Lecture Training 1: Introduction to CS2109S and AI

1. Consider a task in which an agent plays rock-paper-scissors against a human multiple times with no clock. The task environment(s) is/are:
   A. Deterministic
   B. Stochastic
   C. Static
   D. Discrete

   Answer: B, C, D

Explanation:
• The outcome of the game is not fully determined by the current state and action executed by the agent. • Games do not update until the agent shows one of rock, paper, or scissor. • Each game is a distinct, clearly defined action.

2. Consider the PEAS for a robot agent that is playing chess. Which of the following is the actuator?
   A. How close to checkmating the enemy
   B. Chess board
   C. Robotic body and arms
   D. Camera

   Answer: C

3. Every AI agent requires both a performance measure and a utility function.
   A. True
   B. False

   Answer: B (False)

Explanation: Performance measure is needed to assess how well our AI agent is. It is not the case for utility function (Simple-reflex agent doesn't have a utility function).

4. An agent that senses only partial information about the state cannot be perfectly rational.
   A. True
   B. False

Explanation: A rational agent doesn't need to be omniscient. It will try to maximise performance based on available information, which is not necessarily the complete information about the state of the environment

5. A perfectly rational chess-playing agent will always win.
   A. True
   B. False

   Explanation: Make 2 perfectly rational chess-playing agents play against each other, either the game will end with a draw or one of the agents will lose.

Lecture Training 2: Uninformed and Informed Search

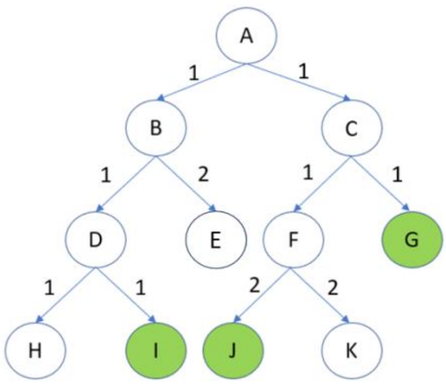
1. Breadth First Search is a special case of Uniform Cost Search.
   A. True
   B. False

   Answer: A (True)

   Explanation:
   When we set the cost = 1 in Uniform Cost Search, it is equivalent to doing Breadth First Search

2. Consider the state space below. A is the start state, and the green-colored states are the goal states. Following the steps taken in the lecture, which of the following is the correct sequence of nodes removed from the frontier and the final path returned by running the **Breadth First Search** algorithm?

   The arrows denote the transition from one state to another and the numbers represent action costs. Note that the transitions are directed; A -> B is a valid transition, but B -> A is not. Moreover, assume that the successor state orderings are such that all ties are broken alphabetically when popped from the queue (i.e. if B and C are the children of A and both are inserted into the queue at the same time, node B is popped first before node C).



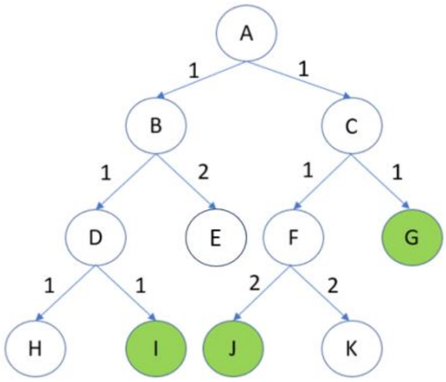
   A. Nodes removed: A – B – C; Final path: A – C – G
   B. Nodes removed: A – B – C – D – E – F – G; Final path: A – C – G
   C. Nodes removed: A – C – B – F – G; Final path: A – C – G
   D. Nodes removed: A – B – D – H – I; Final path: A – B – D – I

   Answer: A / B

   Explanation:
   In BFS, nodes are traversed by depth from left to right, the answer A occurs when the goal test is applied when adding to the queue while B corresponds to the goal test being applied after being popped from the queue

3. Consider the state space below. A is the start state, and the green-colored states are the goal states. Following the steps taken in the lecture, which of the following is the correct sequence of nodes removed from the frontier and the final path returned by running the **Depth First Search** algorithm?

   The arrows denote the transition from one state to another and the numbers represent action costs. Note that the transitions are directed; A -> B is a valid transition, but B -> A is not. Moreover, assume that the successor state orderings are such that all ties are broken alphabetically when popped from the stack (i.e. if B and C are the children of A and both are inserted into the stack at the same time, node B is popped first before node C).



   A. Nodes removed: A – B – D ; Final path: A – B – D – I

   B. Nodes removed: A – C – G; Final path: A – C – G

   C. Nodes removed: A – B – E – D – H – I; Final path: A – B – D – I

   D. Nodes removed: A – B – C – G Final path: A – C – G

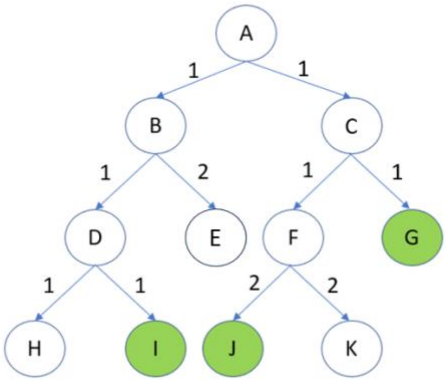
   E. Nodes removed: A – B – D – H – I Final path: A – B – D – I

   Answer: A / C / E

Explanation: In DFS, we traverse to the from left to right and head to the deepest node first. The answer A occurs when we apply the goal test when adding a neighbouring node to the frontier while the answer E corresponds to performing the goal test when popping a node from the frontier.

4. Consider the state space below. A is the start state, and the green-colored states are the goal states. Following the steps taken in the lecture, which of the following is the correct sequence of nodes removed from the frontier and the final path returned by running **Uniform Cost Search** algorithm?

   The arrows denote the transition from one state to another and the numbers represent action costs. Note that the transitions are directed; A -> B is a valid transition, but B -> A is not. Moreover, assume that the successor state orderings are such that all ties are broken alphabetically.



   A. Nodes removed: A – B – C – D – H – I; Final path: A – B – D – I
   B. Nodes removed: A – B – C – D – F – G; Final path: A – C – G
   C. Nodes removed: A – B – D – E; Final path: A – B – E
   D. Nodes removed: A – C – F – G; Final path: A – C – G

   Answer: B

Explanation:
Uniform Cost Search is like BFS except we prioritize lower path costs.

5. Sally, a student studying CS2109S, decided to use the Breadth First Search algorithm to find a solution to solve the 3x3 Rubik's cube. Is Sally doing the right thing? Note: Any Rubik's cube can be solved within 20 moves
   A. Yes
   B. No

   Answer: B (No)

Explanation:
In a 3 by 3 Rubik's cube there are 43,252,003,274,489,856,000 states, making it not feasible to store that many states in the frontier as we will run out of memory.

6. Consider a state space with finite branching factor and infinite depth. Which one of the following will save space for BFS while preserving its completeness?
   A. Apply goal-test when PUSHING a successor state to the frontier
   B. Apply goal-test when POPPING a state from the frontier
   C. Use LIFO stack instead of FIFO queue
   D. The three variants will have similar space usage.

   Answer: A

   Explanation:
   If we check whether a state is the goal BEFORE pushing to frontier queue, then we can immediately return, no need to expand everything else in the frontier before reaching it.

7. What is the overhead if we use Iterative Deepening Search instead of Depth Limited Search (branching factor b = 8 and goal depth d = 4)?
   A. 11.1%
   B. 12.5%
   C. 14.2%
   D. 14.3%

   Answer: D

   Explanation:
   To calculate the overhead of IDS over DLS, we calculate the number of states expanded by each algorithm and calculate $\frac{IDS}{DLS}$

   Number of states expanded by IDS = $8^0 + (8^0 + 8^1) + (8^0 + 8^1 + 8^2) + \cdots (8^0 + 8^1 + 8^2 + 8^3 + 8^4) = 5 \times 8^0 + 4 \times 8^1 + 3 \times 8^2 + 2 \times 8^3 + 8^4 = 5349$
   Number of states expanded by DLS = $8^0 + 8^1 + \cdots + 8^3 + 8^4 = 4681$
   $$\frac{5349}{4681} = 1.1427 \approx 114.27\% \approx 114.3\%$$

8. Iterative Deepening Search is always faster than Depth First Search in terms of time complexity.
   A. True
   B. False

   Answer: B (False)

   Explanation:
   Consider a state space such that each state have only a single successor and the goal node is at depth n. IDS will run in O(n^2) while DFS will run in O(n).

9. True or False: *Uniform Cost Search is a special case of A* search.*
   A. True
   B. False

   Answer: A (True)

Explanation:
Set the heuristic function h(n) = 0

10. True or False: *a heuristic is admissible if and only if it is consistent.*
    A. True
    B. False

    Answer: B (False)

    Explanation

    A consistent heuristic is admissible, but the converse is not true.

    A heuristic is consistent only when $h(n) \leq c(n, a, n') + h(n')$ which is equivalent to $h(n) - h(n') \leq c(n, a, n')$. This means that if the value of the heuristic function going from state n to state n' reduces by more than the actual cost to transit to the new state $h(n) - h(n') > c(n, a, n')$, then the heuristic is not consistent.

    Let's give an example where the heuristic is admissible but not consistent. Suppose there are only three states, A, B and C. Let C be the goal state.

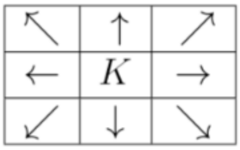
    - $h^*(A) = 7, h^*(B) = 4, h^*(C) = 0$
    - $h(A) = 7, h(B) = 3, h(C) = 0$
    - $h(n)$ is admissible because $h(n) \leq h^*(n)$ for all n.
    - $h(A) - h(B) = 7 - 3 = 4 > h^*(A) - h^*(B) = 7 - 4 = 3$

11. Consider an A* search algorithm which utilises f(n) = a * g(n) + (1 - a) * h(n) (instead of f(n) = g(n) + h(n)) where 0 <= a <= 1. For any value of a, the solution found using this algorithm will be optimal given that h(n) is a consistent heuristic.
    A. True
    B. False

    Answer: B (False)

Explanation:
Setting a = 0 will lead to greedy best-first search algorithm, which is not optimal.

12. Given that a King in a chess game can move one square in any of the eight possible directions (see picture below), the Manhattan Distance heuristic is an admissible heuristic for the problem of moving the King from a tile to another.



A. True

B. False

Explanation:
The actual cost of moving diagonally is 1, while using Manhattan distance, the cost would be 2, so the Manhattan distance overestimates the actual cost.

13. Consider the 8-puzzle problem and 2 heuristics, h1 = number of misplaced tiles and h2 = total Manhattan Distance. Which of the following statements is/are true?

Notes: The empty space is not considered a tile. By total Manhattan Distance, we refer to the sum of the Manhattan Distances between the current position of each tile and the position the tile should be in. As an illustrating example, consider the following diagram where h1 = 4 and h2 = 5.

