CS540 Midterm Exam

October 25

Fall 2024

1 Student Information

- Exam version: 1
- First, Last Name:
- 10-digit Campus ID:
- Section: $\boxed{001}$ MW 4 : 00, $\boxed{002}$ TR 2 : 30, $\boxed{003}$ TR 1 : 00
- Net ID (Wisc Email ID):

2 Instructions

1. You cannot sit next (front, back, left (within two seats including an empty seat), right (within two seats including an empty seat)) to someone you know. Switch seats now if that is the case.
2. Fill in these fields (left to right) on the Scantron sheet using pencil:
   - LAST NAME (family name) and FIRST NAME (given name), fill in bubbles
   - IDENTIFICATION NUMBER is your Campus ID number, fill in bubbles
   - Under A, B, C of SPECIAL CODES, fill in your three-digit section number (001 , 002, 003)
   - Under D of SPECIAL CODES, fill in Exam version above. This is very important!

   Make sure you fill all the above accurately in order to get graded.
3. Mark your answers on the Scantron sheet and this handout. Use a pencil to mark all answers. Each question has exactly one correct answer.
4. You may only reference your one-sheet notes. Please turn off and put away portable electronics now.
5. If you need to use the restroom, your phone must remain in the room, placed face down on your desk and kept visible to the proctors.
6. You have 90 minutes to take the exam. Once you have finished, please show your student ID to one of the proctors and submit both your Scantron and this handout. You do not need to submit your one-sheet note.

Good luck!

-

3 Questions

1. Given the training document "I am Groot I am Groot am I Groot Groot I am I am" , what is the maximum likelihood estimate of the bigram probability $\mathbb{P}\{\text{ Groot} \mid \text{I }\}$ (that is $\mathbb{P}\{w_t = \text{ Groot} \mid w_{t-1} = \text{ I }\}$)?

   - A: $\frac{1}{5}$
   - B: $\frac{2}{4}$
   - C: $\frac{1}{4}$
   - D: $\frac{2}{5}$
   - E: None of the above (or more information is needed)

2. Suppose the phrase "Nice Pool" never appeared in a training document with 1000 word tokens and 100 unique word types, what is the maximum likelihood estimate of $\mathbb{P}\{\text{ Proposal} \mid \text{Nice Pool }\}$ with add-one Laplace smoothing (that is $\mathbb{P}\{w_t = \text{ Proposal} \mid w_{t-1} = \text{ Pool}\,, w_{t-2} = \text{ Nice }\}$)? The words "Nice", "Pool", and "Proposal" are in the vocabulary.

   - A: $\frac{2}{101}$
   - B: $\frac{1}{100}$
   - C: $\frac{2}{1002}$
   - D: $\frac{1}{1000}$
   - E: None of the above (or more information is needed)

3. There are $n = 5$ training documents in a training set, and the TF-IDF feature of the training document 2 for word "Joker" is 0. If in document 2, there are 100 words, and the word "Joker" appeared 3 times. How many documents (out of 5) contains the word "Joker"?

   - A: 2
   - B: 5
   - C: 3
   - D: 1
   - E: None of the above (or more information is needed)

4. Consider the problem of classifying whether a painting is art (A) or trash (T) based on the price (P), which can be high (H), medium (M), or low (L), and a naive Bayes model trained on a training set has the prior probability $\mathbb{P}\{A\} = \frac{1}{3}$ and conditional probabilities $\mathbb{P}\{H|A\} = \frac{3}{5}, \mathbb{P}\{L|A\} = 0, \mathbb{P}\{H|T\} = \frac{1}{5}, \mathbb{P}\{L|T\} = \frac{3}{5}$. Now given a new painting with medium price, what is the conditional probability it is art (that is, $\mathbb{P}\{A|M\}$)?

   - A: $\frac{2}{3}$

- B: $\frac{1}{5}$
- C: $\frac{3}{5}$
- D: $\frac{1}{2}$
- E: None of the above (or more information is needed)

5. If the eigenvalues of the variance covariance matrix of a training set with 5 features are $4, 3, 2, 1$, and 0, in order to explain 80 percent or more of the variation in the training data with the first k principal components, what is the smallest value of k?

