CSE343/CSE543/ECE363/ECE563: Machine Learning Sec A (Monsoon 2023)
END-Sem

Date of Examination: 6.12.2023 Duration: 2 hours Total Marks: 30 marks

---

Instructions –

- Attempt all questions. MCQs have a single correct option.
- State any assumptions you have made clearly.
- Standard institute plagiarism policy holds.
- No evaluation without suitable justification.

0 marks if either option or explanation is incorrect.

Question 1: [1 Mark] In the context of batch learning in neural networks, does the order in which input data is presented to the network influence the learning process?

A. Yes, the order significantly affects learning, leading to different models.

B. No, the order of input data does not influence the learning outcome in batch learning.

C. Yes, but only in networks with more than three layers.

D. The order only matters in real-time data streaming, not in batch learning.

Answer: (B)No, the order of input data does not influence the learning outcome in batch learning. In batch learning, the entire dataset is considered in one go, or in large batches, for the training process. This approach means that the order in which the data points are presented does not affect the learning process, as the model processes the entire dataset or large subsets of it collectively. The training is based on the aggregate error over the entire dataset or the batch, rather than individual data points, making the order of data presentation irrelevant in this context.

Question 2: [1Marks]What impact will increasing the number of nodes in a neural network's hidden layer have?

A. It will always improve the network's accuracy on both training and test sets.

B. It may lead to underfitting, resulting in poor accuracy on both training and test sets.

C. It may result in overfitting, leading to higher accuracy on the training set but poor accuracy on the test set, and increased training time.

D. The number of nodes in the hidden layer has no impact on the network's performance.

(C). It may result in overfitting, leading to higher accuracy on the training set but poor accuracy on the test set and increased training time.

Question 3: [1Marks]You are designing a deep learning system to detect driver fatigue in cars, where it is crucial to detect fatigue accurately to prevent accidents. Which of the following evaluation metrics would be the most appropriate for this system?

1. Precision
2. Recall
3. F1 score
4. Loss value

(B)Explanation: Recall is the most appropriate metric in this context because it measures the proportion of actual positive cases (fatigue detected) that were correctly identified. In safety-critical applications like driver fatigue detection, it's crucial to minimize false negatives (i.e., not detecting fatigue when it is present), which is what Recall focuses on.

Question 4: [2 Marks] Assuming $\delta L/\delta x_3$ is known, write the weight update for w_1 ($\delta L/\delta w_1$ should be in the expanded form). Input x_1 and all weights are positive. All neurons have ReLU Activation.Figure attached below.

Ans:

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial y_1}\frac{\partial y_1}{\partial w_1}$$

$$= \{\frac{\partial L}{\partial z_4}\frac{\partial z_4}{\partial y_1} + \frac{\partial L}{\partial z_3}\frac{\partial z_3}{\partial y_1} + \frac{\partial L}{\partial z_2}\frac{\partial z_2}{\partial y_1}\}\frac{\partial y_1}{\partial z_1}\frac{\partial z_1}{\partial w_1}$$

$$= \{\frac{\partial L}{\partial x_3}\frac{\partial x_3}{\partial z_4}\frac{\partial z_4}{\partial y_1} + \frac{\partial L}{\partial z_4}\frac{\partial z_4}{\partial y_2}\frac{\partial y_2}{\partial z_3}\frac{\partial z_3}{\partial y_1} + \frac{\partial L}{\partial z_3}\frac{\partial z_3}{\partial x_2}\frac{\partial x_2}{\partial z_2}\frac{\partial z_2}{\partial y_1}\}.1.x_1$$

$$= \{\frac{\partial L}{\partial x_3}.1.w_7 + \frac{\partial L}{\partial x_3}.w_4.1.w_8 + \frac{\partial L}{\partial x_3}.w_4.w_3.1.w_2\}.x_1$$

$$= \frac{\partial L}{\partial x_3}(w_7 + w_4w_8 + w_4w_3w_2).x_1$$

$$\text{Given}\frac{\partial y_1}{\partial z_1} = \frac{\partial x_2}{\partial z_2} = \frac{\partial y_2}{\partial z_3} = \frac{\partial x_3}{\partial z_4} = 1$$

Figure 1: Solution to Q4 [1 mark for derivation, 1 mark for correct answer]

Question 5: [2+2 Marks] Consider an activation function $\rho(x) = x \cdot \sigma(x)$, where $\sigma(x)$ is the sigmoid function.