Current state

Goal state

A. (h1 + h2)/2 is admissible

B. h1 + h2 is admissible

C. (h1+h2)/2 dominates h1

D. (h1+h2)/2 dominates h2

Answer: A and C

Explanation:
We know h2 is admissible because each tile, at the very least, needs to be moved the Manhattan distance to its goal state, and each action can only move one tile by one space. We know h1 is admissible because each action can only move one tile. Thus, the average whose value lies between the two heuristics is admissible as well We know that h2 dominates h1 since every misplaced tile at the very least contributes a Manhattan distance of 1. Thus, h2 >= h1 and (h1+h2)/2 >= h1

14. Consider 2 admissible heuristics, h1 and h2, and 2 inadmissible heuristics, h3 and h4. Which of the following heuristic(s) will always be admissible?
    A. f(x) = max(h1(x), h2(x))
    B. f(x) = min(h2(x), h4(x))
    C. f(x) = max(h1(x), h3(x))
    D. f(x) = min(h3(x), h4(x))

    Answer: A and B

    Explanation:
    Since h1 and h2 are admissible, neither one of them will ever over-estimate the true cost, thus their max will not over-estimate as well
    Since h2 is admissible, it will never over-estimate the true cost, thus no matter the value of h4, the min of the two functions will also never over-estimate the true cost

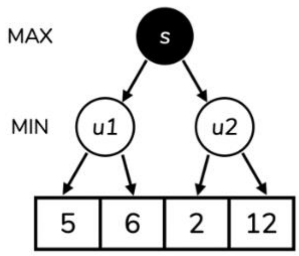
Lecture Training 3: Informed, Local and Adverserial Search

1. True or False: *The minimax algorithm is optimal against both optimal and non-optimal MIN players*.
   A. True
   B. False

   Answer: B (False)

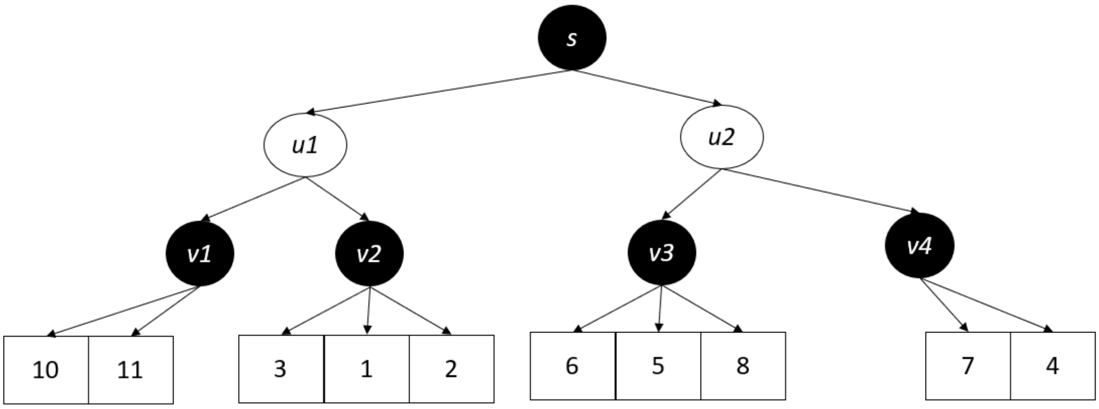
Explanation:
The minimax algorithm is not optimal against a non-optimal MIN player.

In the following example:



The minimax algorithm will pick u1. However, against an extremely non-optimal MIN player that maximizes the score instead, for example, picking u2 can actually lead to a higher score of 12.

2. Assume that we iterate over nodes from **right to left**; which ARCS are pruned by α-β pruning? **Note**: Black node represents MAX player, white node represents MIN player



A. Arc that connects 5 and v3

B. Arc that connects 6 and v3

C. Arc that connects 3 and v2

D. Arc that connects u1 and v1

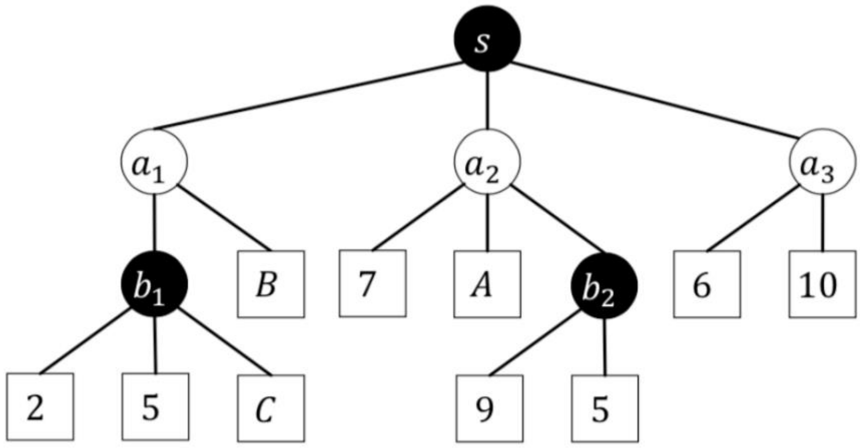
E. Arc that connects 7 and v4

Answer: A, B, D

Explanation:

- 5-v3 is pruned because u2 already has a better option of 7, so after encountering 8, it knows that nothing in this subtree can give a better result.
- 6-v3 is pruned because u2 already has a better option of 7, so after encountering 8, it knows that nothing in this subtree can give a better result.
- u1-v1 is pruned because s already has a better option of 7, so after encountering 3, it knows that nothing in this subtree can give a better result.

For questions 3 to 5, refer to the following minimax tree.



3. Assume that we iterate over nodes from right to left and that each node has an INTEGER value; find the largest range* of values of A that ensures no arcs are pruned by α-β pruning. Choose the best option. For example, A ≥ 0 would represent a larger range than A ≥ 1.
   A. $A \geq 6$
   B. $\boldsymbol{A \geq 7}$
   C. $A \geq 8$
   D. $A \geq 9$

   Answer: B

Explanation:
If A is less than or equal to 6 the MAX player at s would have an equivalent or better option by choosing a3, so there would be no need to continue searching this subtree.

4. Assume that we iterate over nodes from right to left and that each node has an INTEGER value; find the largest range* of values of B that ensures no arcs are pruned by α-β pruning. Choose the best option. For example, B ≥ 0 would represent a larger range than B ≥ 1.
   A. $B \geq 6$
   B. $B \geq 7$
   C. $\boldsymbol{B \geq 8}$
   D. $B \geq 9$

   Answer: C

Explanation:
If B is less than or equal to 7 the MAX player at s would have an equivalent or better option by choosing a2, so there would be no need to continue searching this subtree

5. Assume that we iterate over nodes from right to left and that each node has an INTEGER value; find the largest range* of values of C to ensure no arcs are pruned by α-β pruning. Choose the best option. For example, C ≤ 11 would represent a larger range than C ≤ 10.

   A. $C < 6$
   B. $C < 7$
   C. $\boldsymbol{C < B}$
   D. $C \leq B$
   E. $C \leq A$

   Answer: C

Explanation: If C is greater than or equal to B, the MIN player at a1 would have an equivalent or better option by choosing B, so there would be no need to continue searching this subtree

Lecture Training 4: Intro to Machine Learning & Decision Trees

1. Given the following confusion matrix for a binary classification problem:

	Actual Positive	**Actual Negative**
Predicted Positive	50	10
Predicted Negative	5	35

Which of the following metrics is the highest for this confusion matrix?

A. Accuracy

B. Precision

C. F1 Score

D. Recall

Answer: D

Explanation:

Accuracy = $\frac{50+35}{50+10+5+35} = \frac{17}{20} = 0.85$

Precision = $\frac{50}{50+10} = \frac{5}{6} = 0.8333$

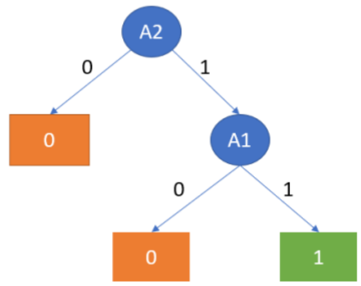
F1 Score = $\frac{2}{\frac{6}{5}+\frac{11}{10}} = \frac{2}{\frac{23}{10}} = \frac{20}{23} = 0.8695$

Recall = $\frac{50}{50+5} = \frac{10}{11} = 0.9091$

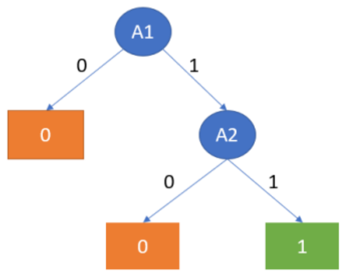
2. Consider the data below. Which of the following is the tree produced by the DTL algorithm in the lecture?

Example	A_1	A_2	A_3	**Output** y
x_1	1	0	0	0
x_2	1	0	1	0
x_3	0	1	0	0
x_4	1	1	1	1
x_5	1	1	0	1

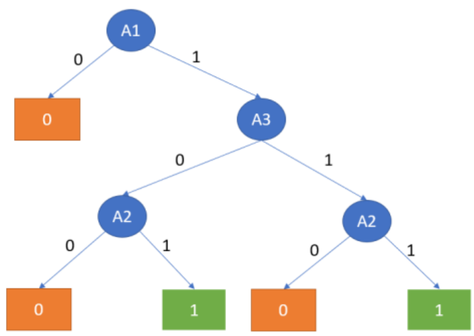
A.



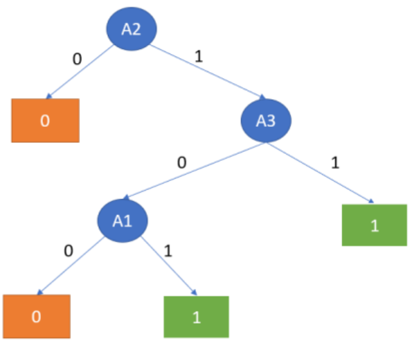
B.



C.



D.



Answer: A

Explanation:

$Initial\ entropy\ =\ I(\frac{2}{5},\frac{3}{5})\ =\ 0.97095$

$remainder(A1)\ =\frac{4}{5}\times I\left(\frac{1}{2},\frac{1}{2}\right)+\frac{1}{5}\times I(0,1)\ =\ 0.8$

$remainder(A2) =\frac{2}{5}\times I(0,1)+\frac{3}{5}\times I(\frac{2}{3},\frac{1}{3})\ =\ 0.55$

$remainder(A3) =\frac{2}{5}\times I(0.5,0.5)+\frac{3}{5}\ \times\ I(\frac{1}{3},\frac{2}{3})\ =\ 0.95$

Since maximum IG = 0.97095 - remainder(A2), first split is on A2.
Since on the A2 = 0, all samples have y = 0, the 0 branch is assigned the classification of 0.