   - A: 5
   - B: 4
   - C: 2
   - D: 3
   - E: None of the above (or more information is needed)

6. The first two principal components based on a training set with 3 features is $u_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ and $u_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix}$. The PCA weights (or PCA features) of a training item is $\begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$, what are the original features of this item?

   - A: $\begin{bmatrix} -\frac{1}{\sqrt{2}} \\ 1 \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ or $\begin{bmatrix} -\frac{1}{\sqrt{2}} \\ -1 \\ \frac{1}{\sqrt{2}} \end{bmatrix}$
   - B: $\begin{bmatrix} -\frac{1}{\sqrt{2}} \\ 2 \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ or $\begin{bmatrix} -\frac{1}{\sqrt{2}} \\ -2 \\ \frac{1}{\sqrt{2}} \end{bmatrix}$
   - C: $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{3}{\sqrt{2}} \end{bmatrix}$
   - D: $\begin{bmatrix} -\frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix}$
   - E: None of the above (or more information is needed)

7. Which of the following feature vectors have a Manhattan L_1 distance of exactly 5 away from $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$?

- A: Only $\begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}$
- B: More than one of $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ or $\begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}$ or $\begin{bmatrix} 0 \\ -1 \\ -2 \end{bmatrix}$
- C: Only $\begin{bmatrix} 0 \\ -1 \\ -2 \end{bmatrix}$
- D: Only $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$
- E: None of the above

8. The pairwise distance matrix between the four items $\{1, 2, 3, 4\}$ is given by $\begin{bmatrix} 0 & 2 & 4 & 6 \\ 2 & 0 & 3 & 5 \\ 4 & 3 & 0 & 1 \\ 6 & 5 & 1 & 0 \end{bmatrix}$. What is the complete linkage distance between clusters $\{1, 2\}$ and $\{3, 4\}$ (items 1 and 2 in the first cluster and items 3 and 4 in the second cluster)?

- A: 6
- B: 2
- C: 5
- D: 1
- E: None of the above

9. The pairwise distance matrix between the four clusters $\{1, 2, 3, 4\}$ is given by $\begin{bmatrix} 0 & 2 & 4 & 6 \\ 2 & 0 & 3 & 5 \\ 4 & 3 & 0 & 1 \\ 6 & 5 & 1 & 0 \end{bmatrix}$. Which two clusters are merged in the next iteration of hierarchical clustering?

- A: $\{1, 4\}$ using single linkage, $\{3, 4\}$ using complete linkage
- B: $\{3, 4\}$ using single linkage, $\{1, 4\}$ using complete linkage
- C: $\{1, 4\}$ using both single and complete linkage distances
- D: $\{3, 4\}$ using both single and complete linkage distances
- E: None of the above (or more information is needed)

10. During k-means clustering with $k = 2$ and Euclidean distance on a training set with 1 feature and 4 items $\{1, 3, 5, 7\}$, the cluster centers are $c_1 = 2$ and $c_2 = x$. If cluster 2 is empty after the points are reassigned to the centers, which of the following are possible values of x? In case of tie in distances, the item is assigned to cluster 1.

    - A: Only $x = 15$
    - B: Only $x = 13$
    - C: More than one of $x = 11$ or $x = 13$ or $x = 15$
    - D: Only $x = 11$
    - E: None of the above (or more information is needed)

11. Given a training set containing 100 images of cats and 10 images of flerkens. There are 2 features for each image. A K-nearest neighbor classifier trained on this training set classifies every possible new image as a cat. Which of the following are possible values of K for this to happen? The answer should apply for any training set, not just training sets with special feature values.

