Midterm Version A

CS540

July 10, 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 $\hat{y}_i = a_i = \mathbb{1}_{\{wx_i+b\geqslant 0\}}$ with initial weights $w = 1$ and bias $b = -1$. Given a new input $x_i = 2$ and $y_i = 0$. Let the learning rate be 1, what is the updated weight w after one iteration of the perceptron algorithm?

   - A: -1
   - B: 0
   - C: 1
   - D: 2
   - E: 3

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

   - A: -3
   - B: -2
   - C: -1
   - D: 0
   - E: 1

3. Let $C(w) = w^T w, w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$. What is the Hessian matrix of C at $w = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$?

   - A: $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$
   - B: $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$
   - C: $\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$
   - D: $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
   - E: $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

4. Continue from the previous question, what is the Laplacian (trace of Hessian matrix) of C at $w = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$?

   - A: 0
   - B: 1
   - C: 2
   - D: 4
   - E: 8

5. Consider a linear model without bias term $a_i = wx_i$ with the log cost function. The initial weight is w_0. What is the updated weight after one stochastic gradient descent step for w if the chosen training data is $x_1 = 1, y_1 = 1$? The learning rate is α.

$$C\left(w\right) = -y_i \log\left(a_i\right) - \left(1 - y_i\right) \log\left(1 - a_i\right)$$

- A: $w_0 - \dfrac{\alpha}{w_0}$
- B: $w_0 - \alpha w_0$
- C: w_0
- D: $w_0 + \alpha w_0$
- E: $w_0 + \dfrac{\alpha}{w_0}$

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

- A: $w_0 - \dfrac{\alpha}{w_0}$
- B: $w_0 - \alpha w_0$
- C: w_0
- D: $w_0 + \alpha w_0$
- E: $w_0 + \dfrac{\alpha}{w_0}$

7. Perceptron algorithm will NOT terminate on which one of the following training set?

x_{i1}	x_{i2}	x_{i3}	$y_i^{(A)}$	$y_i^{(B)}$	$y_i^{(C)}$	$y_i^{(D)}$	$y_i^{(E)}$
0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	1	0	0	1	1	0	1
0	1	1	0	0	0	1	1
1	0	0	0	1	0	0	1
1	0	1	0	0	1	1	1
1	1	0	1	1	1	1	1
1	1	1	1	1	1	1	1

- A: Training set $\left\{\left(x_i, y_i^{(A)}\right)\right\}_{i=1}^{8}$
- B: Training set $\left\{\left(x_i, y_i^{(B)}\right)\right\}_{i=1}^{8}$
- C: Training set $\left\{\left(x_i, y_i^{(C)}\right)\right\}_{i=1}^{8}$
- D: Training set $\left\{\left(x_i, y_i^{(D)}\right)\right\}_{i=1}^{8}$
- E: Training set $\left\{\left(x_i, y_i^{(E)}\right)\right\}_{i=1}^{8}$

8. Suppose $(x-a)^2$ and $(x-b)^2$ are convex, which of the following functions is not convex? The domain is $(-\infty, \infty)$.

   - A: $(x-a)^2 + (x-b)^2$
   - B: $(x-a)^2 - (x-b)^2$
   - C: $(x-a)^2 + 2(x-b)^2$
   - D: $(x-a)^2 - 2(x-b)^2$
   - E: $2(x-a)^2 - (x-b)^2$

9. For a fully connected neural network with 10 input units $(x_{i1}, x_{i2}, ..., x_{i10})$, and 10 hidden units in the first layer $\left(a_{i1}^{(1)}, a_{i2}^{(1)}, ..., a_{i10}^{(1)}\right)$, and 5 hidden units in the second layer $\left(a_{i1}^{(2)}, a_{i2}^{(2)}, ..., a_{i5}^{(2)}\right)$, and a single unit in the last layer (binary classification) $\left(a_i^{(3)}\right)$. How many weights (not including biases) are updated during training?