Entropy at 1 split after A2 = $I\left(\frac{1}{3},\frac{2}{3}\right) = 0.9183$

remainder(A1) = $\frac{1}{3}\times I(0,1)+\frac{2}{3}\times I(1,0)\ =\ 0$

remainder(A3) = $\frac{2}{3}\ \times I\left(\frac{1}{2},\frac{1}{2}\right)+\frac{1}{3}\times I(1,0)\ =\ 2/3$

Since remainder(A1) is 0, IG is maximized if the split is on A1

3. A new food stall has opened and sells 3 items: Burgers, Fish & Chips, and Fried Chicken. To better plan for how many of each menu item they need to prepare, the stall attempts to make a model to predict what menu item a student will buy. After collecting orders from 160 students on a certain day, the stall noted that the students ordered 70 burgers, 60 fish & chips, and 30 fried chicken. What is the entropy of these 160 examples?

   The formula for calculating the entropy of a variable is given below:

$$I\big(P(v_1),\dots,P(v_n)\big) = -\sum_{i=1}^{n} P(v_i)\log_2(P(v_i))$$

   A. $Entropy\ <\ 0.8$
   B. $0.8 \leq Entropy < 1$
   C. $1 \leq Entropy < 1.2$
   D. $1.2 \leq Entropy < 1.4$
   E. $\boldsymbol{1.4 \leq Entropy < 1.6}$
   F. $1.6 \leq Entropy$

   Answer: E

Explanation:

$$I\left(\frac{70}{160},\frac{60}{160},\frac{30}{160}\right) = -\frac{70}{160}\log_2\frac{70}{160}-\frac{60}{160}\log_2\frac{60}{160}-\frac{30}{160}\log_2\frac{30}{160} = 1.505$$

4. A new food stall has opened and sells 3 items: Burgers, Fish & Chips, and Fried Chicken. To better plan for how many of each menu item they need to prepare, the stall attempts to make a model to predict what menu item a student will buy. After collecting orders from 160 students on a certain day, the stall noted that students ordered 70 burgers, 60 fish & chips, and 30 fried chicken. The stall also logged which year of study the student is currently in right now in the following table:

	Burger	Fish & Chips	Fried Chicken	Total
Year 1	20	20	0	40
Year 2	30	10	30	70
Year 3	20	30	0	50
Total	70	60	30	160

The stall tries to use which year the student is in as an attribute to predict which dish the student will order. What's the Information Gain (IG) of this attribute?

Recall from the slides that IG(***A***) = Information Content of this node – remainder(***A***)

If an attribute A divides a sample S into v sets: $S_1, S_2, \dots, S_v$
$remainder(A) = \sum_{i=1}^{v} \frac{|S_i|}{|S|} I(S_i)$, where $|X|$ denotes the size of set X.

As an example $S_{Year\ 3}$ contains 20 burger orders, 30 fish & chips orders and 0 fried chicken orders.

A. $IG < 0.3$
B. $\mathbf{0.3 \leq IG < 0.4}$
C. $0.4 \leq IG < 0.5$
D. $0.5 \leq IG < 0.6$
E. $0.6 \leq IG$

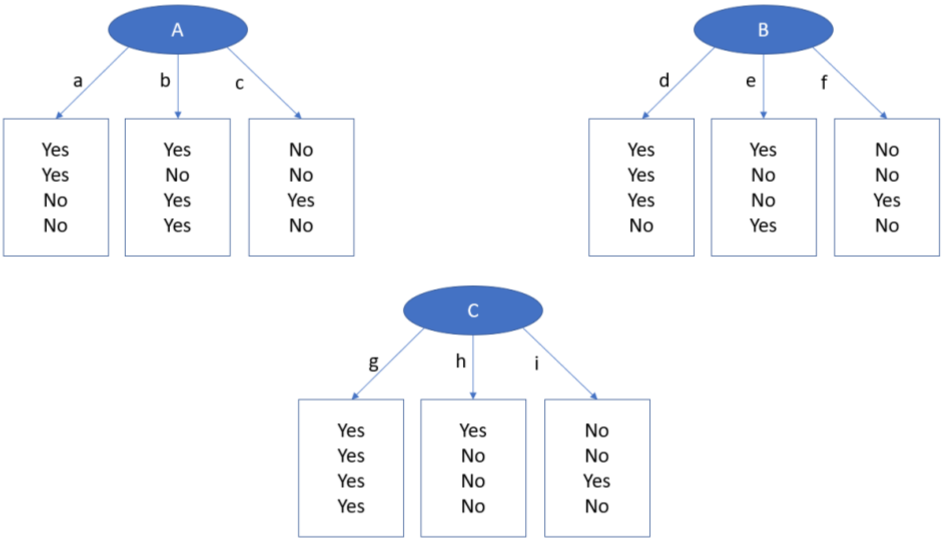
Answer: B

Explanation:
$IG = Information\ Content - remainder(A)$
From the previous question, Information Content ≈ 1.505

$$remainder(A) = \frac{4}{16} I\left(\frac{2}{4}, \frac{2}{4}, 0\right) - \frac{7}{16} I\left(\frac{3}{7}, \frac{1}{7}, \frac{3}{7}\right) - \frac{5}{16} I\left(\frac{2}{5}, \frac{3}{5}, 0\right) \approx 1.187$$

$$IG = 1.505 - 1.187 = 0.318$$

5. Which of the following feature will lead to the largest information gain?



A. Feature A
B. Feature B
C. Feature C
D. All features lead to the same information gain.

Answer: C

Explanation:
Intuitively, it's better to choose an attribute that splits the examples into subsets where most/all of them have the same target.

Mathematically, we can verify this claim by computing the information gain of every feature:

- $IG(A) = I\left(\frac{6}{12}, \frac{6}{12}\right) - remainder(A)$
  $= 1 - \left(\frac{1}{3}\right) * I\left(\frac{1}{2}, \frac{1}{2}\right) - \left(\frac{1}{3}\right) * I\left(\frac{1}{4}, \frac{3}{4}\right) - \left(\frac{1}{3}\right) * I\left(\frac{1}{4}, \frac{3}{4}\right) \approx 1 - 0.874 = 0.126$
- $IG(B) = I\left(\frac{6}{12}, \frac{6}{12}\right) - remainder(B)$
  $= 1 - \left(\frac{1}{3}\right) * I\left(\frac{1}{4}, \frac{3}{4}\right) - \left(\frac{1}{3}\right) * I\left(\frac{1}{2}, \frac{1}{2}\right) - \left(\frac{1}{3}\right) * I\left(\frac{1}{4}, \frac{3}{4}\right) \approx 1 - 0.874 = 0.126$
- $IG(C) = I\left(\frac{6}{12}, \frac{6}{12}\right) - remainder(C)$
  $= 1 - \left(\frac{1}{3}\right) * I(0, 1) - \left(\frac{1}{3}\right) * I\left(\frac{1}{4}, \frac{3}{4}\right) - \left(\frac{1}{3}\right) * I\left(\frac{1}{4}, \frac{3}{4}\right) \approx 1 - 0.541 = 0.459$

6. Suppose you create a training set from a decision tree T_1. Then, you create a decision-tree T_2 that learns from the training set. T_2 will always be the same as the original tree T_1.
   A. True
   B. False

   Answer: B (False)

Explanation: The algorithm might return a different tree that is logically equivalent to T_1

7. Suppose we split our data into a training set and a testing set. While we cannot guarantee its performance on the testing test, we can always find at least one decision tree that can perfectly label every example in the training set, no matter the amount of data and attributes we have.
   A. True
   B. False

   Answer: B (False)

Explanation: The data may not be consistent. Suppose we are to train a decision tree that determines whether a student passes or fails, but the only attribute we have is gender. There is no way to cleanly split using the training set with this attribute.

8. Suppose we're training a decision tree with possible labels True and False. The more attributes we have in the final decision tree, the better our decision tree will be at labelling new data.
   A. True
   B. False

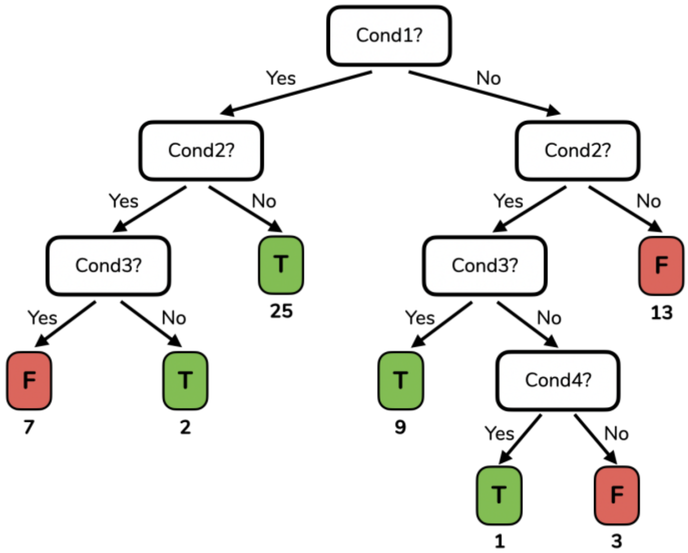
   Answer: B (False)

Explanation: Using too many attributes, thus building a complex hypothesis, makes our model susceptible to overfitting, resulting in being worse at predicting new data.

9. Which of the following attribute(s) can a Decision Tree model use?
   A. Binary valued attributes
   B. Discrete valued attributes
   C. Continuous valued attributes
   D. Binned / Discretized continuous valued attributes

   Answer: A, B and D

10. For this decision tree, which pruning method would leave fewer branches?



A. Pruning with max depth = 2

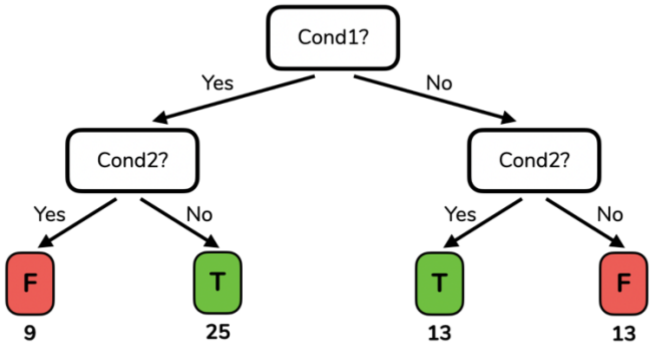
B. Pruning with min-sample = 10

C. Both pruning methods will leave the same number of branches

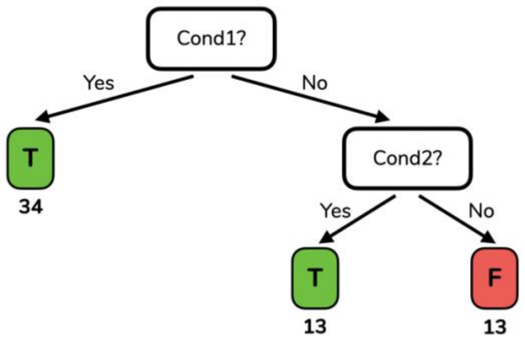
Answer: B

Explanation:

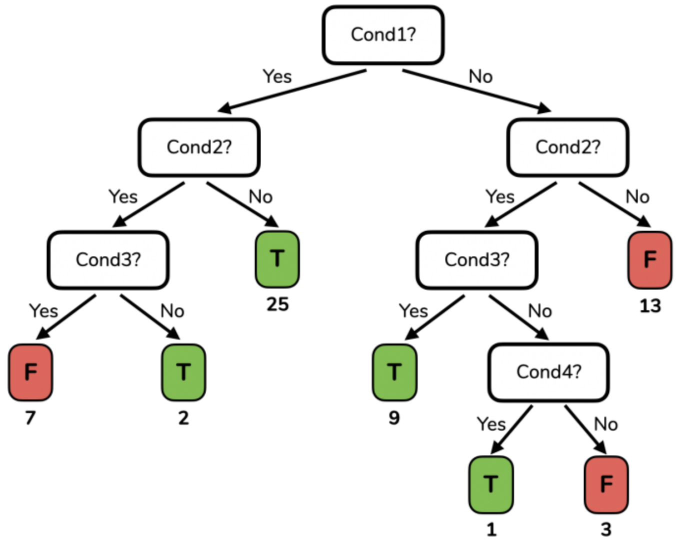
Decision tree after pruning with Max-depth = 2:



Decision tree after pruning with Min-sample = 10:



11. Assume that before pruning, this decision tree correctly labels every example as T or F. If we prune away the **leaf nodes of the attribute** that gives the least Information Gain, how many samples will be mislabelled by the new tree?