    - A: Only $K = 1$
    - B: More than one of $K = 1$ or $K = 11$ or $K = 21$
    - C: Only $K = 21$
    - D: Only $K = 11$
    - E: None of the above (or more information is needed)

12. What is the 6-fold cross validation accuracy of the K-nearest neighbor classifier with $K = 1$ on the following training set $\{(x_i, y_i)\}_{i=1}^{6} = \{(1, C), (3, C), (6, F), (7, C), (9, C), (10, F)\}$? There is one numerical feature and one binary label C for cat and F for flerken. Tie-breaking: if two training items have the same distance to the test item, the nearest neighbor is the item with the smaller x_i value.

    - A: $\frac{1}{3}$
    - B: $\frac{2}{3}$
    - C: 1
    - D: $\frac{1}{2}$
    - E: None of the above (or more information is needed)

13. There are 10 minions: 3 of them (Bob, Carl and Stuart) are short and have one eye; 1 of them (John) is tall and has one eye; n of them are short and have two eyes, $6 - n$ of them are tall and have two eyes. Which of the following value of n would maximize the information gain when predicting the number of eyes (one or two) based on the height (short or tall)?

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

- E: More information is needed

14. There are 4 binary features (each feature takes value of either 0 or 1) and one binary label in a training set, and a decision tree is trained on this training set (ID3 algorithm, leafs are created when maximum information gain is 0 among all possible splits, and no pruning is done). What is maximum possible depth of the decision tree? The root is at depth 0.

    - A: 16
    - B: 4
    - C: Infinity
    - D: 5
    - E: None of the above (or more information is needed)

15. What is the validation accuracy of the decision tree with depth 1 : $\begin{cases} C & \text{if } x_{i2} \leqslant 1 \\ F & \text{if } x_{i2} > 1 \end{cases}$ on the validation set: $\{(x_i, y_i)\}_{i=1}^{6} = \left\{ \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, C \right), \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, C \right), \left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}, F \right), \left(\begin{bmatrix} 2 \\ 1 \end{bmatrix}, C \right), \left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, C \right), \left(\begin{bmatrix} 2 \\ 2 \end{bmatrix}, F \right) \right\}$? There are two numerical features $\begin{bmatrix} x_{i1} \\ x_{i2} \end{bmatrix}$ and one binary label C for cat and F for flerken.

    - A: $\frac{2}{3}$
    - B: $\frac{5}{6}$
    - C: 1
    - D: $\frac{1}{3}$
    - E: None of the above (or more information is needed)

16. In one iteration of the linear threshold unit perceptron algorithm, an item with feature vector $\begin{bmatrix} 1 \\ 0 \\ -1 \\ 1 \end{bmatrix}$ has true label 0 but is predicted as 1 by the perceptron based on its current weights. When these weights are updated, how many of the weights, including the bias, are decreased? Do not count the ones that stay the same.

    - A: 1
    - B: 3
    - C: 4
    - D: 0
    - E: None of the above (or more information is needed)

17. Let $a_i = \frac{1}{1 + e^{-z_i}}$ be the logistic activation where $z_i = w_1 x_{i1} + w_2 x_{i2} + ... + w_m x_{im} + b$. Which of the following expressions are equivalent to $a_i \geqslant 0.5$?

- A: $z_i \leqslant 0$
- B: $z_i \leqslant 1$
- C: $z_i \geqslant 1$
- D: $z_i \geqslant 0$
- E: None of the above (or more information is needed)

18. For a linear threshold unit perceptron for binary classification problem (that is, $y_i \in \{0, 1\}$), which of the following loss functions are the same (that is, the loss is the same for all possible values of a_i and y_i)? The activation function is $a_i = \begin{cases} 1 & \text{if } wx_i + b \geqslant 0 \\ 0 & \text{if } wx_i + b < 0 \end{cases}$.

$$
\begin{aligned}
C\left(a_i, y_i\right) &= \left(a_i - y_i\right)^{2,} \\
C'\left(a_i, y_i\right) &= \left|a_i - y_i\right|, \\
C''\left(a_i, y_i\right) &= \frac{1}{2} \max\left\{0, 1 - \left(2a_i - 1\right)\left(2y_i - 1\right)\right\}, \\
C'''\left(a_i, y_i\right) &= -y_i \log\left(a_i\right) - \left(1 - y_i\right) \log\left(1 - a_i\right).
\end{aligned}
$$