(a) Compute $\rho'(x)$ in terms of $\sigma(x)$.

(b) For large x, compare $\rho(x)$ and $\rho'(x)$ with standard activation functions. (No derivation required).

Part (a) Given that: $\sigma'(x) = \sigma(x)(1 - \sigma(x))$

By the multiplication rule:

$$\rho'(x) = x'\sigma(x) + x\sigma'(x) = 1\sigma(x) + x\sigma(x)(1 - \sigma(x)) = \sigma(x)[1 + x(1 - \sigma(x))]$$

Part (b)

1. For large values of x, $\sigma(x) \approx 1$, thus $\rho(x) \approx x$ and the function mimics RELU activation.
2. For large values of x, $\sigma(x) \approx 1$, or $(1 - \sigma(x)) \approx 0$, and overall $[1 + x(1 - \sigma(x))] \approx 1$, $\rho'(x)$ function mimics Sigmoid activation.

Question 6: [2 Marks] Consider an ensemble of 5 models which are trained on the same architecture but have different initializations for a handwritten digit classification task. Does the guarantee of better performance in expectation (in terms of cross-entropy loss) by averaging the predictions of all five networks hold if you instead average the weights and biases of the networks? Why or why not?

Solution: No, the guarantee does not hold because the loss is not convex with respect to the weights and biases. Networks starting from different initializations might learn different hidden representations, so it makes no sense to average the weights and biases.

Question 7: [2 Marks] Prove that approximately 63% of the entire original dataset (total training set) is present in any of the sampled bootstrap datasets using the Bagging method.

Solution: We have a dataset with n observations. In bootstrap sampling, we draw with replacement from this original dataset n times to create a new dataset of the same size n.
For any single observation, the probability of not being chosen on the first draw is: P(not chosen) = 1 - 1/n [1 mark]
Since each draw is independent of the others, the probability of not being chosen in all n draws is: P(not chosen in all n draws) = $(1 - 1/n)^{\text{n}}$
Taking the limit: As n gets large (which is typically the case in real applications),
$limit_{\text{n to infinity}}\ (1 - 1/n)^{\text{n}} = e^{-1} = 0.368$
Probability of appearing = 1-0.368 = 0.63 [1 mark]

Question 8:[2 marks] Consider the given distance metric:

$$d(x, y) = \sum \left| \frac{x_i - y_i}{x_i + y_i} \right|$$

Enumerate and explain the desirable properties of a distance metric. Evaluate the given distance metric against these properties, demonstrating mathematically or through examples how the metric adheres to or fails to meet each of the properties.

Desirable Properties of a Distance Metric:

- Symmetry:
- Property: D(A, B) = D(B, A)
- Example: Euclidean distance is symmetric; d(A, B) = d(B, A).
- Positivity and Self-Similarity:
-
- Property: D(A, B) ≥ 0, D(A, B) = 0 if and only if A = B.
- Example: Manhattan distance is positive, and d(A, B) = 0 only if A = B.
- Triangle Inequality:
- Property: D(A, B) + D(B, C) ≥ D(A, C)
- Example: Euclidean distance satisfies the triangle inequality.

Given metric is Canberra distance. Let's consider four data points (A, B, C, and D) in 2D space with coordinates:

$A(2,5), B(4,3), C(6,7), D(8,2)$

The Canberra distance between two points A and B is given by:

$$d(x,y) = \sum \frac{|x_i - y_i|}{|x_i + y_i|}$$

While Canberra distance exhibits symmetry and satisfies positivity and self-similarity, it fails the triangle inequality, making it unsuitable for applications where the triangle inequality is a critical property of the distance metric.