   - A: $5 \cdot 5$
   - B: $5 \cdot 5 + 1$
   - C: $10 \cdot 10 + 10 \cdot 5$
   - D: $10 \cdot 10 + 10 \cdot 5 + 5$
   - E: $10 \cdot 10 \cdot 5$

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

    - A: $1 + 1 + 1$
    - B: $10 + 5$
    - C: $10 + 5 + 1$
    - D: $5 \cdot 5$
    - E: $5 \cdot 5 + 1$

11. Continue from the previous question. The notation $w_{j'j}^{(l)}$ is the weight from unit j' in layer $l-1$ to unit j in layer l, and the notation $a_j^{(l)}$ is the unit j activation on layer l for the current instance. Which of the following stochastic gradient expression is correct?

- A: $\frac{\partial C}{\partial w_{12}^{(1)}} = \sum_{j=1}^{10} \sum_{j'=1}^{5} \frac{\partial C}{\partial a_j^{(2)}} \frac{\partial a_j^{(2)}}{\partial a_{j'}^{(1)}} \frac{\partial a_{j'}^{(1)}}{\partial w_{1j'}^{(1)}}$
- B: $\frac{\partial C}{\partial w_{12}^{(1)}} = \sum_{j=1}^{5} \frac{\partial C}{\partial a_j^{(2)}} \frac{\partial a_j^{(2)}}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial w_{12}^{(1)}}$
- C: $\frac{\partial C}{\partial w_{12}^{(1)}} = \sum_{j=1}^{5} \frac{\partial C}{\partial a_j^{(2)}} \frac{\partial a_j^{(2)}}{\partial a_2^{(1)}} \frac{\partial a_2^{(1)}}{\partial w_{12}^{(1)}}$
- D: $\frac{\partial C}{\partial w_{12}^{(1)}} = \frac{\partial C}{\partial a_1^{(2)}} \frac{\partial a_1^{(2)}}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial w_{12}^{(1)}}$
- E: $\frac{\partial C}{\partial w_{12}^{(1)}} = \frac{\partial C}{\partial a_2^{(2)}} \frac{\partial a_2^{(2)}}{\partial a_2^{(1)}} \frac{\partial a_2^{(1)}}{\partial w_{12}^{(1)}}$

12. Continue from the previous question. Which of the following stochastic gradient expression is correct?

- A: $\frac{\partial C}{\partial b_2^{(1)}} = \sum_{j=1}^{10} \sum_{j'=1}^{5} \frac{\partial C}{\partial a_j^{(2)}} \frac{\partial a_j^{(2)}}{\partial a_{j'}^{(1)}} \frac{\partial a_{j'}^{(1)}}{\partial b_{j'}^{(1)}}$
- B: $\frac{\partial C}{\partial b_2^{(1)}} = \sum_{j=1}^{5} \frac{\partial C}{\partial a_j^{(2)}} \frac{\partial a_j^{(2)}}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial b_2^{(1)}}$
- C: $\frac{\partial C}{\partial b_2^{(1)}} = \sum_{j=1}^{5} \frac{\partial C}{\partial a_j^{(2)}} \frac{\partial a_j^{(2)}}{\partial a_2^{(1)}} \frac{\partial a_2^{(1)}}{\partial b_2^{(1)}}$
- D: $\frac{\partial C}{\partial b_2^{(1)}} = \frac{\partial C}{\partial a_1^{(2)}} \frac{\partial a_1^{(2)}}{\partial a_2^{(1)}} \frac{\partial a_2^{(1)}}{\partial b_2^{(1)}}$
- E: $\frac{\partial C}{\partial b_2^{(1)}} = \frac{\partial C}{\partial a_2^{(2)}} \frac{\partial a_2^{(2)}}{\partial a_2^{(1)}} \frac{\partial a_2^{(1)}}{\partial b_2^{(1)}}$

