Final exam COMS 474/574

Saturday Nov. 21, 9AM to Sunday Nov. 26 11:59PM (Submission will close automatically by due time. No late submissions. Do not wait until the last minute to submit in case of system glitches.)

Submit your answers as a PDF to Canvas.

The exam is open note and open book. Links to precompiled PDFs for the course are on the landing page of the course on Github. Other help materials (click links) include: Discussion thread for final exam on Canvas, Review video for HW5, Notes during HW5 review

You cannot talk with anyone about how to solve the problems. You can however, ask the instructor on Canvas or during Zoom help sessions (Zoom meeting ID: 928 4811 4250 and password: 459503) for clarifications:

- Help session 1: Sunday, Nov 22, 8-9pm central time
- Help session 2: Wednesday, Nov 25, 3-4pm central time
- Let the instructor know if the times doesn't work for you.

By default, **you must show steps**. You'll get no point on a problem if you only provide a final answer. However, you do not need to show details for matrix operations. For example, you don't have to show how every element in a matrix multiplication is obtained.

This document will be updated to reflect feedback from students.

Last updated: 2020/11/21 at 16:51:33

Pledge

Please put (copy-and-paste, or hand write) this statement on top of your answers. It's better if you add your signature there – but it is not necessary. Please replace "YOUR_FIRST_NAME" and "YOUR_LAST_NAME" with your first and last names.

"I affirm that the work on this exam is my own and I will not use any people to help me nor will I share any part of this exam or my work with others without permission of the instructor. – YOUR_FIRST_NAME YOUR_LAST_NAME"

Regular problems (13 points)

1. [1pt] What is the Hadamard product $A \circ B$ between the following two matrixes?

   $A = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/3 & 1/2 & 1 \end{pmatrix}$

   $B = \begin{pmatrix} 0.5 & 1 & 6 \\ 3 & -4 & 2 \end{pmatrix}$

   Hint: HW5, Problem 1.

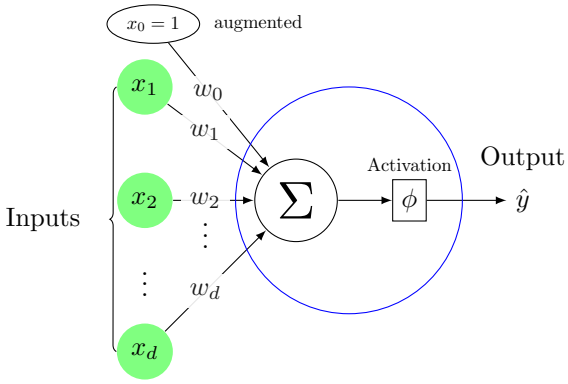
2. [2pt] Continuing from Problem 1 above, what is the product AB^T? The operator $\cdot^T$ means matrix transpose. What is the product BA^T? The operator $\cdot^T$ means matrix transpose.

   Hint: HW5, Problem 2.

3. [1pt] Continuing from Problems 1 and 2, above, given $f(x) = x + 1$, what is the value of $f(AB^T)$? Note that $f(AB^T)$ means applying f to every element of AB^T.

   Hint: HW5, Problem 4.

4. [2pt] Here is a diagram of a neuron.



Suppose $d = 3$. If the augmented input vector $\mathbf{x} = [x_0, x_1, x_2, x_3]^T = [1/2, 1/3, 1/4, 1/5]^T$, and the weight vector $\mathbf{w} = [w_0, w_1, w_2, w_3]^T = [2, 3, 4, 5]^T$, and the activation function $\phi(x) = x^2$ (note that in function notation, the x in $\phi(x)$ here can be any number or vector. not to be confused with the input vector $\mathbf{x}$), what is the value of the prediction $\hat{y} = \phi(\mathbf{w}^T\mathbf{x})$?

Hint: HW5, Problem 6.

5. [3pt] Continuing from Problem 4 above, if the loss is defined as $E = y + \hat{y}$,
   a. what is the value of $\frac{\partial E}{\partial \hat{y}}$?
   b. what is the value of $\frac{\partial \hat{y}}{\partial \mathbf{w}^T\mathbf{x}}$?
   c. what is the value of $\frac{\partial \mathbf{w}^T\mathbf{x}}{\partial x_1}$?
   d. what is the value of $\frac{\partial E}{\partial x_1} = \frac{\partial E}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{w}^T\mathbf{x}} \frac{\partial \mathbf{w}^T\mathbf{x}}{\partial x_1}$?