Figure 2: Solution to Q8

Alternate sol for proving non validity:- when $x_i = -y_i$, distance is infinite.

Question 9: [3 Marks] For a CNN-based classifier, calculate the number of weights, number of biases, and the size of the associated feature maps for each layer, following the notation:

- CONV-K-N denotes a convolutional layer with N filters, each of size $K \times K$. Padding and stride parameters are always 0 and 1, respectively.
- POOL-K indicates a $K \times K$ pooling layer with stride K and padding 0.
- FC-N stands for a fully-connected layer with N neurons.

Successively:

$$120 \times 120 \times 32 \quad \text{and} \quad 32 \times (9 \times 9 \times 3 + 1)$$
$$60 \times 60 \times 32 \quad \text{and} \quad 0$$
$$56 \times 56 \times 64 \quad \text{and} \quad 64 \times (5 \times 5 \times 32 + 1)$$
$$28 \times 28 \times 64 \quad \text{and} \quad 0$$
$$24 \times 24 \times 64 \quad \text{and} \quad 64 \times (5 \times 5 \times 64 + 1)$$
$$12 \times 12 \times 64 \quad \text{and} \quad 0$$
$$3 \quad \text{and} \quad 3 \times (12 \times 12 \times 64 + 1)$$

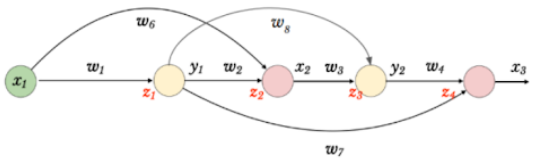


Figure 3: Question 4

Layer	Activation map dimensions	Number of weights	Number of biases
INPUT	$128 \times 128 \times 3$	0	0
CONV-9-32			
POOL-2			
CONV-5-64			
POOL-2			
CONV-5-64			
POOL-2			
FC-3			

Figure 4: Question 9

Question 10: [4 Marks] Discuss the role of activation functions in mitigating exploding and vanishing gradient problems. Provide examples of activation functions that are more or less prone to these issues and explain why.

Exploding and Vanishing Gradient Issues:

Exploding Gradient: Occurs when gradients become extremely large during backpropagation, leading to unstable learning. This occurs when the weights become very large.

Vanishing Gradient: Happens when gradients become too small, causing slow or halted learning for deep networks.

Activation Functions:

Sigmoid Activation: Prone to vanishing gradients, particularly for deep networks.

Hyperbolic Tangent (tanh) Activation: Similar to sigmoid, still susceptible to vanishing gradients.