13. Given the following weights of a two layer neural network, which of following logical operators does it represent?

$$w_{11}^{(1)} = +2, w_{21}^{(1)} = -20, b_1^{(1)} = -1$$
$$w_{12}^{(1)} = -2, w_{22}^{(1)} = +20, b_2^{(1)} = -10$$
$$w_{11}^{(2)} = -2, w_{21}^{(2)} = -2, b_1^{(2)} = +1$$

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

- A: XOR
- B: NOR
- C: XNOR
- D: $\Rightarrow$
- E: $\Leftarrow$

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 five data points.

x	1	2	3	4	5	6	7	8	9	10
y	1	1	0	1	1	0	0	1	0	0

- A: 0 percent
- B: 10 percent
- C: 20 percent
- D: 50 percent
- E: 100 percent

15. Continue from the previous question. What is 2 fold cross validation accuracy (percentage of correct classificationn) if 3 nearest neighbor classifier with Manhattan distance is used?

- A: 0 percent
- B: 10 percent
- C: 20 percent
- D: 50 percent
- E: 100 percent

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

- A: $w_1 = 1, w_2 = 0$
- B: $w_1 = 0, w_2 = -1$
- C: $w_1 = -1, w_2 = 0$
- D: $w_1 = 0, w_2 = 1$
- E: $w_1 = -1, w_2 = -1$

17. Find the weights w_1, w_2 for the support vector machine classifier $\mathbb{1}_{\{w_1 x_{i1} + w_2 x_{i2} - 2 \geqslant 0\}}$ given the following training data. Note that the bias $b = -2$ is fixed.

x_1	x_2	y
0	0	0
1	1	0
2	2	1
3	3	1

- A: $w_1 = \dfrac{1}{2}, w_2 = \dfrac{1}{2}$
- B: $w_1 = -\dfrac{1}{2}, w_2 = -\dfrac{1}{2}$
- C: $w_1 = \dfrac{3}{4}, w_2 = \dfrac{3}{4}$
- D: $w_1 = -\dfrac{3}{4}, w_2 = -\dfrac{3}{4}$
- E: $w_1 = \dfrac{2}{3}, w_2 = \dfrac{2}{3}$

18. Continue from the previous question. What is the margin? Use Euclidean distance.

- A: $\dfrac{1}{2}$
- B: $\dfrac{\sqrt{2}}{2}$
- C: 1
- D: $\sqrt{2}$
- E: 2

19. What is the convolution between the following two matrices (use zero padding, i.e. set nonexistent values to 0 around the edges of the first matrix)? Remember to flip the filter first.

$$\begin{bmatrix}1 & 2 & 3\\4 & 5 & 6\\7 & 8 & 9\end{bmatrix} * \begin{bmatrix}0 & 0 & 0\\0 & 0 & 0\\0 & 0 & 1\end{bmatrix}$$

- A: $\begin{bmatrix}1 & 2 & 0\\4 & 5 & 0\\0 & 0 & 0\end{bmatrix}, B: \begin{bmatrix}0 & 0 & 0\\0 & 1 & 2\\0 & 4 & 5\end{bmatrix}$
- C: $\begin{bmatrix}9 & 8 & 0\\6 & 5 & 0\\0 & 0 & 0\end{bmatrix}, D: \begin{bmatrix}0 & 0 & 0\\0 & 9 & 8\\0 & 6 & 5\end{bmatrix}$
- E: $\begin{bmatrix}5 & 6 & 0\\8 & 9 & 0\\0 & 0 & 0\end{bmatrix}$

20. Continue from the previous question, what is the convolution between the following two matrices? Remember to flip the filter first.

$$\begin{bmatrix}0 & 0 & 0\\0 & 0 & 0\\0 & 0 & 1\end{bmatrix} * \begin{bmatrix}1 & 2 & 3\\4 & 5 & 6\\7 & 8 & 9\end{bmatrix}$$

