Midterm Version C

CS540

July 8, 2019

1 Instruction

1. Each incorrect answer receives -0.25, each correct answer receives 1, blank answers receives 0.
2. Check to make sure your name and (numerical) student ID (if you have it) is on the (Scantron) answer sheet. Also write your Wisc email ID on the answer sheet.
3. Check to make sure you completed question 41 and 42.
4. If you think none (or more than one) of the answers are correct, choose the best (closest) one.
5. Please submit this midterm, the answer sheet, the formula sheet, and all your additional notes when you finish.
6. Good luck!

2 Questions

41. Calculator?
    - A: Yes.
    - B: No.
42. Number of pages of additional notes? Please submit them at the end of the exam.
    - A: 0
    - B: 1
    - C: 2
    - D: 3
    - E: 4 or more.

3 Questions

1. Consider a linear threshold perceptron without the bias term $\hat{y}_i = a_i = \mathbb{1}_{\{w^T x_i \geqslant 0\}}$ with initial weights $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$. Given a new input $x_i = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $y_i = 0$. Let the learning rate be 1, what is the updated weight w_1 after one iteration of the perceptron algorithm?

- 0

2. Continue from the previous question, what is the updated weight w_2?

- -2

3. Let $C(w) = \text{tr}\left(ww^T\right), w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$. Here, $\text{tr}(A) = a + d$ for $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is the sum of diagonal entries of a matrix. What is the Hessian matrix of C at $w = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$?

- $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

4. Continue from the previous question, how many of the following vectors are eigenvectors of Hessian matrix?

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

- 6

5. Consider a linear model without bias term $a_i = wx_i$ with the hinge cost function $\max\{0, 1 - a_i y_i\}$. The initial weight is $w = 0$. What is the updated weight after one stochastic (sub)gradient descent step for w if the chosen training data is $x_1 = 1, y_1 = 1$? The learning rate is $\alpha = 1$.

- 1

6. Continue from the previous question, what if the chosen training data is $x_1 = -1, y_1 = 1$? Everything else is the same.

- -1

7. Let the hypothesis space (set of possible functions to choose from) be $\mathcal{H} = \left\{f : f(x) = \mathbb{1}_{\{x>b\}}, \text{ for some } b \in \mathbb{Z}\right\}$. This means $\hat{y}_i = 1$ if $x_i > b$ and $\hat{y}_i = 0$ if $x_i \leqslant b$ for some integer b. Given the following training set, what is C?

$$\begin{aligned} C &= \min_{\hat{f} \in \mathcal{H}} \sum_{i=1}^{5} \mathbb{1}_{\{\hat{f}(x_i) \neq y_i\}} \\ &= \min_{b} \sum_{i=1}^{5} \mathbb{1}_{\{\hat{y}_i \neq y_i\}} \end{aligned}$$

x_i	1	2	3	4	5
y_i	0	0	0	1	1

- 0

8. Continue from the previous question. What if the training set is changed to the following?

x_i	1	2	3	4	5
y_i	1	1	0	0	0

- 2

9. Continue from the previous question. What if the training set is changed to the following?

x_i	1	2	3	4	5
y_i	1	0	0	0	1

- 1

10. Suppose ax^2 is convex, how many of the following functions must be (weakly) convex?

$$a\left|x\right|, ax^3, ax^4, ax^{-1}, ax^{-2}$$

- 2, first and third

11. Given a fully connected neural network with inputs being flattened $3x3$ image pixel intensities. The network has one hidden layer and a single unit in the last (output) layer for binary classification. There are 90 weights (not including bias terms) that are updated during training. How many hidden units in the hidden layer are there?

- 9

12. Continue from the previous question. How many biases are updated during training?

- 10

13. Given the following weights, which of following logical operators does it represent?

$$\begin{aligned} w_{11}^{(1)} &= +2, w_{12}^{(1)} = -4, b_1^{(1)} = -1 \\ w_{21}^{(1)} &= -2, w_{22}^{(1)} = +4, b_2^{(1)} = -2 \\ w_{11}^{(2)} &= -2, w_{21}^{(2)} = -2, b_1^{(2)} = +1 \end{aligned}$$

The activation functions are LTU, $\mathbb{1}_{\{w^T x+b \geqslant 0\}}$ for all units. The notation $w_{ij}^{(l)}$ represents the weight in layer l from unit i in the previous layer to unit j in the next layer.

x_1	x_2	XOR	NOR	XNOR	$\Rightarrow$	$\Leftarrow$
1	1	0	0	1	1	1
1	0	1	0	0	0	1
0	1	1	0	0	1	0
0	0	0	1	1	1	1

- XNOR

14. Given the following training data. What is 2 fold cross validation accuracy (percentage of correct classificationn) if 1 nearest neighbor classifier with Manhattan distance is used? The first fold is the first two data points.

x_{i1}	1	2	3	5
x_{i2}	5	3	2	1
y	1	0	0	1

- 50 percent

15. Continue from the previous question. What is 4 fold cross validation accuracy (percentage of correct classificationn)?

    - 50 percent

16. What is w that minimizes $w_1 w_2$ subject to the constraint that $|w_1| + |w_2| = 1$?

    - $\begin{bmatrix} -\frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$ or $\begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}$

17. What is w that minimizes $w_1 w_2$ subject to the constraint that $w_1^2 + w_2^2 = 1$?

    - $\begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ or $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$

18. Find the weights w for the support vector machine classifier $\mathbb{1}_{\{w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i3} + 1 \geqslant 0\}}$ given the following training data.

x_1	x_2	x_3	y
0	0	0	1
0	0	2	0

    - $w = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}$

19. Continue from the previous question. What if the labels are flipped.

x_1	x_2	x_3	y
0	0	0	0
0	0	2	1

    - Impossible.