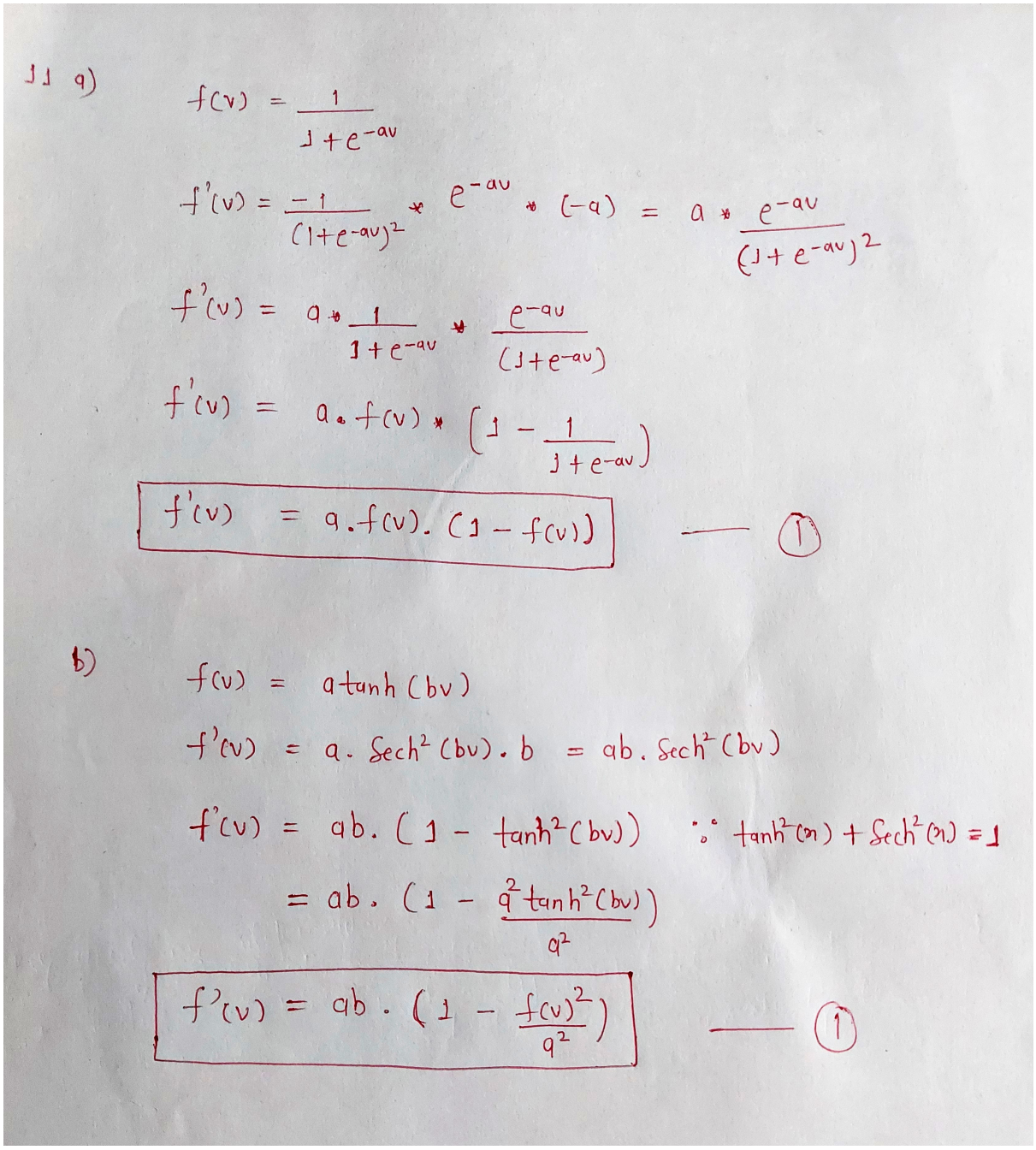
Sigmoid and tanh: More prone to vanishing gradient, limiting their use in deep networks.

ReLU, Leaky ReLU, PReLU, ELU: Designed to alleviate vanishing gradients, making them suitable for deep architectures.

Question 11: [2 Marks] Express the derivative of a sigmoid in terms of the sigmoid itself for positive constants a and b:

(a) A purely positive sigmoid: $\varphi_j(v) = \frac{1}{1+\exp(-av)}$

(b) An antisymmetric sigmoid: $\varphi_j(v) = a \tanh(bv)$



Question 12:[2 Marks] Consider the following patterns, each having four binary-valued attributes:

ω_1	1100	0000	1010	0011
ω_2	1100	1111	1110	0111

Note: especially that the first patterns in the two categories are the same. Identify the root node feature for a binary classification tree for this data so that the leaf nodes have the lowest impurity possible.