- A: $\begin{bmatrix}1 & 2 & 0\\4 & 5 & 0\\0 & 0 & 0\end{bmatrix}, B: \begin{bmatrix}0 & 0 & 0\\0 & 1 & 2\\0 & 4 & 5\end{bmatrix}$
- C: $\begin{bmatrix}9 & 8 & 0\\6 & 5 & 0\\0 & 0 & 0\end{bmatrix}, D: \begin{bmatrix}0 & 0 & 0\\0 & 9 & 8\\0 & 6 & 5\end{bmatrix}$
- E: $\begin{bmatrix}5 & 6 & 0\\8 & 9 & 0\\0 & 0 & 0\end{bmatrix}$

21. In a convolutional neural network, suppose the resulting activation matrix of the convolution layer is the following $A^{(1)}$. What is the activation matrix after a non-overlapping 2×2 max pooling layer? Non-overlapping is also called stride 2.

$$A^{(1)} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix}$$

- A: $\begin{bmatrix} 6 & 6 & 8 & 8 \\ 6 & 6 & 8 & 8 \\ 14 & 14 & 16 & 16 \\ 14 & 14 & 16 & 16 \end{bmatrix}$
- B: $\begin{bmatrix} 6 & 7 & 8 \\ 10 & 11 & 12 \\ 14 & 15 & 16 \end{bmatrix}$
- C: $\begin{bmatrix} 6 & 8 \\ 14 & 16 \end{bmatrix}$
- D: $\begin{bmatrix} 13 & 14 & 15 & 16 \end{bmatrix}$
- E: $\begin{bmatrix} 4 \\ 8 \\ 12 \\ 16 \end{bmatrix}$

22. Given the following 3×3 image, what is the gradient magnitude of the center pixel using the derivative filters $\begin{bmatrix} -1 & 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 0 \\ -1. \end{bmatrix}$

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

- A: $\sqrt{(1)^2 + (1)^2}$
- B: $\sqrt{(1)^2 + (2)^2}$
- C: $\sqrt{(1)^2 + (3)^2}$
- D: $\sqrt{(2)^2 + (4)^2}$
- E: $\sqrt{(2)^2 + (6)^2}$

23. Given the magnitude and direction for a 2×2 cell on an image, and suppose there are 2 bins. What is the histogram of oriented gradient for this 2×2 cell? Opposite directions, for example 0 and π and -π, are considered the same orientation. (Different from original HOG paper: do not split a single gradient magnitude into two bins.)

$$M = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \Theta = \begin{bmatrix} 0.25\pi & 0.75\pi \\ -0.25\pi & -0.75\pi \end{bmatrix}$$

- A: $\begin{bmatrix} 1 & 3 \end{bmatrix}$
- B: $\begin{bmatrix} 2 & 2 \end{bmatrix}$
- C: $\begin{bmatrix} 3 & 7 \end{bmatrix}$
- D: $\begin{bmatrix} 4 & 6 \end{bmatrix}$
- E: $\begin{bmatrix} 5 & 5 \end{bmatrix}$

24. The histogram of oriented gradient feature vector for a 4×4 image with 2×2 cells, 1×1 blocks has length 16. How many bins are there?

- A: 1
- B: 2
- C: 4
- D: 8
- E: 16

25. Bagging uses subsamples (with replacement) of the training set to train different decision trees. Given the training set $\left(x_i, y_i\right)_{i=1...4} \in \left\{(1,1), (2,0), (3,1), (4,0)\right\}$, how many of the following sets are possible subsamples?

$$\left\{(1,1), (2,0), (3,1), (4,0)\right\}$$
$$\left\{(1,1), (1,1), (2,0), (2,0)\right\}$$
$$\left\{(1,0), (2,1), (3,1), (4,0)\right\}$$
$$\left\{(1,1), (1,1), (1,1), (1,1)\right\}$$

- A: 0
- B: 1
- C: 2
- D: 3
- E: 4

26. What is the accuracy (on the training set) of the decision stump with 1 split on x_1 trained on the following training set?

