of 10 examples? Justify your answer.

22. (3 points) What are the two key aspects that distinguish convolutional layers from fully-connected layers? Which of these two distinguishes convolutional layers from locally connected layers as well?

23. (4 points) Compute the numerical values of A convolved with B where $A = \begin{pmatrix} 0 & -3 & 4 \\ -1 & -2 & 5 \\ -3 & -3 & 0 \end{pmatrix}$ and $B = \begin{pmatrix} 4 & 5 \\ 3 & -2 \end{pmatrix}$. That is compute $A \star B$ per the lecture notation.

24. (1 point) What are the three components included in a typical convolutional block?

25. (1 point) Does maxpooling provide invariance to small or large translations of images in the input domain?

26. (2 points) Write down the analytical expression and numerical value for the number of (a) weights and (b) connections in a fully-connected layer with $M = 18$ inputs and $N = 21$ outputs. Write down the analytical expression and numerical value for the number of (a) weights and (b) connections in a 2D convolutional layer with $I = 8$ input channels and $J = 9$ output channels, a filter size of $K = 5$, and an output width of $W = 50$ and height of $H = 44$.

27. (3 points) Define bias and define variance (1 point). Indicate whether underfitting translates to [low or high] bias and [low or high] variance (1 point). Indicate whether overfitting translates to [low or high] bias and [low or high] variance (1 point).