Question 13: [2 Marks] In the context of behavior simulation and content simulation tasks, discuss the implications of the observed performance gap between LCBM and large content-only models like GPT-3.5 and GPT-4. What insights can be drawn from this performance difference, and how does it contribute to our understanding of the effectiveness of including behavior tokens in language model training? Provide a brief analysis of the observed trends as presented during

To select the query at the root node, we investigate queries on each of the four attributes. The following shows the number of patterns sent to the "left" and "right" for each value, and the entropy at the resulting children nodes:

query	sent left	left entropy	sent right	right entropy
a_1	$2\omega_1, 1\omega_2$	0.9183	$2\omega_1, 3\omega_2$	0.9710
a_2	$3\omega_1, 0\omega_2$	0	$1\omega_1, 4\omega_2$	0.7219
a_3	$2\omega_1, 1\omega_2$	0.9183	$2\omega_1, 3\omega_2$	0.9710
a_4	$3\omega_1, 2\omega_2$	0.9710	$1\omega_1, 2\omega_2$	0.9183

Because query a_2 leads to the greatest weighted reduction in impurity, a_2 should be the query at the root node. We continue and grow the tree shown in the figure, where the "?" denotes the leaf having equal number of ω_1 and ω_2 patterns.

Figure 5: Q12, All the calculations for entropy/impurity should be there.

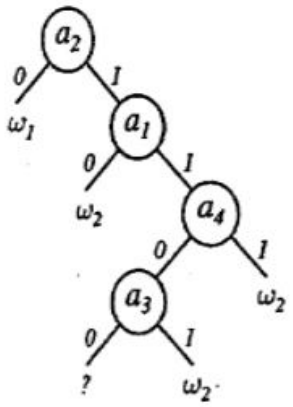


Figure 6: Q12. Final Decision Tree

the lecture.

Q13:- Refer this for more details: https://arxiv.org/pdf/2309.00359.pdf

- LCBM, while being 10x smaller than GPT-3.5 and 4, performs better than them on all behavior-related tasks.
- Further, we see that there is no significant difference between 10-shot and 2-shot GPT-4 or between GPT-3.5 and GPT-4, indicating that unlike other tasks, it is harder to achieve good performance through in-context learning on the behavior modality.
- It can be observed that often GPT-3.5 and 4 achieve performance comparable to (or worse than) random baselines. Interestingly, the performance of GPTs on the content simulation task is also substantially behind LCBM.
- The way we formulate the content simulation task (Listing 5), it can be seen that a substantial performance could be achieved by strong content knowledge, and behavior brings in little variance.

- We still see a substantial performance gap between the two models. All of this indicates that large models like GPT-3.5 and 4 are not trained on behavior tokens.

Question 14: [2 Marks] Consider a scenario where a company is developing an AI-based conversational chatbot focused on mental health support. Before deploying this chatbot, outline the critical considerations the company should take into account and discuss each of the following aspects:

- Controlled Generation vs. Free Flow QA: Compare and contrast the advantages and disadvantages of implementing a controlled generation approach versus a free-flow question-answer model in the context of a mental health chatbot.
- Industry-Specific Approvals: Adherence to industry-specific regulations such as HIPAA, GDPR, NIST, etc., holds significant implications. Explain the implications of adhering to these regulations and the measures the company should take to ensure compliance in the development and deployment of the mental health chatbot.

Controlled Generation: Advantages:

- Precision and predictability in responses, Lower risk of generating inappropriate or harmful content, Aligns with ethical and regulatory standards.

Disadvantages:

- Limited flexibility in responding to diverse user inputs, Potential to miss nuanced expressions and unique user needs, May feel less conversational and empathetic.

Free Flow QA: Advantages:

- Enhanced flexibility in addressing various user inputs, Can simulate more natural and empathetic conversations, Better adaptation to users' emotional states.

Disadvantages:

- Higher risk of generating inappropriate or unsafe content, Difficulty in maintaining control over the conversation, Challenges in aligning with regulatory standards.

0.25 marks for any one advantage and disadvantage of each type(0.25*4=1 mark)

- Implications: Ensures the protection of sensitive health information, Guarantees user privacy and control over personal data,Sets standards for cybersecurity and data protection. Give 0.5 for any one mentioned.

- To ensure compliance : Implement robust encryption for data transmission, Ensure secure storage and access controls for user data, Obtain clear consent for data collection and usage, or something along these lines. 0.5 marks