$$\left(x_{i1}, x_{i2}, y_i\right)_{i=1...5} = \left\{\left(0,0,0\right), \left(0,0,0\right), \left(1,1,1\right), \left(1,1,1\right), \left(1,1,0\right)\right\}$$

- A: 20 percent
- B: 40 percent
- C: 60 percent
- D: 80 percent
- E: 100 percent

27. Continue from the previous question, what is accuracy if the decision stump is split on $x_{2?}$

- A: 20 percent
- B: 40 percent
- C: 60 percent
- D: 80 percent
- E: 100 percent

28. Given the following training set S, suppose n instances are removed, and 1 nearest neighbor with Manhattan distance is trained on the remaining set S', and tested on the original training set S. If an instance is equally distant from two points with different labels, the 1 nearest neighbor labels the instance as 0 with probability $\dfrac{1}{2}$ and 1 with probability $\dfrac{1}{2}$. If the accuracy is 100 percent (for certain) on S, what is the maximum possible value for n?

$$S = \left(x_i, y_i\right)_{i=1...5} = \left\{\left(-2,0\right), \left(-1,0\right), \left(0,0\right), \left(1,1\right), \left(2,1\right)\right\}$$

- A: 0
- B: 1
- C: 2
- D: 3
- E: 4

29. Continue from the previous question. Same assumptions as the previous question: what is the maximum possible value for n if the training set S is changed to the following?

$$S = \left(x_i, y_i\right)_{i=1...5} = \left\{\left(-2,1\right), \left(-1,0\right), \left(0,0\right), \left(1,1\right), \left(2,1\right)\right\}$$

- A: 0
- B: 1
- C: 2
- D: 3
- E: 4

30. Suppose there are 10 training instances, and the feature mapping for a support vector machine with kernel trick is given by $\phi(x_1, x_2) = \left(x_1^2, 2x_1x_2, x_2^2\right)$. What is the size of the kernel matrix (Gram matrix)?

- A: 2×2
- B: 3×3
- C: 10×10
- D: 20×20
- E: 30×30

31. Two documents A and B contain only words H and T. Document A is $HHTTTTTTTT$ and document B is $HHHHHTTTTT$. There are 2 copies of document A and 3 copies of document B. One document is chosen at random (each copy with equal probability) and one word is chosen at random (each word with equal probability). What is the probability that the word is H?

- A: $0.2 + 0.5$
- B: $0.2 \cdot 0.5 + 0.5 \cdot 0.5$
- C: $0.2 \cdot 0.5 + 0.8 \cdot 0.5$
- D: $0.2 \cdot 0.4 + 0.5 \cdot 0.6$
- E: $0.8 \cdot 0.4 + 0.5 \cdot 0.6$

32. Continue from the previous question, suppose the answer to the question is p. Given the chosen word is H, what is the probability that the document is A?

- A: $\dfrac{0.2 \cdot 0.5}{p}$
- B: $\dfrac{0.2 \cdot 0.4}{p}$
- C: $\dfrac{0.2 \cdot 0.5 + 0.5 \cdot 0.5}{p}$
- D: $\dfrac{0.2 \cdot 0.4 + 0.5 \cdot 0.6}{p}$
- E: 0.2

33. Given the counts, find the maximum likelihood estimate of $\mathbb{P}\left\{A | \neg B, \neg C\right\}$ with Laplace smoothing. The event A means $A = T$ and the event $\neg A$ means $A = F$ in the following table.