20. Consider a filter discribed by $F_{t,t'} = tt', t = -1, 0, 1, t' = -1, 0, 1$. What is the convolution between the following matrix and this filter? Use zero padding, i.e. set nonexistent values to 0 around the edges of the first matrix.

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} * F$$

    - $\begin{bmatrix} 4 & -3 \\ -2 & 1 \end{bmatrix}$

21. Continue from the previous question. What is the convolution between the filter in the previous question and itself?

$$F * F$$

- $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

22. Given the following x and y gradient of a 3×3 image, what is the gradient magnitude for the center pixel?

$$\nabla_x = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & -1 \\ -1 & -1 & -1 \end{bmatrix}, \nabla_y = \begin{bmatrix} 1 & -1 & 1 \\ -1 & 2 & 1 \\ -1 & 1 & -1 \end{bmatrix}$$

   - $2\sqrt{2}$

23. Continue from the previous question. Suppose 3×3 cells (1×1 blocks) without normalization are used for the previous image, what is the HOG feature descriptor with 2 bins? The first bin contains gradients with directions from 0 to $\frac{\pi}{2}$ and $-\frac{\pi}{2}$ to -π. (Different from original HOG paper: do not split a single gradient magnitude into two bins.)

   - $\begin{bmatrix} 4\sqrt{2} + 2\sqrt{2} & 4\sqrt{2} \end{bmatrix}$

24. Given the following counts according to labels in the training set, how many instances are used to train One vs One support vector machine for class 0 vs class 1.

y_i	count
0	10
1	20
2	30
3	40

   - 30

25. Continue from the previous question. How many instances are used to train One vs All support vector machine for class 0?

   - 100

26. What is the accuracy (on the training set) of the decision tree that first splits on x_2 trained on the following training set?

$$\left(x_{i1}, x_{i2}, y_i\right)_{i=1...4} = \{(0,1,0), (0,1,0), (1,0,0), (1,0,1)\}$$

   - 75 percent

27. Continue from the previous question, what is accuracy if the decision tree is first split on $x_{1?}$

   - 75 percent

28. Continue from the previous question, what is the conditional entropy of Y given X_1 ?

$$H\left(Y|X_1\right)$$

- $\frac{1}{2}$

29. Given two instances $x_1 = 1, x_2 = -1$, suppose the feature map for support vector machine is $\phi(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$, what is the kernel (Gram) matrix?

    - $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

30. Continue from the previous question. What is the kernel matrix if the feature map is $\phi(x) = \begin{bmatrix} x \\ x^2 \\ x^3 \end{bmatrix}$, everything else the same.

    - $\begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}$

31. Three documents A, B and C. The conditional probabilities of observing word H are $\mathbb{P}\{H|A\} = 0.1, \mathbb{P}\{H|B\} = 0.2, \mathbb{P}\{H|C\} = 0.3$. One document is chosen at random (each document with equal probability) and one word is chosen at random according to the conditional probabilities. What is the probability that the word is H?

    - 0.2

32. Continue from the previous question, given the chosen word is H, what is the probability that the document is A?

    - $\frac{1}{6}$

33. Given the counts, find the maximum likelihood estimate of $\mathbb{P}\{A = 1|B + C = 2\}$ without smoothing.

A	B	C	count
0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	4
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	2

    - $\frac{1}{3}$

34. Continue from the previous question. Find the maximum likelihood estimate of $\mathbb{P}\{A = 1|B + C = 1\}$ without smoothing.

    - 1

35. Suppose A is the common cause of B and C. All variables are binary. What is $\mathbb{P}\{B = 1, C = 1\}$?

$$\mathbb{P}\{A = 1\} = 0.1, \mathbb{P}\{A = 0\} = 0.9$$
$$\mathbb{P}\{B = 1|A = 1\} = 0.2, \mathbb{P}\{B = 1|A = 0\} = 0.4$$
$$\mathbb{P}\{C = 1|A = 1\} = 0.3, \mathbb{P}\{C = 1|A = 0\} = 0.5$$

- $0.1 \cdot 0.2 \cdot 0.3 + 0.9 \cdot 0.4 \cdot 0.5$

36. Continue from the previous question. Suppose the answer to the previous question is p, what is $\mathbb{P}\{C = 1|B = 1\}$?

- $\dfrac{p}{0.1 \cdot 0.2 + 0.9 \cdot 0.4}$

37. Continue from the previous question, what is $\mathbb{P}\{B = 1|C = 1\}$?

- $\dfrac{p}{0.1 \cdot 0.3 + 0.9 \cdot 0.5}$

38. Given the following transition matrix for a bigram model with characters "a" "b" "c". Two (uniform) random numbers between 0 and 1 are generated to simulate the characters after "a", say $u_1 = 0.5, u_2 = 0.5$. Using the CDF inversion method, which two characters are generated? For example, row "b" column "c" is $\mathbb{P}\{"c"|"b"\}$ the probability that the next character is "c" given the current one is "b".

–	a	b	c
a	0.1	0.2	0.7
b	0.2	0.3	0.5
c	0.3	0.4	0.3

- "cb"

39. Continue from the previous question. Given the previous bigram model, suppose a string starts with "a", what is the probability that the next two characters are "b" and "c".

- $0.2 \cdot 0.5$

40. Given the information gain (also called mutual information) between variables in the following table. Which edges are not included in the Bayesian network generated using Chow Liu Algorithm (maximum spanning tree based on mutual information as the edge weights)?

vertex	vertex	information gain
A	B	0.1
A	C	0.2
A	D	0.3
B	C	0.4
B	D	0.5
C	D	0.6

- Included edges: $(C, D), (B, D), (A, D)$