# FINAL EXAM

CS 178, SPRING 2023

June 14th 2023

**YOUR STUDENT ID:**

**YOUR NAME:**

**YOUR SEAT NUMBER:**

## Instructions

- DO NOT GO BEYOND THIS PAGE TO START UNTIL INSTRUCTED TO DO SO.
- Please write your student ID, full name, and seat number in the spaces provided above.
- Put away everything (including any electronic devices, written materials, etc) except for a pen or pencil.
- If you think there is a typo in a question, please read the question carefully at least twice. If you still think there is a typo please raise your hand and we will respond.
- The exam will last for 2 hours. There are 6 questions on the exam for a total of 160 points. The number of points per question is listed, but this value is subject to change during grading.
- If you need to use a rest-room please raise your hand and let one of the TAs know. You will need to drop off your cellphone at the front of the class before you leave: you can pick it up after the exam is over.
- When instructed that the exam is over, put down your pen/pencil and stop writing. Remain in your seat and we will collect exams row by row.
- If you finish with more than 10 minutes of the exam remaining you can leave your seat and drop your exam off at the front of the lecture hall. Otherwise wait in your seat until we collect the exams.

**Feel free to use this page for your rough work: it will be ignored during grading**

**Feel free to use this page for your rough work: it will be ignored during grading**

## Problem 1 (2 points/question; 20 points total.)

State TRUE or FALSE to indicate if each of the following statements are true or false:

1. When using stacking to create an ensemble, the weights used to perform stacking should be learned using the training data:
   **False**
2. The kMeans clustering algorithm has a time complexity of $O(dn^2)$ per training iteration:
   **False**
3. For any given classification problem (where there are no repeated datapoints), there always exists a decision tree achieving perfect accuracy:
   **True**
4. When creating bootstraps with the bagging method, we should sample without replacement so that the same datapoint does not appear more than once:
   **False**
5. In kMeans clustering, there is always a value for k (the number of clusters) that will result in a squared error of zero:
   **True**
6. Logistic regression models have low variance but high bias:
   **True**
7. A model that achieves low training error but high testing error is likely to be underfitting the data:
   **False**
8. The K-means clustering algorithm is always guaranteed to converge to an optimal solution:
   **False**
9. Neural networks with only a single hidden layer can learn non-linear decision boundaries:
   **True**
10. Given enough iterations, the kMeans clustering algorithm is guaranteed to converge to the global optimum of the squared error:
    **False**

## Problem 2: Neural Networks (30 points.)

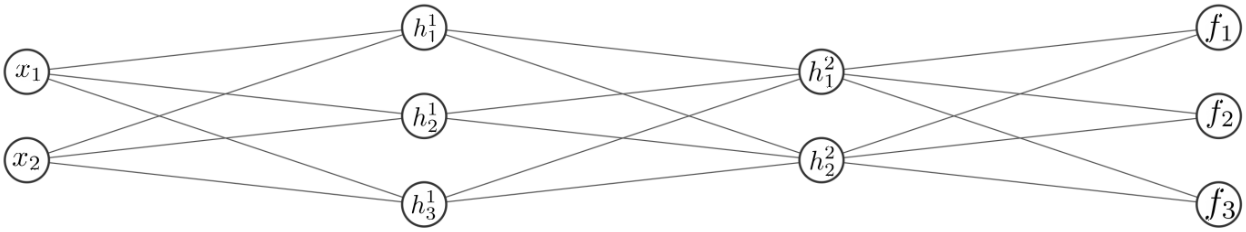


Figure 1:

A small neural network is drawn in Figure 1. This neural network takes in a two-dimensional feature vector $\mathbf{x} = (x_1, x_2)$ and has two hidden layers. The first hidden layer $\mathbf{h}^1 = (h_1^1, h_2^1, h_3^1)$ has three hidden units, and the second hidden layer $\mathbf{h}^2 = (h_1^2, h_2^2)$ has two hidden units. This network is being used for a classification task with three classes, and hence has three output nodes. The network has no bias units. Consider the following inputs and weight matrices, where $\mathbf{W}_1$ are the weights in the first layer, $\mathbf{W}_2$ are the weights in the second layer, and $\boldsymbol{\beta}$ are the weights in the final layer:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

$$\mathbf{W}_1 = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 0 & 2 \\ 1 & -4 \end{bmatrix} \qquad \mathbf{W}_2 = \begin{bmatrix} b_{11} & b_{21} & b_{31} \\ b_{12} & b_{22} & b_{32} \end{bmatrix} = \begin{bmatrix} 8 & 1 & -1 \\ 4 & -1 & 12 \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{11} & \beta_{21} \\ \beta_{12} & \beta_{22} \\ \beta_{13} & \beta_{23} \end{bmatrix} = \begin{bmatrix} -1 & -4 \\ 4 & 3 \\ -2 & 1 \end{bmatrix}$$

For example, $a_{21}$ is the weight connecting $x_2$ to $h_1^1$ and $\beta_{23}$ is the weight connecting $h_2^2$ to $f_3$.

1. (**20 points.**) Compute the values of all hidden and output nodes for this network. Use the ReLU activation function for all hidden layers, and the softmax activation function for the output layer. You do **not** need to simplify the computations for the final layer. Your answer should be written as vectors below. (The next page is blank for scratch work.)

**ANSWER:** $\mathbf{h}^1 = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}$ $\qquad \mathbf{h}^2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$ $\qquad \mathbf{f} = \begin{bmatrix} e^{-2}/(e^{-2} + e^{8} + e^{-4}) \\ e^{8}/(e^{-2} + e^{8} + e^{-4}) \\ e^{-4}/(e^{-2} + e^{8} + e^{-4}) \end{bmatrix}$

## Problem 2 Continued

## Problem 2 Continued

Suppose you are now given a filter **F** in a convolutional neural network and a single-channel input image **X**:

$$\mathbf{F} = \begin{bmatrix} 1 & -1 \\ 2 & 3 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 3 & 1 & 4 \\ 1 & 5 & 9 \\ 2 & 6 & 5 \end{bmatrix}$$

2. (**10 points.**) Perform convolution with the filter **F** on the input **X**. Your answer should be in the form of a matrix.

$$\textbf{ANSWER} = \begin{bmatrix} 19 & 34 \\ 18 & 23 \end{bmatrix}$$

## Problem 3: K-Means Clustering (25 points.)

Consider the following dataset of seven 1-dimensional data points: $\{2, 6, 7, 4, -1, 0, 1\}$. We would like to run the K-means algorithm on this data, with $K = 2$. Suppose we randomly select our initial 2 centroids as datapoints $x = 0$ and $x = 1$ (i.e., the 6th and 7th elements of the dataset above).

1. (**5 points.**) What are the initial assignments of points to each cluster? For example, the answer would be cluster 1 = {2, 7 } if points $x = 2$ and $x = 7$ are in the first cluster. Cluster 1 in the initial assignment should correspond to the centroid value of 0, and Cluster 2 to the centroid value of 1.

   Cluster 1 = { -1, 0 } Cluster 2 = { 2, 6, 7, 4, 1 }

2. (**20 points.**) Using the above as a starting point for the algorithn, for each subsequent iteration of the algorithm, write out below what the centroids are (up to 2 decimal places, e.g., 4.25), for each iteration of the algorithm.

   Also indicate with "yes" or "no" if the algorithm has converged at the end of each iteration. Assume that each iteration consists of centroid calculation, assignment, and convergence check, in that order.

   Note that the algorithm might not need the full 5 iterations indicated below (if not, leave the remaining iterations blank), or might not have converged at all after 5 iterations (in which case you would answer "no" for convergence at each iteration).

   - Iteration 1: centroid 1 $= -0.5$ centroid 2 $= 4$ converged? No
   - Iteration 2: centroid 1 $= 0$ centroid 2 $= 4.75$ converged? No
   - Iteration 3: centroid 1 $= 0.5$ centroid 2 $= 5.67$ converged? Yes
   - Iteration 4: centroid 1 $=$ centroid 2 $=$ converged?
   - Iteration 5: centroid 1 $=$ centroid 2 $=$ converged?