- A: Only C, C'
- B: Only C, C', C''
- C: All four are the same
- D: Only C, C'', C''''
- E: None of the above

19. At the beginning of one gradient descent step to find the weight and the bias for a linear regression problem with a quadratic loss function, the weight is $w = 2$ and the bias is $b = -1$. Consider the following training set $\{(x_i, y_i)\}_{i=1}^{5} = \{(-2, 0), (-1, 0), (0, 0), (1, 1), (2, 1)\}$. Which point lead to the largest loss C_i? The answer should not depend on the constant in front the quadratic loss function.

- A: Only $(2, 1)$
- B: Only $(-2, 0)$
- C: $(-2, 0)$ and $(2, 1)$ lead to the same largest loss
- D: Only $(0, 0)$
- E: None of the above (or more information is needed)

20. How many of the following training items would violate the hard margin support vector machine constraints with $w = \frac{1}{2}, b = 0$? The training set is $\{(x_i, y_i)\}_{i=1}^{4} = \{(-3, 0), (-1, 0), (1, 1), (2, 1)\}$. The support vector machine classifies items with $wx_i + b \geqslant 0$ as class 1.

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

- E: None of the above (or more information is needed)

21. Hard margin support vector machine is trained on the following training set and all five items are support vectors: $\{(x_i, y_i)\}_{i=1}^{5} = \left\{ \left(\begin{bmatrix} 1 \\ -1 \end{bmatrix}, 0 \right), \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, 0 \right), \left(\begin{bmatrix} -1 \\ 1 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} -1 \\ 2 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} a \\ b \end{bmatrix}, 1 \right) \right\}$. Which of the following values of $\begin{bmatrix} a \\ b \end{bmatrix}$ are possible?

- A: Only $\begin{bmatrix} -1 \\ -1 \end{bmatrix}$
- B: More than one of $\begin{bmatrix} -1 \\ -1 \end{bmatrix}$ or $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ or $\begin{bmatrix} -1 \\ 3 \end{bmatrix}$
- C: Only $\begin{bmatrix} -1 \\ 3 \end{bmatrix}$
- D: Only $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$
- E: None of the above (or more information is needed)

22. A two-layer fully connected neural network has 1 hidden layer with 5 units and the softmax output layer has 3 units. There are a total of 65 weights and 8 biases in the network. How many features (input layer units) do the network has?

- A: 7
- B: It is impossible for such a network to have 65 weights and 8 biases
- C: 3
- D: 10
- E: None of the above (or more information is needed)

23. In a three-layer fully connected neural network, let $a_{i2}^{(1)}$ be the second tanh activation unit in the first hidden layer and $a_{i3}^{(2)}$ be the third sigmoid activation unit in the second hidden layer. What is $\frac{\partial a_{i3}^{(2)}}{\partial a_{i2}^{(1)}}$? For tanh activation function $g(z)$, the derivative $g'(z) = 1 - g(z)^2$, and for sigmoid activation function $g(z)$, the derivative $g'(z) = g(z)(1 - g(z))$.

- A: $a_{i2}^{(1)} \left(1 - a_{i2}^{(1)}\right) w_{23}^{(2)}$
- B: $\left(1 - \left(a_{i2}^{(1)}\right)^2\right) w_{23}^{(2)}$
- C: $\left(1 - \left(a_{i2}^{(1)}\right)^2\right) w_{32}^{(2)}$
- D: $a_{i2}^{(1)} \left(1 - a_{i2}^{(1)}\right) w_{32}^{(2)}$
- E: None of the above (or more information is needed)

24. In a three-layer fully connected neural network, let $a_{i2}^{(1)}$ be the second tanh activation unit in the first hidden layer and $a_{i3}^{(2)}$ be the third sigmoid activation unit in the second hidden layer. Which of the following expressions for $\frac{\partial a_{i3}^{(2)}}{\partial w_{12}^{(1)}}$ are correct?