A. 1
B. 2
C. 3
D. None of the above

Answer: B

Explanation:
There are only two candidate nodes that we can consider for removal. The left node which decides between 7 False and 2 True or the right node that decides between 3 False and 1 True.

IG(Left Node) = I(2/9, 7/9) - remainder(Left Node) ≈ 0.764
IG(Right Node) = I(1/4, 3/4) - remainder(Right Node) ≈ 0.811

Note that remainder(Left Node) = 0 and remainder(Right Node) = 0.
Since the left node has less information gain, we prune away the left node to 9 False. Hence the 2 True nodes will be mislabelled.

Lecture Training 5: Linear Regression

1. Which of the following would you apply unsupervised learning to? (Select all that apply)
   A. Predicting Housing Prices in Singapore
   B. Given a large dataset of medical records of patients who have diabetes, try to learn whether there might be different clusters of such patients for which we might tailor different treatments.
   C. Given medical patients respond to an experimental drug, discover whether there are different categories or "types" of patients in terms of how they react to the medicine.
   D. Check whether an email is spam or not

   Answer: C and D

Explanation:
Unsupervised Learning tasks often have no correct answer.

2. Which of the following would you treat as classification problem? (Select all that apply)
   A. Predicting Housing Prices in Singapore
   B. Given a large dataset of medical records of patients who have diabetes, try to learn whether there might be different clusters of such patients for which we might tailor different treatments.
   C. Given medical patients respond to an experimental drug, discover whether there are different categories or "types" of patients in terms of how they react to the medicine.
   D. Check whether an email is spam or not

   Answer: D

3. Which of the following would you treat as regression problem? (Select all that apply)
   A. Predicting Housing Prices in Singapore
   B. Given a large dataset of medical records of patients who have diabetes, try to learn whether there might be different clusters of such patients for which we might tailor different treatments.
   C. Given medical patients respond to an experimental drug, discover whether there are different categories or "types" of patients in terms of how they react to the medicine.
   D. Check whether an email is spam or not

   Answer: A

4. What is the maximum degree of polynomial one should use to fit a data set of 10 points?
   A. 20
   B. 1
   C. 10
   D. 9
   E. 11

   Answer: C

   Explanation:
   The maximum degree of polynomial needed to fit any set of n points is n-1. If we use more than n-1, we will overfit.

5. What is the value of w?

$$wX^T = y^T$$
$$X = \begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix}$$
$$y = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$$

   A. $w = \begin{bmatrix} -2 \\ 2 \end{bmatrix}$
   B. $\boldsymbol{w} = [\mathbf{-2} \;\; \mathbf{2}]$
   C. $w = [2 \;\; -2]$
   D. $w = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$

   Answer: B

   Explanation:

$$w \cdot X^T = w \cdot \begin{bmatrix} 1 & 3 \\ 2 & 5 \end{bmatrix} = [2 \;\; 4]$$
$$[-2 \;\; 2] \cdot \begin{bmatrix} 1 & 3 \\ 2 & 5 \end{bmatrix} = [-2+4. \;\; -6+10] = [2 \;\; 4]$$

6. Consider a linear regression problem. We have a training set and we managed to find w1 and w0 such that mean squared error of the training set is 0. Which of the statements must be true? (Check all that apply)
   A. w1 and w0 must be 0
   B. y = 0 for each training data
   C. The training data can be fit perfectly on a straight line

   Answer: C

   Explanation:
   A loss of 0 means every data perfectly fits on the line y = w1 x + w0.

7. Which of the following is true regarding gradient descent? (Select all that apply)
   A. Setting the learning rate as small as possible is not harmful. It can speed up the convergence rate.
   B. Gradient descent algorithm will always lead to global minimum
   C. If w1 and w2 are initialized such that w1 = w2, after one step of gradient descent, w1 = w2 still holds due to symmetry
   D. None of the above

   Answer: D

   Explanation:
   Option 1: A smaller learning rate takes longer rather than faster.
   Option 2: Gradient Descent doesn't guarantee to lead to the global minimum. For example, it can get stuck in a local minimum.
   Option 3: They likely won't have the same value since their gradients should be different.

8. Consider the three images below. Which one of the following statements is true?

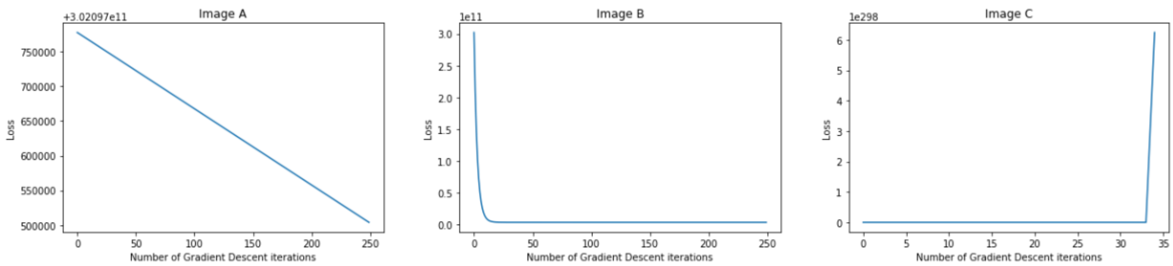


	Image A	Image B	Image C
A.	Learning Rate too large	Learning Rate is optimal	Learning Rate too small
B.	**Learning Rate too small**	**Learning Rate is optimal**	**Learning Rate too large**
C.	Learning Rate is optimal	Learning Rate too small	Learning Rate too large
D.	None of the above		

Answer: B

9. Suppose we have a training set with m samples: $x^{(1)}$, $x^{(2)}$, $x^{(3)}$, ..., $x^{(m)}$. In batch gradient descent we use all m samples with each update, while in stochastic gradient descent, we only use $x^{(1)}$ for the first update, $x^{(2)}$ for the second update,... and so on.
   A. True
   B. False

   Answer: B (False)

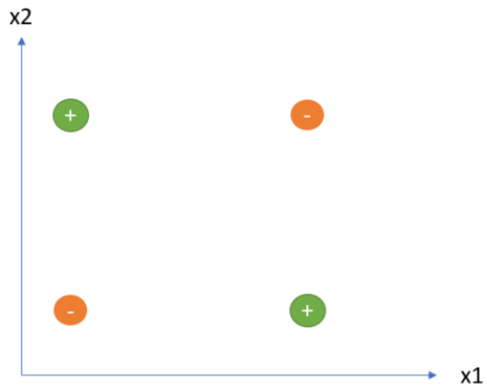
Explanation
"Stochastic" means we want some randomness to help escape local minima. The order in which we use the samples should be randomly decided.

Lecture Training 6: Logistic Regression

1. Logistic regression needs binning/discretization to handle continuous-valued attributes.
   A. True, it cannot handle raw continuous values
   B. False, it can handle continuous values directly without binning

   Answer: B

2. The logistic regression can classify the data below perfectly WITHOUT needing new features (e.g. x1 * x2, x1 * x1, x2 * x2, etc)



   A. True
   B. False

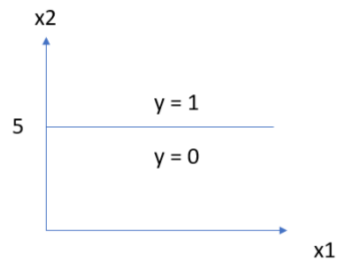
   Answer: B

Explanation: The data is not linearly separable.

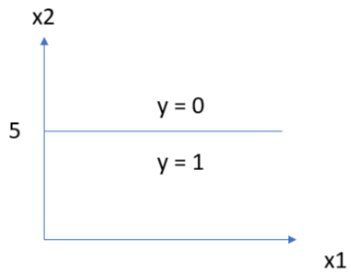
3. Suppose you have a logistic classifier $y = h_w(x) = g(w_0 + w_1 * x1 + w_2 * x2)$. Assume $w_0 = 10$, $w_1 = 0$, and $w_2 = -2$. Which of the following represents the decision boundary found by the classifier?

   g is a threshold function that takes a value x and returns:
   - 1 if x > 0
   - 0 otherwise

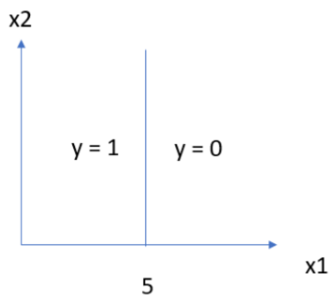
   A.



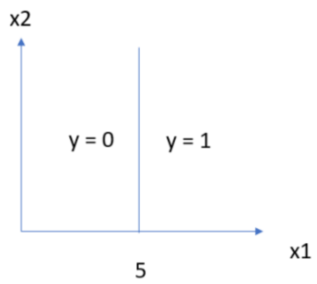
   B.



   C.



   D.



Answer: B

Explanation: The threshold function g returns 1 when 10 – 2 * x2 > 0 or when x2 < 5

4. Suppose you have a trained logistic classifier, and it outputs on a new example a prediction $h_w(x) = 0.4$. Which of the following is/are true? Select all that apply.
   A. $\boldsymbol{P(\, y = 1 \mid x, w\,) = 0.4}$
   B. $P(\, y = 0 \mid x, w\,) = 0.4$
   C. $P(\, y = 1 \mid x, w\,) = 0.6$
   D. $\boldsymbol{P(\, y = 0 \mid x, w\,) = 0.6}$

   Answer: A and D

Explanation:
$h_w(x)$ can be seen as the probability of the output being labelled as 1 given the input.