## Problem 4: Decision Trees (30 points.)

1. Consider the table of data given below. The dataset consists of three binary features and a binary label.

| $x_1$ | $x_2$ | $x_3$ | $y$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 |
| 0 | 1 | 0 | 1 |
| 0 | 1 | 1 | 1 |
| 1 | 0 | 0 | 1 |
| 1 | 0 | 0 | 1 |

(a) (**5 points.**) What is the Gini index of the dataset (without doing any splitting)? You can express your answer as a fraction.

$1 - (1/3)^2 - (2/3)^2 = 4/9$

(b) (**5 points.**) Using the Gini index as our splitting criterion, which feature should we split on to create the root of our decision tree?

Split on $x_1$ ($x_2$ also is a valid answer, giving the same Gini index)

(c) (**10 points.**) Based on your choice in part (b), draw the resulting complete decision tree (i.e., a decision tree where examples at each leaf are all from the same class) based on this data. If there are any ties, use the feature with the smaller index.

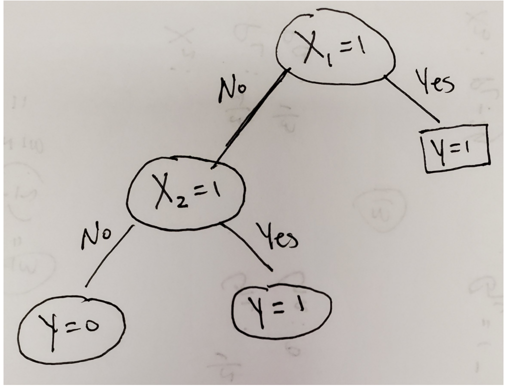
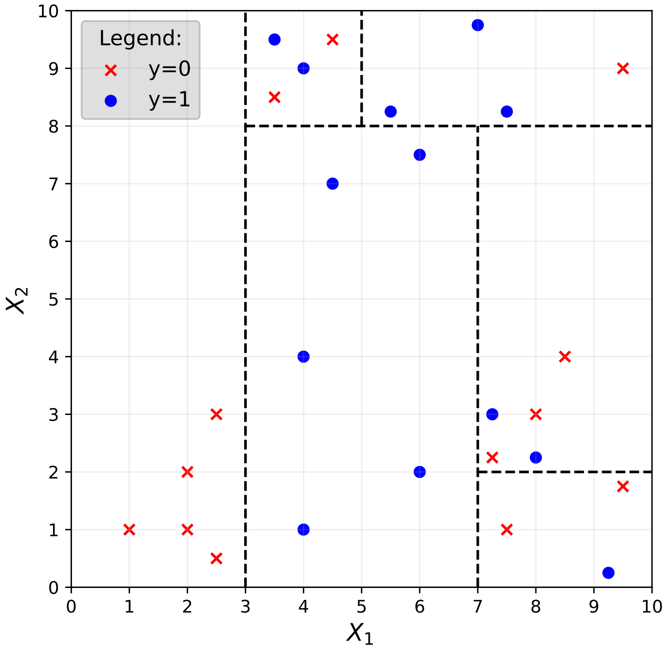
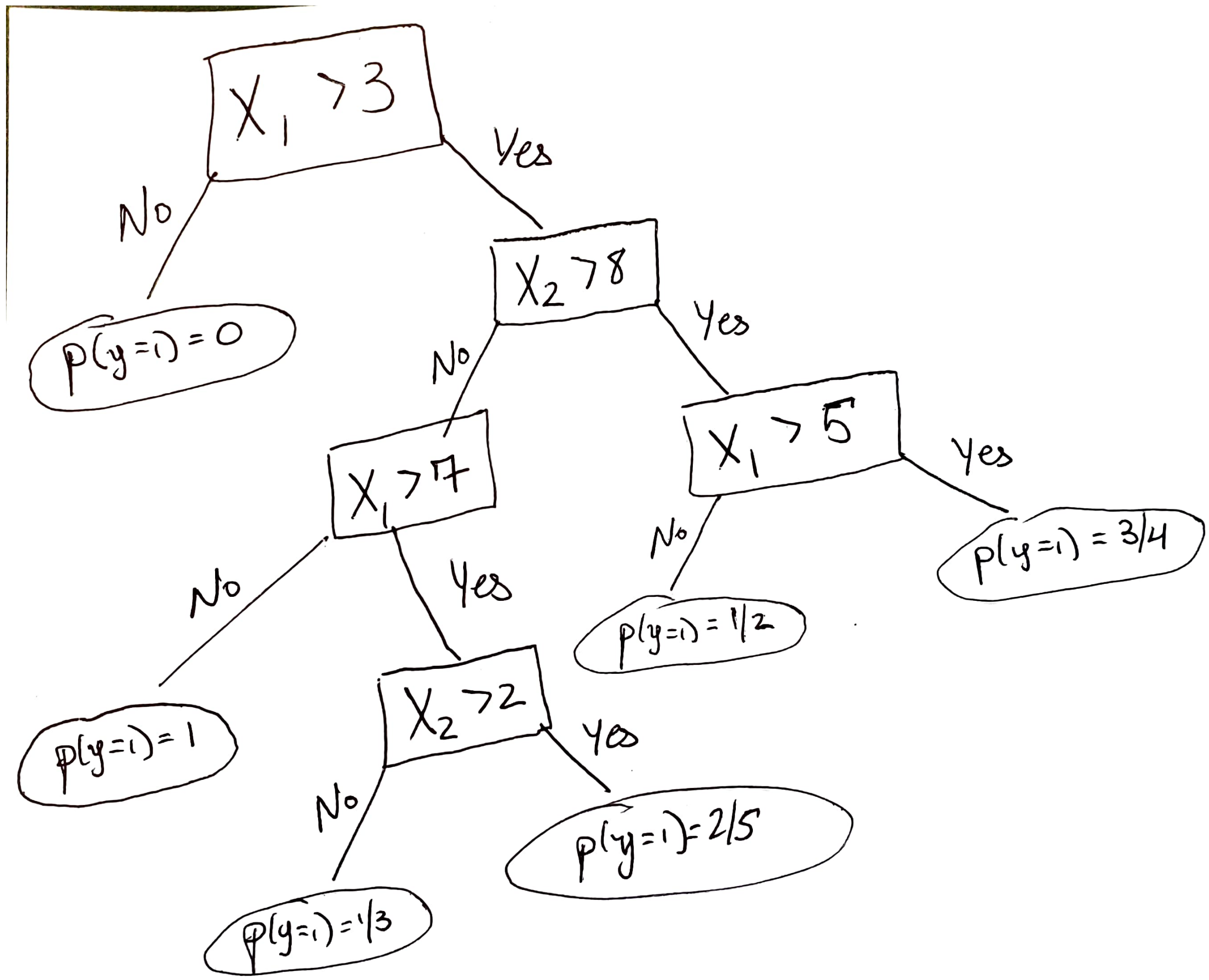


Figure 2: Note: you can also swap the roles of $x_1$ and $x_2$ to get a valid solution.

## Problem 4: continued

2. (**10 points.**) In the following figure, you are given the splits of a decision tree on a dataset with two features. The splits are depicted by dashed lines. Draw the corresponding decision tree and at each leaf node, provide the value for $P(y = 1)$. For your tree, use the convention that all splits are of the form "$X_i > t$?" where $i = 1$ or $i = 2$ is the feature index and $t$ is a threshold.