- A: $\frac{\partial a_{i3}^{(2)}}{\partial a_{i2}^{(1)}} a_{i1}^{(1)}$
- B: $\frac{\partial a_{i3}^{(2)}}{\partial a_{i1}^{(1)}} a_{i2}^{(1)}$
- C: $\frac{\partial a_{i3}^{(2)}}{\partial a_{i1}^{(1)}} x_{i2}$
- D: $\frac{\partial a_{i3}^{(2)}}{\partial a_{i2}^{(1)}} x_{i1}$
- E: None of the above (or more information is needed)

25. For some loss function $C = C_1 + C_2$ with $\frac{\partial C_i}{\partial w} = x_i w$, the training set (after shuffling) with two items with $x_1 = -2$ and $x_2 = 4$ is used in one epoch of stochastic gradient descent to minimize C. The weight at the beginning of this epoch is $w = 1$. What is the weight after this epoch (two stochastic gradient descent steps, x_1 first, x_2 next)? The learning rate is $\alpha = 1$.

- A: 3
- B: -9
- C: -1
- D: 1
- E: None of the above (or more information is needed)

26. Which of the following can be the activation values $a_i^{(L)}$ of the softmax output layer of a neural network with 3 output units for some training item i? The hidden units have the sigmoid activation function.

- A: Only $\begin{bmatrix} 0.2 \\ 0.3 \\ 0.5 \end{bmatrix}$
- B: Only $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$
- C: More than one of $\begin{bmatrix} 0.2 \\ 0.3 \\ 0.5 \end{bmatrix}$ or $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ or $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$
- D: Only $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$
- E: None of the above (or more information is needed)

27. Consider a stochastic gradient descent step of linear regression with quadratic loss $\frac{1}{2}(a_i - y_i)^2$ plus L_2 regularization, where a_i is the linear (or identity) activation value, if the weight is $w = 2$, the bias is $b = 0$, the randomly selected item for this iteration has $x_i = 1$ and $y_i = 1$, the regularization parameter $\lambda = 0.1$, the learning rate is $\alpha = 1$, what is the weight w after this stochastic gradient descent step?

- A: 0.6
- B: 1
- C: 1.6
- D: 2
- E: None of the above (or more information is needed)

28. What is the convolution between the image $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ and the filter $\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$ (same when flipped)? Assume zero padding and a stride of 1.
    - A: $\begin{bmatrix} -3 & 2 \\ 3 & 3 \end{bmatrix}$
    - B: $\begin{bmatrix} -4 & 0 \\ 0 & -1 \end{bmatrix}$
    - C: $\begin{bmatrix} 1 & -1 \\ 1 & 4 \end{bmatrix}$
    - D: $\begin{bmatrix} -2 & 0 \\ 0 & 0 \end{bmatrix}$
    - E: None of the above (or more information is needed)
29. A convolutional neural network has input (grayscale) image of size 30×30 that is connected to a convolutional layer with 2 activation maps, where each uses a 5×5 filter, with zero padding and a stride of 1. The convolutional layer is then connected to a pooling layer that uses 3×3 max pooling, non-overlapping and no padding. The pooling layer is then fully connected to an output softmax layer with 3 units. How many weights, not including biases, are learned (updated during training) in this network?
    - A: 37
    - B: 650
    - C: 74
    - D: 325
    - E: None of the above (or more information is needed)
30. For a recurrent neural network with 3 input features, 2 hidden recurrent units, and 1 output unit at the end of the sequence. During one step of stochastic gradient descent, for one training item, a sequence of length 5, the values of 11 weights are updated. How many weights will be updated for another training sequence of length 7? Note: a weight is considered updated if one gradient descent step is applied, including when the gradient has value 0.
    - A: 19
    - B: 11
    - C: 23
    - D: 15
    - E: None of the above (or more information is needed)