5. Suppose you have trained a logistic classifier. Consider a new data point: (x, y) with $y = 1$. The classifier returns $h_w(x) = 0.4$ for the data point. What is the value of Binary Cross Entropy? (The base of the log is e)
   A. 0.511
   B. 0.693
   C. 0.799
   D. 0.916

   Answer: D

   Explanation:
   $$\log_e 0.4 - (1-1)\log_e 0.6 = 0.916$$

6. Logistic regression can be used on multi-class classification tasks.
   A. True
   B. False

   Answer: A

7. Why do we do feature scaling?
   A. It speeds up solving for *w* using the Normal Equation.
   B. It prevents the matrix X^T X (used in the normal equation) from being non-invertible.
   C. It speeds up gradient descent by making it require fewer iterations to get to a good solution.
   D. In classification tasks, it helps make our data linearly separable.
   E. It transforms a multi-class classification problem into a binary classification problem.

   Answer: C

8. We're trying to decide which hypothesis to use. Models 1 through 4 are different candidates (linear, non-linear, multivariate...).

   Given the errors, which model should we pick?

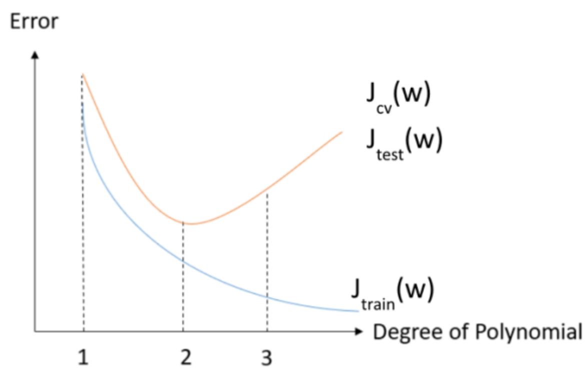
Model	Training Error	Validation Error	Testing Error
1	1500	1750	2500
2	1700	1900	1800
3	1000	2000	1500
4	2000	2100	2100

   A. Model 1
   B. Model 2
   C. Model 3
   D. Model 4

   Answer: A

   Explanation:
   We should pick the model that minimises cross-validation error. Usually, we wouldn't even run the other 3 models on the testing set.

9. Consider the graph below. Which of the following is/are true? (Select all that apply)



   A. Degree = 1 will cause the model to underfit the data
   B. Degree = 1 will cause the model to overfit the data
   C. Degree = 2 will cause the model to underfit the data
   D. Degree = 2 will cause the model to overfit the data
   E. Degree = 3 will cause the model to underfit the data
   F. Degree = 3 will cause the model to overfit the data

   Answer: A and F

   Explanation:
   At degree 1, the train error rate is high.
   At degree 3, there is a high variance between the train and the test error rates.

10. For the machine learning algorithms we've learned, which of the following is/are hyperparameters?
    A. Learning rate of gradient descent
    B. Max depth of a decision tree
    C. Min sample of a decision tree
    D. Loss function of a regression model

    Answer: A, B, C and D

Lecture Training 7: Support Vector Machines

1. Which of the following is/are true regarding regularization? (Select all that apply)

   The regression cost function with regularization:

$$J(w) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_w\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{i=1}^{n} w_i^2\right]$$

   A. Using too large λ can cause your model to **overfit**
   B. Using too large λ can cause your model to underfit
   C. Consider a classification problem. Adding regularization may cause your model to misclassify some training examples.

   Answer: B and C

Explanation: Refer to lecture 7 slide 16 and 17

2. Suppose you train a linear regression model twice with two different lambda values. You get two different sets of w values: w_A and w_B. However, you forgot which w values correspond to which λ value. Which w do you think corresponds to λ_2?

   λ_1= 0 and λ_2= 1. w_A = [80.12, 72.23] and w_B = [13.22, 10.55]

   The regression cost function with regularization:

$$J(w) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_w\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{i=1}^{n} w_i^2\right]$$

   A. W_A
   B. W_B

   Answer: B

Explanation: Regularization discourages large weights.

3. It is important to perform feature scaling before we train a linear regression model with regularization.
   A. True
   B. False

   Answer: A

   Explanation: Performing feature scaling is critical when using regularization, as it ensures that the penalty is applied uniformly across all features, stabilizes the optimization process, and prevents any one feature from disproportionately influencing the model.

4. In SVM, what is the margin?
   A. The distance between the two closest data points of different classes
   B. The distance between the decision boundary and the closest data points of both classes
   C. The distance between the decision boundary and the farthest data point of any class
   D. The distance between the two farthest data points of different classes

   Answer: B

5. Suppose you have trained a SVM model on a specific dataset. What should we do if we suspect that the SVM model is underfitting the dataset?

   The SVM solves:

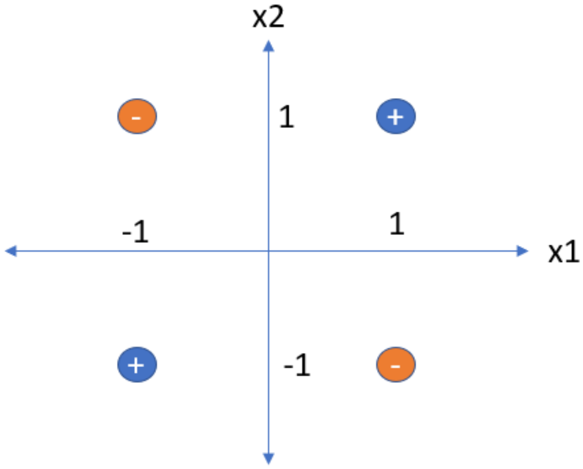
$$\min_{w} C\left[\sum_{i=1}^{m} y^{(i)} cost_1(w^T x) + \left(1 - y^{(i)}\right) cost_0(w^T x)\right] + \frac{1}{2}\sum_{i=1}^{n} w_i^2$$

   A. Increase the value of C
   B. Decrease the value of C
   C. Changing C will not change anything

   Answer: A

   Explanation: Higher C means less slack, which means we reduce the rate of allowance for misclassifications. By doing so, we allow the SVM model to better fit towards a dataset. (Refer to slide 49)

6. It is possible to construct a support vector machine that computes the XOR function consisting of 4 data:
   1. x1 = 1, x2 = 1, and label = +
   2. x1 = 1, x2 = -1, and label = -
   3. x1 = -1, x2 = 1, label = -
   4. x1 = -1, x2 = -1, label = +



   A. True
   B. False

   Answer: A

   Explanation: Use a non-linear kernel

7. In general, when training a SVM with kernel, what does the kernel trick do?
   A. It maps the data from the input space to a higher-dimensional feature space
   B. It maps the data from the input space to a lower-dimensional feature space.
   C. It performs regularization to prevent overfitting

   Answer: A

8. Which of the following is/are true regarding Gaussian transformed features? (Select all that apply)

   Gaussian transformed features can be written as, where l_j are some numbers.

   $$f(x, l_j) = e^{-\frac{|x-l_j|^2}{2\sigma^2}}$$

   A. It is important to perform feature scaling before using the Gaussian Kernel

   B. The maximum value of the Gaussian transformed features is 1

   C. The maximum value of the Gaussian transformed features is e

   Answer: A and B

   Explanation:
   A: The Gaussian kernel in SVM uses a similarity function that measures the distances between a pair of examples. If features take a different range of values Euclidean distance will be dominated by the features that have a huge range of values and consequently, will ignore other features whose range of values are small. Thus, feature scaling has to be performed before using the Gaussian kernel.
   B: RBF kernel decreases with distance and ranges between zero (in the infinite-distance limit) and one.

9. Assume data with one-dimensional feature x. Suppose you have trained a soft-margin SVM model WITH Gaussian transformed features on a specific dataset. What should we do if we suspect that the SVM model is overfitting the dataset?

   With $y^{(i)}$ being 0 or 1, the soft-margin SVM with Gaussian transformed features solves: (Important: $cost_0$ equals exactly $cost_{-1}$ defined in class. It is just a notation difference.)

$$\min_{w} C\left[\sum_{i=1}^{m} y^{(i)} cost_1\left(w^T f^{(i)}\right) + \left(1 - y^{(i)}\right) cost_0\left(w^T f^{(i)}\right)\right] + \frac{1}{2}\sum_{i=1}^{n} w_i^2$$

   Gaussian transformed features can be written as follows, where l_j are some predefined numbers:

$$f\left(x, l_j\right) = e^{-\frac{|x-l_j|^2}{2\sigma^2}}$$

   A. Increase σ^2
   B. Decrease σ^2
   C. Changing σ^2 will not change anything

   Answer: A

   Explanation: As σ^2 increases, the bandwidth of the kernel function increases.
   If the sigma value is very small, then the decision boundary is highly non-linear. On the other hand, if the sigma value is large, then the decision boundary tends to be linear.

   Read More: https://towardsdatascience.com/support-vector-machines-under-the-hood-c609e57a4b09

10. Which of the following is true regarding Multi-Class Classification (num_of_class > 2) using SVM?
   A. We cannot perform multi-class classification using SVM
   B. For multi-class classification with num_of_class = K, we need to train K - 1 SVM models
   C. For multi-class classification with num_of_class = K, we need to train K SVM models
   D. For multi-class classification with num_of_class = K, we need to train K + 1 SVM models

   Answer: C

   Explanation: One vs All requires K classifiers for K classes.

Lecture Training 8: Introduction to Neural Networks

1. Which of the following is/are non-linear function(s)? Select all that apply
   A. $f(x) = x$
   B. $\boldsymbol{f(x) = \log x}$
   C. $f(x) = ax + b$ where a, b are real numbers
   D. $\boldsymbol{f(x) = e^x}$

   Answer: B and D

2. Assume we have number of iterations = infinity. Perceptron Learning Algorithm (PLA) will always converge (no more weight update).
   A. True
   B. False

   Answer: B

   Explanation: If the data is not linearly separable, PLA will not converge (Slide 16)

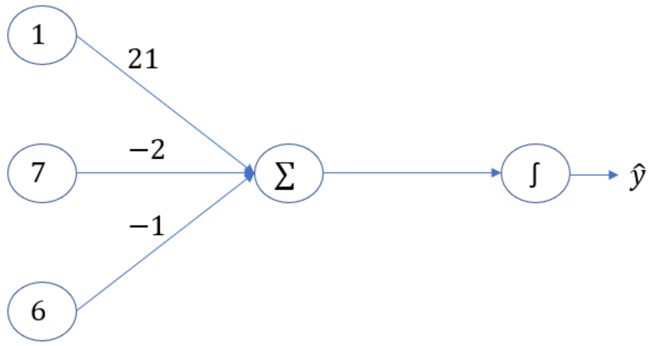
3. I was given a linearly separable data set and I used it to train my classification model. After the training is done, I found a new data point. Fortunately, this data point is classified correctly by my trained model. Furthermore, this data point is far away from the decision boundary.

   After adding this new data point to my training data and re-training my model using the weights from the old model as the initial weight to the new model (if applicable), I found that its decision boundary has shifted.

   Out of the following classification algorithms we have learned in CS2109S, which one(s) did I pick for my model? Select all likely answers.
   A. Perceptron learning algorithm
   B. Logistic regression
   C. Linear SVM

   Answer: B

4. What is the output of the perceptron below? Assume the non-linear activation function is the step function sgn(x) from lecture.