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

- A: $\dfrac{1 + 1}{2 + 1}$
- B: $\dfrac{1 + 1}{2 + 2}$
- C: $\dfrac{1 + 1}{2 + 4}$
- D: $\dfrac{0 + 1}{2 + 2}$
- E: $\dfrac{0 + 1}{2 + 4}$

34. Suppose A is the common effect of B and C, which means the Bayesian network is $B \rightarrow A \leftarrow C$. All variables are binary. What is $\mathbb{P}\{B = 1, C = 1\}$?

$$\mathbb{P}\{C = 1\} = 0.4, \mathbb{P}\{B = 1\} = 0.9$$
$$\mathbb{P}\{A = 1|B = 1, C = 1\} = 0.8, \mathbb{P}\{A = 1|B = 0, C = 1\} = 0.2$$
$$\mathbb{P}\{A = 1|B = 1, C = 0\} = 0.3, \mathbb{P}\{A = 1|B = 0, C = 0\} = 0.5$$

- A: $0.4 \cdot 0.9$
- B: $0.4 \cdot 0.9 \cdot 0.8$
- C: $0.4 \cdot 0.9 \cdot 0.8 + 0.6 \cdot 0.9 \cdot 0.3 + 0.4 \cdot 0.1 \cdot 0.2 + 0.6 \cdot 0.1 \cdot 0.5$
- D: $0.4 \cdot 0.1 + 0.6 \cdot 0.9$
- E: $0.4 \cdot 0.1 \cdot 0.2 + 0.6 \cdot 0.9 \cdot 0.3$

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

- A: 0.4
- B: $\dfrac{0.4}{0.9}$
- C: $\dfrac{0.4 \cdot 0.1 + 0.6 \cdot 0.9}{0.9}$
- D: $\dfrac{0.4 \cdot 0.1 + 0.6 \cdot 0.9}{0.9}$
- E: $\dfrac{0.4 \cdot 0.9 \cdot 0.8 + 0.6 \cdot 0.9 \cdot 0.3 + 0.4 \cdot 0.1 \cdot 0.2 + 0.6 \cdot 0.1 \cdot 0.5}{0.9}$

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

- A: 0.9
- B: $\dfrac{0.9}{0.4}$
- C: $\dfrac{0.4 \cdot 0.1 + 0.6 \cdot 0.9}{0.4}$
- D: $\dfrac{0.4 \cdot 0.9 + 0.6 \cdot 0.1}{0.4}$
- E: $\dfrac{0.4 \cdot 0.9 \cdot 0.8 + 0.6 \cdot 0.9 \cdot 0.3 + 0.4 \cdot 0.1 \cdot 0.2 + 0.6 \cdot 0.1 \cdot 0.5}{0.4}$

37. Given the following training set, what is the most likely character if a new string starts with "baa" assuming the bigram for chracters model?

$$(z_1, z_2, ..., z_{47}) = \text{"aaaaabaacaadaaeaaeaadaadaaabbbbbbbbbbbbbbbbbbbb.."}$$

- A: a
- B: b
- C: c
- D: d
- E: e

38. Given the following training set, what is the most likely character if a new string starts with "baa" assuming the trigram for chracters model? Note the previous question uses BIgram and this question uses TRIgram, the string is the same.

$$(z_1, z_2, ..., z_{47}) = \text{"aaaaabaacaadaaeaaeaadaadaaabbbbbbbbbbbbbbbbbbbb.."}$$

- A: a
- B: b
- C: c
- D: d
- E: e

39. Given the following transition matrix for a bigram model with characters "a" "b" "c", what is the (estimated) probability that the third character is "c" given the first is "a". For example, row b column c of the matrix is the probability that a "c" follows an "b": $\mathbb{P}\{c|b\} = 0.5$.

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

- A: 0.7
- B: $0.2 \cdot 0.4 + 0.3 \cdot 0.3$
- C: $0.2 \cdot 0.5 + 0.7 \cdot 0.3$
- D: $0.1 \cdot 0.7 + 0.2 \cdot 0.5 + 0.7 \cdot 0.3$
- E: $0.3 \cdot 0.3 + 0.2 \cdot 0.4 + 0.1 \cdot 0.3$

40. Continue from the previous question, suppose the model is trained (without smoothing) on a training set with 100 characters in total, and there are 20 "a" and 20 "c". How many "ac" substring are there? The last character in the training set string is "c".

- A: 6
- B: 7
- C: 14
- D: 15
- E: 20