   Hint: HW5, Problem 7.

6. [4pt] Continuing from Problems 4 and 5 above, what is the value of $\frac{\partial E}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial E}{\partial x_0} \\ \frac{\partial E}{\partial x_1} \\ \frac{\partial E}{\partial x_2} \\ \frac{\partial E}{\partial x_3} \end{pmatrix}$? And what is the value of $\frac{\partial E}{\partial \mathbf{w}} = \begin{pmatrix} \frac{\partial E}{\partial w_0} \\ \frac{\partial E}{\partial w_1} \\ \frac{\partial E}{\partial w_2} \\ \frac{\partial E}{\partial w_3} \end{pmatrix}$?

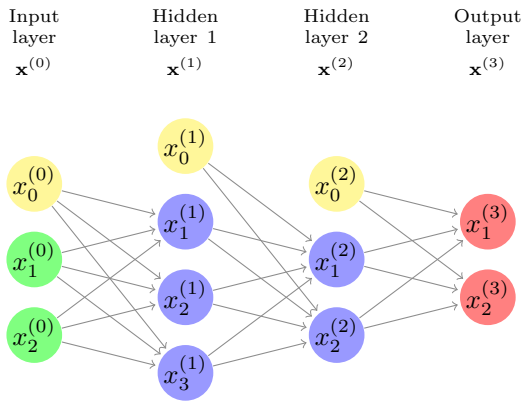
   Your answers should be two column real-valued vectors.

   Hint: HW5, Problem 8.

Bonus problems [11 points]

Hint: The slides "Recap:..." and "A grounded example..."

7. [2pt] Here is a neural network.



Denote $\mathbb{W}^{(l)}$ as the transfer matrix from layer l to layer $l+1$, for all $l \in [0..2]$. Suppose that $\mathbb{W}^{(0)} = \begin{pmatrix} 1 & 1 & 1 \\ -1 & -1 & -1 \\ 0.1 & 0.1 & 0.1 \end{pmatrix}$, $\mathbb{W}^{(1)} = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \\ 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$, $\mathbb{W}^{(2)} = \begin{pmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \\ 0.25 & 0.25 \end{pmatrix}$, all biases are 1, i.e., $x_0^{(l)} = 1, \forall l \in [0..2]$, and the **activation function is rectified linear unit (ReLU)** $\phi(x) = \max(x, 0)$.

If the input vector $\mathbf{x}^{(0)} = [1, 1, 1]^T$, what are the values of all activations $\mathbf{x}^{(l)}$ for all $l \in [1..3]$? Express activations at any layer l as a column vector.

8. [4pt] Continuing from Problem 7 (activation is ReLU) above, if the loss is squared error $E = (\hat{\mathbf{y}} - \mathbf{y})^2$, what is $\boldsymbol{\delta}^{(3)}$? Then express $\boldsymbol{\delta}^{(l-1)}$ in terms of $\boldsymbol{\delta}^{(l)}$ for layers $l \in \{0, 1, 2\}$. Note that $\boldsymbol{\delta}^{(\boldsymbol{l})} = \frac{\partial E}{\partial \mathbb{W}^{(l-1)} \mathbf{x}^{(l-1)}}$ is the derivative of loss over the pre-activation weighted sum at any layer l. Also note that there is no bias term for the output layer whose layer index is 3. Analytical result is sufficient and that's why $\mathbf{y}$ is not provided on purpose. All variables and functions used in your answer must be defined.

9. [1pt] In supervised learning, given the same training data, normally will a model WITHOUT regularization do better than a model WITH regularization **on the training data**? Why? Suppose the two models share the same hyperparameters except those related to regularization.

10. [1pt] In binary classification problems, a balanced dataset means that there are equal number of samples of both classes. If a dataset is unbalanced (the opposite case), will plain accuracy (i.e., percentage of samples predicted correctly) still be a good metric to gauge the performance of a model? Why?

11. [3pt] What is the time complexity of k-means clustering? Using the big-O notation, express it in terms of k, the number of clusters, and m, the number of samples. Show your proof.

THE END!