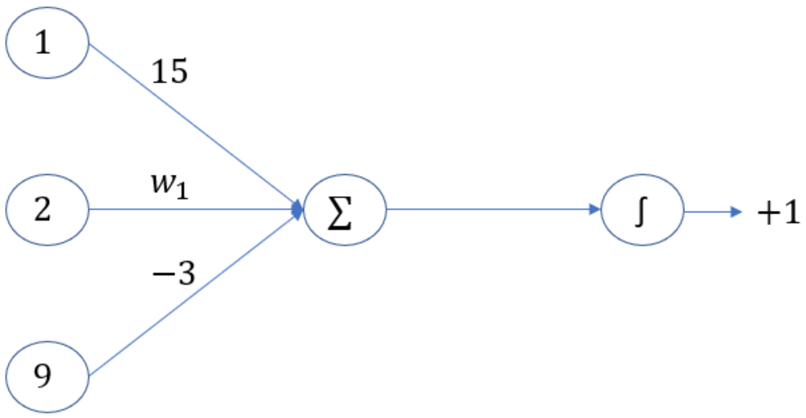
A. $\boldsymbol{\hat{y} = +1}$
B. $\hat{y} = -1$

Answer: A

Explanation:

$$(21 \times 1) + (-2 \times 7) + (-1 \times 6) = 1$$
$$sgn(1) = +1$$

5. Which of the following is/are valid value(s) for w_1? Select all that apply



Assume the non-linear activation function is the step function sgn(x).

A. $w_1 = 4$
B. $w_2 = 5$
C. $\boldsymbol{w_3 = 6}$
D. $\boldsymbol{w_4 = 7}$

Answer: C and D

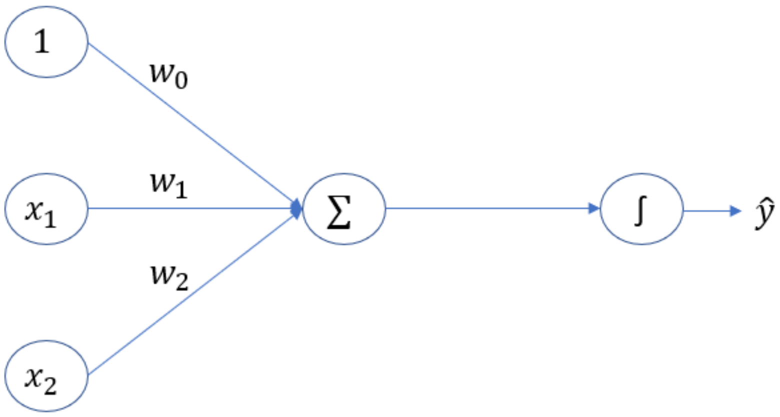
Explanation

$$15 \times 1 + 2w_1 + (-3 \times 9) \geq 0$$
$$2w_1 - 12 \geq 0$$
$$w_1 \geq 6$$

6. Consider two perceptrons:

Perceptron A	Perceptron B
$w_0 = 0$ $w_1 = 2$ $w_2 = 1$	$w_0 = 1$ $w_1 = 2$ $w_2 = 1$

True or False: For any x_1 and x_2, if the output of perceptron A = +1, then perceptron B will have output = +1.



Assume the non-linear activation function is the step function sgn(x).

A. True

B. False

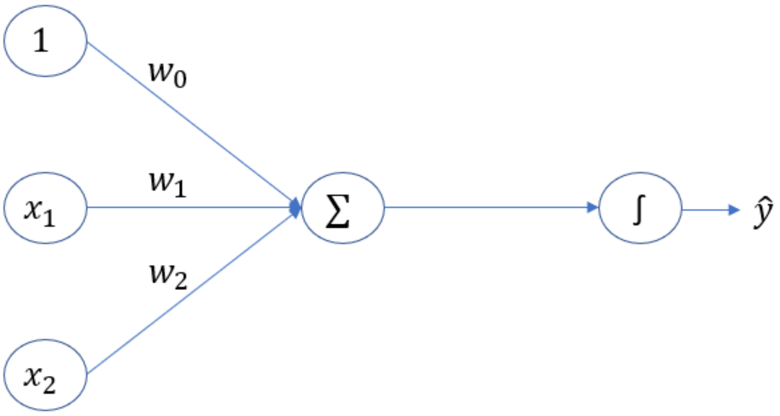
Answer: A

Explanation:
Output of A = $2x_1 + x_2 \geq 0$
Output of B = $1 + \ 2x_1 + x_2 \geq 0$

7. Consider two perceptrons:

Perceptron A	**Perceptron B**
$w_0 = 0$ $w_1 = 2$ $w_2 = 1$	$w_0 = 1$ $w_1 = 2$ $w_2 = 1$

True or false: for any x_1 and x_2, if the output of perceptron A = -1, then perceptron B will have output = -1.



Assume the non-linear activation function is the step function sgn(x).

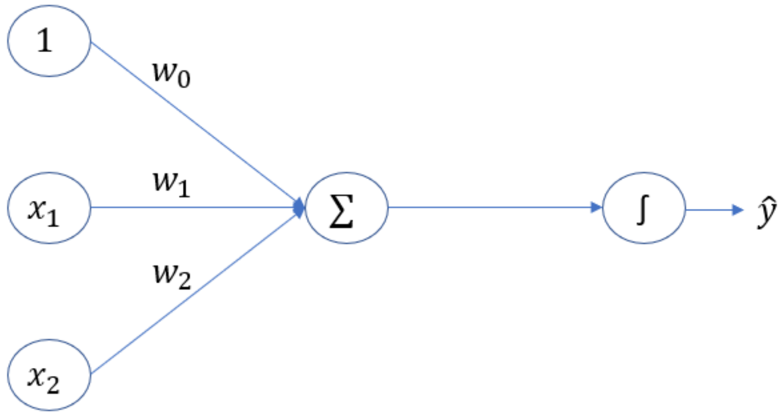
A. True

B. False

Answer: B

Explanation:
Output of A = $2x_1 + x_2 < 0$
Output of B = $1 + \ 2x_1 + x_2$
Set $x_1 = -1, x_2 = 1$ then $2x_1 + x_2 = -1$ and $sgn(2x_1 + x_2) = \ -1$
However, $1 + 2x_1 + x_2 = 0$ and $sgn(1 + 2x_1 + x_2) = +1$

8. What happens if we set our perceptron's activation function to be a linear function?



   A. The perceptron will turn into a linear regression model.
   B. The perceptron will turn into a decision tree.
   C. The perceptron will turn into a logistic regression model.
   D. None of the above

   Answer: A (Refer to Slide 45)

9. Similar to the previous question, what happens if we set our perceptron's activation function to be a sigmoid function?

$$g(z) = \frac{1}{1 + e^{-z}}$$

   A. The perceptron will turn into a linear regression model.
   B. The perceptron will turn into a decision tree.
   C. The perceptron will turn into a logistic regression model.
   D. None of the above

   Answer: C (Refer to Slide 45)

10. You have trained a perceptron and the weights converged. However, you realized that the data are mislabelled (+1 becomes -1 and vice versa) and you lost the training data.

    True or false: we can still create a new perceptron which produce correct labels with the weights from the old perceptrons.

    You may assume that there are no data points that fall exactly on the boundary between the separator line (sum of wx = 0).

    A. True

    B. False

    Answer: A

    Explanation:
    Since it's the labelling that were mislabelled (+1 becomes -1 and vice versa), we can just negate the weights of the old perceptrons.

11. Consider a perceptron with w = [0; 1; 0.5]. We collected training data with x_1 = 4 and x_2 = -2 which has label = -1.

    What is the new updated weight of the perceptron using this training data? (Set learning rate = 0.1)

    Perceptron Update Rule:

$$\overset{\text{new weight}}{w} \leftarrow \underset{\text{old weight}}{w} + \overset{\text{learning rate}}{\eta}\underbrace{(y - \hat{y})}_{\text{learning error}}x$$

    Assume the non-linear activation function is sgn function.

    A. w' = [0; 1; 0.5]
    B. w' = [-0.2; 0.2; 0.9]
    C. w' = [0.2; -0.2; 0.9]
    D. w' = [0.2; 1.8; 0.1]

    Answer: B

Explanation

Output of perceptron:

$(1 \ \times 4) + (0.5 \ \times -2) = 2$

$sgn(2) = \ +1$

The data would be misclassified with label 1 when its actual label is -1

$$\begin{bmatrix} 0 \\ 1 \\ 0.5 \end{bmatrix} + \big(0.1 \times (-1 - 1)\big) \cdot \begin{bmatrix} 1 \\ 4 \\ -2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0.5 \end{bmatrix} + (-0.2) \cdot \begin{bmatrix} 1 \\ 4 \\ -2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0.5 \end{bmatrix} + \begin{bmatrix} -0.2 \\ -0.8 \\ 0.4 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 1 \\ 0.5 \end{bmatrix} + \begin{bmatrix} -0.2 \\ -0.8 \\ 0.4 \end{bmatrix} = \begin{bmatrix} -0.2 \\ 0.2 \\ 0.9 \end{bmatrix}$$

12. What is the derivative of σ($2x$) relative to x?
    Hint: Use chain rule and remember that σ'(x)=σ(x)(1-σ(x))
    A. f'(x) = σ(2x)(1-σ(2x))
    B. f'(x) = 2σ(2x)(1-σ(2x))
    C. f'(x) = 2σ(x)(1-σ(x))
    D. f'(x) = σ(x)(1-σ(x))

Answer: B

Explanation:
Since $\sigma(x) = \frac{1}{1+e^{-x}}$ then $\sigma(2x) = \frac{1}{1+e^{-2x}}$

$$\sigma'(2x) = \frac{d(\frac{1}{1+e^{-2x}})}{dx} = \frac{d}{dx}((1+e^{-2x})^{-1})$$

$$\frac{d}{dx}((1+e^{-2x})^{-1}) = -(1+e^{-2x})^{-2} \cdot (-2e^{-2x}) = 2 \cdot \frac{e^{-2x}}{(1+e^{-2x})^2}$$

$$2 \cdot \frac{e^{-2x}}{(1+e^{-2x})^2} = 2 \cdot \frac{1}{1+e^{-2x}} \cdot \frac{e^{-2x}}{1+e^{-2x}} = 2 \cdot \frac{1}{1+e^{-2x}} \cdot (1 - \frac{1}{1+e^{-2x}})$$

$$2 \cdot \frac{1}{1+e^{-2x}} \cdot \left(1 - \frac{1}{1+e^{-2x}}\right) = 2\sigma(2x)\big(1-\sigma(2x)\big)$$

Lecture Training 9: More Neural Networks

1. What is the role of the **activation function** in a multi-layer perceptron?
   A. To compute the dot product of the input and weight vectors
   B. To introduce non-linearity into the output of each neuron
   C. To compute the derivative of the loss function with respect to the weights
   D. To regulate the amount of regularization applied to the weights