$X_1 > 3$
No
Yes
$p(y=1) = 0$
$X_2 > 8$
No
Yes
$X_1 > 7$
$X_1 > 5$
No
Yes
No
Yes
$p(y=1) = 1$
$p(y=1) = 1/2$
$p(y=1) = 3/4$
$X_2 > 2$
No
Yes
$p(y=1) = 1/3$
$p(y=1) = 2/5$

## Problem 5: Markov Reward Processes (30 points.)

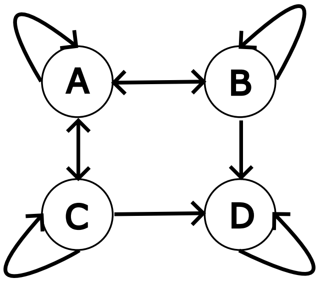


Figure 3:

In Figure 3, you are given a Markov process with four states $(A, B, C, D)$. Each state has the following reward associated with it:

$$R_A = 4 \qquad R_B = 3 \qquad R_C = 2 \qquad R_D = 1$$

1. (**10 points.**) In Table 1, you are given five episodes sampled from this Markov process. Using these samples, estimate the state transition matrix for this Markov process. You should leave your answers as fractions. The rows in your answer indicate the starting state, and the columns indicate the ending state.

| Episode | States |
|---|---|
| 1 | AABDD |
| 2 | CAAAB |
| 3 | CAACD |
| 4 | ACDDD |
| 5 | BAACD |

Table 1:

**ANSWER:**

$$\begin{array}{c|cccc|} & A & B & C & D \\ \hline A & 5/10 & 2/10 & 3/10 & 0 \\ B & 1/2 & 0 & 0 & 1/2 \\ C & 2/5 & 0 & 0 & 3/5 \\ D & 0 & 0 & 0 & 1 \end{array}$$

2. (**10 points.**) Suppose now you are given three more episodes in Table 2 from the Markov chain, all starting from the state A. Using these episodes and the rewards given above, estimate the value $v(A)$ of the state $A$. Use a discount factor of $\gamma = 0.1$. You do not need to simplify your answer.

| Episode | States |
|---|---|
| 1 | AABDD |
| 2 | ACCAB |
| 3 | ACABD |

Table 2:

**ANSWER:** $v(A) = \frac{1}{3}(4.4311 + 4.2243 + 4.2431)$

3. (**10 points.**) Suppose we know that states $A$ and $D$ have values $v(A) = 5$, $v(D) = 2$, and that the probability of transitioning from $C$ to $A$ is $P_{CA} = 0.2$ and the probability of transitioning from $C$ to $D$ is $P_{CD} = 0.5$. Using a discount of $\gamma = 0.1$, compute the value $v(C)$ of the state $C$.

**ANSWER:**

$$v(C) = R_C + \gamma(P_{CA}v(A) + P_{CB}v(B) + P_{CC}v(C) + P_{CD}v(D))$$

$$= 2 + \frac{1}{10}\big(0.2(5) + 0 + 0.3v(C) + 0.5(2)\big)$$

$$= 2.2 + 0.3v(C)$$

$$v(C) = 2.2/0.97$$

# Problem 6: Cross Validation (25 points.)

1. Consider the one-dimensional regression training dataset given in Table 3. We would like to evaluate the performance of the linear regression model on this dataset, where we will measure performance using the mean-squared-error (MSE).

   As we saw in class, one way of assessing model performance on a small dataset is by using cross validation. What is the leave-one-out cross validation MSE of a linear regression model on this data? That is, perform cross-validation for the linear regression model on this dataset with three folds using MSE as the evaluation metric. You do not need to simplify your answer.

| $\mathbf{x}$ | $y$ |
|---|---|
| 5 | 5 |
| 7 | 3 |
| 10 | 15 |

Table 3:

**ANSWER:**

| Train | Test | Line | MSE |
|---|---|---|---|
| $(7,3),(10,15)$ | $(5,5)$ | $y = 4x - 25$ | 100 |
| $(5,5),(10,15)$ | $(7,3)$ | $y = 2x - 5$ | 36 |
| $(5,5),(7,3)$ | $(10,15$ | $y = -x + 10$ | 225 |

Table 4:

Overall cross-validation MSE is $\frac{1}{3}(100 + 36 + 225)$.