   Answer: B

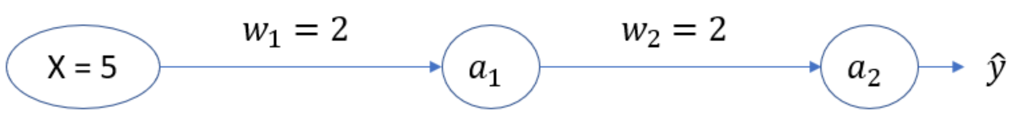
   Explanation: Activation functions like ReLU or Sigmoid are usually non-linear. (Lecture 8, Slide 22)

2. What is the purpose of the bias term in a multi-layer perceptron?
   A. To add additional features to the input data
   B. To introduce non-linearity into the output of each neuron
   C. To adjust the threshold of activation for each neuron
   D. To regularize the weights of the network to prevent overfitting

   Answer: C

   Explanation: https://stackoverflow.com/questions/2480650/what-is-the-role-of-the-bias-in-neural-networks

3. Given the multi-layer neural network below, what is the output of the neural network?

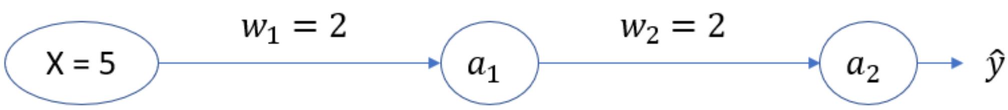


   The activation function at $\mathbf{a_1}$ is the ReLU function and the activation function at $\mathbf{a_2}$ is the linear function f(x) = x
   A. ŷ = 1
   B. ŷ = 9
   C. ŷ = 10
   D. ŷ = 20

   Answer: D

   Explanation:
   $a_1 = \max(0, w_1 x) = \max(0, 5 \times 2) = \max(0, 10) = 10$
   $a_2 = 10 \times 2 = 20$
   $\hat{y} = 20$

4. Given the multi layer neural network below, what is the new weight for w_2 after one step of gradient descent? Assume the true label is y = 18, learning rate = 0.1, and the loss function is MSE.



The activation function at **a_1** is the ReLU function and the activation function at **a_2** is the linear function f(x) = x

A. w_2 = -1

B. w_2 = 0

C. w_2 = 1

D. w_2 = 2

Answer: B

Explanation

$$\frac{\delta L}{\delta w_2} = \frac{\delta L}{\delta \hat{y}} \frac{\delta \hat{y}}{\delta w_2}$$

$$L = \frac{1}{2}(\hat{y} - y)^2$$

$$\frac{\delta L}{\delta \hat{y}} = (\hat{y} - y)$$

$$\hat{y} = a_2 = w_2 a_1 = 2 \times 5 \times 2 = 20$$

$$\frac{\delta \hat{y}}{\delta w_2} = a_1 = 5 \times 2 = 10$$

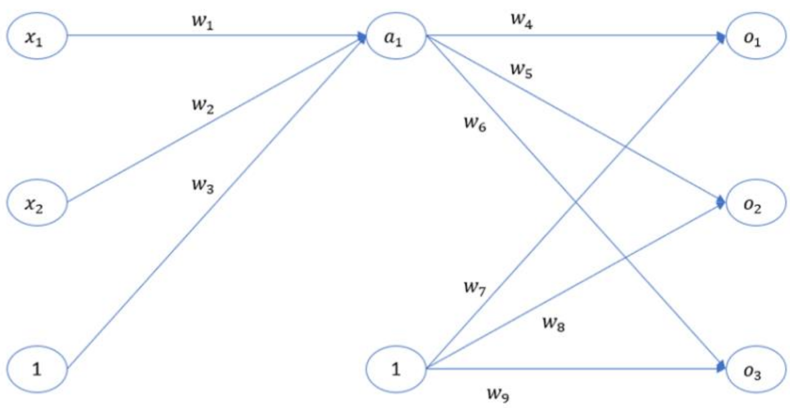
$$\frac{\delta L}{\delta w_2} = \frac{\delta L}{\delta \hat{y}} \frac{\delta \hat{y}}{\delta w_2} = (\hat{y} - y) \cdot a_1$$

$$w_{2Updated} = w_2 - 0.1 \times (20 - 18) \cdot 10$$

$$= w_2 - 2$$

$$= 2 - 2 = 0$$

5. Given the multi-layer neural network below, what is the derivative of the loss L with respect to w_3.



$o_1 = w_4 a_1 + w_7$ $\quad$ $o_2 = w_5 a_1 + w_8$ $\quad$ $o_3 = w_6 a_1 + w_9$

$a_1 = w_1 x_1 + w_2 x_2 + w_3$ $\quad$ $L = \frac{1}{2}[(o_1 - t_1)^2 + (o_2 - t_2)^2 + (o_3 - t_3)^2]$

A. $\frac{\delta L}{\delta w_3} = (o_1 - t_1)w_1 + (o_2 - t_2)w_2 + (o_3 - t_3)w_3$

B. $\frac{\delta L}{\delta w_3} = (o_3 - t_3)w_4 + (o_2 - t_2)w_5 + (o_1 - t_1)w_6$

C. $\boldsymbol{\frac{\delta L}{\delta w_3} = (o_1 - t_1)w_4 + (o_2 - t_2)w_5 + (o_3 - t_3)w_6}$

D. None of the above

Answer: C

Explanation

$$\frac{\delta L}{\delta w_3} = \frac{\delta L}{\delta a_1}\frac{\delta a_1}{\delta w_3}$$

$$\frac{\delta L}{\delta a_1} = \frac{\delta L}{\delta o_1}\frac{\delta o_1}{\delta a_1} + \frac{\delta L}{\delta o_2}\frac{\delta o_2}{\delta a_1} + \frac{\delta L}{\delta o_3}\frac{\delta o_3}{\delta a_1}$$

$$\frac{\delta L}{\delta o_1} = \frac{\delta}{\delta o_1}\left(\frac{1}{2}(o_1 - t_1)^2 + \frac{1}{2}(o_2 - t_2)^2 + \frac{1}{2}(o_3 - t_3)^2\right) = (o_1 - t_1)$$

$$\frac{\delta L}{\delta o_2} = \frac{\delta}{\delta o_2}\left(\frac{1}{2}(o_1 - t_1)^2 + \frac{1}{2}(o_2 - t_2)^2 + \frac{1}{2}(o_3 - t_3)^2\right) = (o_2 - t_2)$$

$$\frac{\delta L}{\delta o_3} = \frac{\delta}{\delta o_3}\left(\frac{1}{2}(o_1 - t_1)^2 + \frac{1}{2}(o_2 - t_2)^2 + \frac{1}{2}(o_3 - t_3)^2\right) = (o_3 - t_3)$$

$$\frac{\delta o_1}{\delta a_1} = \frac{\delta}{\delta a_1}(w_4 a_1 + w_7) = w_4$$

$$\frac{\delta o_2}{\delta a_1} = \frac{\delta}{\delta a_1}(w_5 a_1 + w_8) = w_5$$

$$\frac{\delta o_3}{\delta a_1} = \frac{\delta}{\delta a_1}(w_6 a_1 + w_9) = w_6$$

$$\frac{\delta a_1}{\delta w_3} = \frac{\delta}{\delta w_3}(w_1 x_1 + w_2 x_2 + w_3) = 1$$

$$\frac{\delta L}{\delta w_3} = \frac{\delta L}{\delta a_1}\frac{\delta a_1}{\delta w_3} = [(o_1 - t_1)w_4 + (o_2 - t_2)w_5 + (o_3 - t_3)w_6] \times 1$$
$$= (o_1 - t_1)w_4 + (o_2 - t_2)w_5 + (o_3 - t_3)w_6$$

6. Consider a convolution where the dimension of the image is 11 x 11, the kernel size is 3 x 3, the padding = 1, and the stride = 2. What is the dimension of the output?
   A. 5 x 5
   B. 6 x 6
   C. 7 x 7
   D. 8 x 8

   Answer: B

   Explanation:
   When the padding is 1, the image size is now 13 by 13.
   Number of times sliding window can move = 13 – 3 = 10
   Output Dimension = $1 + floor\left(\frac{10}{2}\right) = 1 + 5 = 6$ (Additional 1 because the first position is also counted)

7. What is the resulting output when we apply the given filter to the image matrix?

Image Matrix

1	1	1	1	0
0	0	0	1	1
0	1	1	1	1
0	0	0	0	1
1	1	0	0	0

Filter

-2	-1	0
-1	1	1
0	1	2

A.

0	1	-1
2	1	2
0	-2	-2

B.

0	1	2
2	1	2
0	-3	-2

C.

0	1	2
2	1	2
-1	-3	-2

D.

0	1	2
3	1	1
0	-3	-1

Answer: B

Explanation: Let W_{ij} be a 3 by 3 matrix

$W_{11} = (-2 \cdot 1) + (-1 \cdot 1) + (0 \cdot 1) + (-1 \cdot 0) + (1 \cdot 0) + (1 \cdot 0) + (0 \cdot 0) + (1 \cdot 1) + (2 \cdot 1) = 0$

$W_{12} = (-2 \cdot 1) + (-1 \cdot 1) + (0 \cdot 1) + (-1 \cdot 0) + (1 \cdot 0) + (1 \cdot 1) + (0 \cdot 1) + (1 \cdot 1) + (2 \cdot 1) = 1$

$W_{13} = (-2 \cdot 1) + (-1 \cdot 1) + (0 \cdot 0) + (-1 \cdot 0) + (1 \cdot 1) + (1 \cdot 1) + (0 \cdot 1) + (1 \cdot 1) + (2 \cdot 1) = 2$

$W_{21} = (-2 \cdot 0) + (-1 \cdot 0) + (0 \cdot 0) + (-1 \cdot 0) + (1 \cdot 1) + (1 \cdot 1) + (0 \cdot 0) + (1 \cdot 0) + (2 \cdot 0) = 2$

$W_{22} = (-2 \cdot 0) + (-1 \cdot 0) + (0 \cdot 1) + (-1 \cdot 1) + (1 \cdot 1) + (1 \cdot 1) + (0 \cdot 0) + (1 \cdot 0) + (2 \cdot 0) = 1$

$W_{23} = (-2 \cdot 0) + (-1 \cdot 1) + (0 \cdot 1) + (-1 \cdot 1) + (1 \cdot 1) + (1 \cdot 1) + (0 \cdot 0) + (1 \cdot 0) + (2 \cdot 1) = 2$

$W_{31} = (-2 \cdot 0) + (-1 \cdot 1) + (0 \cdot 1) + (-1 \cdot 0) + (1 \cdot 0) + (1 \cdot 0) + (0 \cdot 1) + (1 \cdot 1) + (2 \cdot 0) = 0$

$W_{32} = (-2 \cdot 1) + (-1 \cdot 1) + (0 \cdot 1) + (-1 \cdot 0) + (1 \cdot 0) + (1 \cdot 0) + (0 \cdot 1) + (1 \cdot 0) + (2 \cdot 0) = -3$

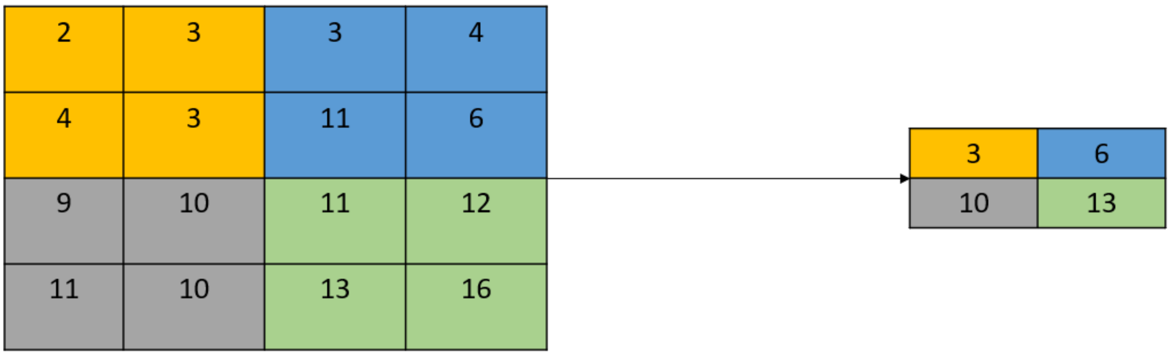
$W_{33} = (-2 \cdot 1) + (-1 \cdot 1) + (0 \cdot 1) + (-1 \cdot 0) + (1 \cdot 0) + (1 \cdot 1) + (0 \cdot 0) + (1 \cdot 0) + (2 \cdot 0) = -2$

8. When we apply convolution, the dimension of the output is always less than the dimension of the input.
   A. True
   B. False

   Answer: B

   Explanation: Using a 1 by 1 convolution layer would result in the output dimension having the same dimensions as the input layer.

9. What kind of pooling is done given the figure below?



   A. Max-pool
   B. Average-pool
   C. Sum-pool
   D. None of the above

   Answer: B

10. Suppose you have an input with size 300 x 300. You have a convolutional layer consisting of 20 kernels, where each kernel has a size 5 x 5. How many parameters are in this layer?
   A. 20
   B. 25
   C. 300
   D. 500

   Answer: D

   Explanation: The number of parameters in the convolutional layer is the sum of the number of weights on each kernel.

   $5 \times 5 \ \times 20 = 500$

Lecture Training 10: Neural Networks on Sequential Data

1. We can use softmax for binary classification problems.
   A. True
   B. False

   Answer: A (True)

   Explanation:
   $$softmax(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$$
   $$P(class\ 0) = \frac{e^{z_0}}{e^{z_0} + e^{z_1}}$$
   $$P(class\ 1) = \frac{e^{z_1}}{e^{z_0} + e^{z_1}}$$

2. MLP vs CNN vs RNN

   Which of the following utilises weight sharing? Choose all that apply.

   Note that weight sharing was only briefly mentioned in lecture. You are encouraged to do some Googling and exploration of the topic on your own.
   A. Multi-layer perceptron
   B. Convolutional neural network
   C. Recurrent neural network

   Answer: B, C

3. What is a major advantage of Recurrent Neural Networks?
   A. It can handle variable-length inputs.
   B. RNNs can process data (e.g. multiple input sentences) in parallel, leading to faster training times.
   C. It is good at extracting features from an image.

   Answer: A

4. When we implement drop out for regularization, we randomly set some weights to be 0 during training.
   A. True
   B. False

   Answer: B (False)

Explanation: During training, we randomly set some **activations** to 0. Keyword here is activations NOT weights!

5. ReLU will not face vanishing gradient or exploding gradient problem.
   A. True
   B. False

   Answer: B (False)

Explanation: Although ReLU primarily addresses the vanishing gradient problem, it can still face the exploding gradient problem if large learning rates are used, especially in recurrent neural networks.

6. What is the use of LSTM over normal RNN?
   A. They can process data in both forward and backward directions, unlike normal RNNs.
   B. They can remember or forget hidden states, unlike normal RNNs.
   C. They prevent overfitting in neural networks, which normal RNNs struggle with.
   D. They require less training data compared to normal RNNs.

   Answer: B

7. What sequence modelling should you use for a handwriting to text function.
   A. One to One
   B. One to Many
   C. Many to One
   D. Many to Many

   Answer: D

Explanation: The input is a sequence of pen strokes or pixel values from an image of handwriting. The output is a corresponding sequence of characters or words.

8. What sequence modelling should you use for a language Identification function.
   A. One to One
   B. One to Many
   C. Many to One
   D. Many to Many

   Answer: C

   Explanation: The input is a sequence of characters or words. The output is a single label corresponding to the type of language. Alternatively you can think of it as a classification task that takes in a sequence of characters and outputs only one label.

Lecture Training 11: Unsupervised Learning

1. We need to perform mean normalization / feature scaling in order to perform k-means clustering meaningfully.
   A. True
   B. False

   Answer: A (True)

   > **Explanation:**
   > K-means clustering relies on Euclidean distance to assign points to clusters. Features with larger scales will have a disproportionately large impact on distance calculations, potentially skewing the clustering results. For example, if one feature has values in the range of thousands (e.g., income) and another in the range of single digits (e.g., age in decades), the k-means algorithm will prioritize the larger-scale feature, as it contributes more to the distance calculation.

2. We have 3 centroids, $\mu_1 = \begin{bmatrix}1\\3\end{bmatrix}, \mu_2 = \begin{bmatrix}-1\\2\end{bmatrix}$ and $\mu_3 = \begin{bmatrix}-1\\4\end{bmatrix}$. Moreover, we have a particular training data $x^{(i)} = \begin{bmatrix}0\\1\end{bmatrix}$. After assigning each training data to a cluster, what is $c^{(i)}$? Note: Use Euclidean distance to calculate the distance between a point and a centroid.
   A. $c^{(i)} = 1$
   B. $\boldsymbol{c^{(i)} = 2}$
   C. $c^{(i)} = 3$
   D. Cannot be determined.

   Answer: B

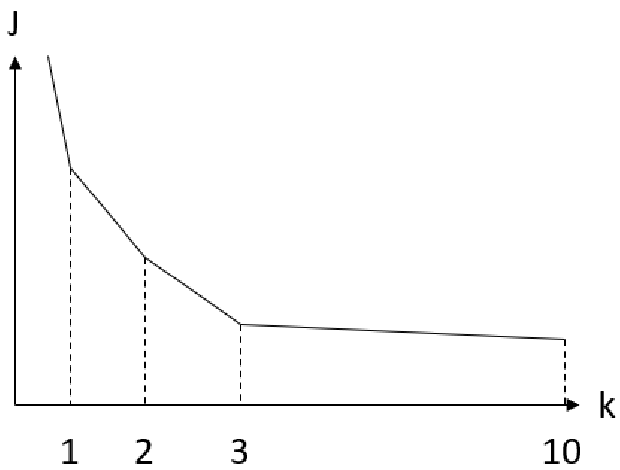
   > **Explanation:**
   > Let δ be the Euclidean distance.
   > $$\delta\left(x^{(i)}, \mu_1\right) = \sqrt{(0-1)^2 + (1-3)^2} = \sqrt{5}$$
   > $$\delta\left(x^{(i)}, \mu_2\right) = \sqrt{(0-(-1))^2 + (1-2)^2} = \sqrt{2}$$
   > $$\delta\left(x^{(i)}, \mu_1\right) = \sqrt{(0-(-1))^2 + (1-4)^2} = \sqrt{10}$$
   >
   > $x^{(i)}$ is assigned to $c^{(i)} = 2$ as the Euclidean distance between $x^{(i)}$ and the centroid μ_2 is the shortest.

3. Suppose we run the k-means algorithm on a certain dataset and get the final centroids A. If we re-run the k-means algorithm on the same dataset and get the final centroids B, A will always be the same as B.
   A. True
   B. False

   Answer: B (False)

   Explanation: The k-means clustering algorithm picks k initial centroids randomly from the points in the data. The final location of the centroids depend on the starting centroids, hence it may be different across re-runs.

4. Based on the elbow method, which k should we choose?



   A. 1
   B. 2
   C. 3
   D. 10

   Answer: C

5. Suppose you run k-means clustering on a dataset 10 times and get 10 different clustering data. How to pick the best clustering data? Note, clustering data consists of $c^{(1)}, c^{(2)}, \dots, c^{(m)}, \mu_1, \mu_2, \dots, \mu_k$.
   A. Select the 1st clustering data
   B. Select the 10th clustering data
   C. Plot all 10 clustering data and visually choose the best one
   D. Use the elbow method
   E. Calculate $J(c^{(1)}, c^{(2)}, \dots, c^{(m)}, \mu_1, \mu_2, \dots, \mu_k)$ for each clustering data and chose the one that minimizes the cost function.
   F. None of the above.

   Answer: E

Explanation: In this question, we assume k is already fixed, so no point using the elbow method. Instead, we use the loss function known as "distortion" that measures the distance of each sample to its centroid formulated as $J(c^{(1)}, c^{(2)}, \dots, c^{(m)}, \mu_1, \mu_2, \dots, \mu_k)$

6. Which of the following is a disadvantage of hierarchical clustering?
   A. It requires the number of clusters to be predefined.
   B. It does not handle noise well.
   C. It does not work well with non-linearly separable data.
   D. It is computationally expensive for large datasets.

   Answer: D

Explanation: High space and time complexity, making it impractical for large datasets. (Slide 34)

7. We can apply SVD to non-square matrix (number of rows != number of columns)
   A. True
   B. False

   Answer: A (True)

Explanation: Take without loss of generality $n > m$, For any $n \times m$ rectangular real-valued matrix X, there exists a factorization $X = U\Sigma V^T$ called SVD, such that • U is $n \times m$ and has m orthonormal columns. • Σ is a $m \times m$ diagonal matrix with $\sigma_j > 0$. • V is $m \times m$ and has m orthonormal columns and rows.

8. When we perform PCA, how do we choose k? Assume we want to retain at least 99% variance.
   A. Select the highest k such that at least 99% of the variance is retained.
   B. Select the lowest k such that at least 99% of the variance is retained.
   C. Set k = 99% * num_of_features + 1
   D. None of the above.

   Answer: B

Explanation: Refer to Slides 70 – 72

In principal component analysis, we want to capture components that maximize the statistical variations of the data.

Steps:

1. Create the covariance matrix of the data $Cov(X) = \frac{1}{m}\hat{X}\hat{X}^T$
2. Compute SVD on the $Cov(X)$ to obtain the U matrix
3. Reduce r components to obtain $\tilde{U}$.

Note that X's singular values^2 are related to the singular values (variances) in Cov:

$$XX^T = U\Sigma V^T(U\Sigma V^T)^T = U\Sigma V^T V\Sigma U^T = U\Sigma\Sigma U^T = U\begin{bmatrix}\sigma_1^2 & & \\ & \sigma_2^2 & \\ & & \ddots & \\ & & & \sigma_m^2\end{bmatrix}U^T$$

Choose minimum r such that $\frac{\sum_{i=1}^{r}\sigma_i^2}{\sum_i^m \sigma_i^2} \geq 0.99$