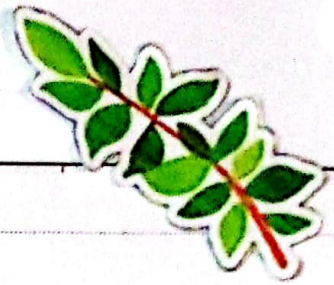


نوايا المذاكرة

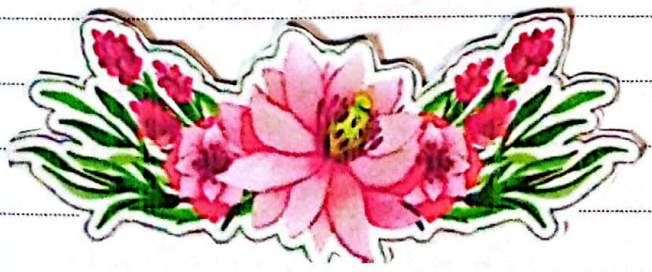
١- رضا الله.
٢- رضا الوالدين وإدخال السرور على قلبهم.
٣- إجابة عن سؤال: وعن عمره فيما أفناه ...
٤- نفع المسلمين بما تعلمت وسد ثغور الأمة.
٥- تحقيق الأهداف وما خلقت لأجله.
٦- تعمير الأرض.
٧- رفع الجهل.
٨- المؤمن القوى خير وأحب إلى الله من المؤمن الضعيف.

" مذاكرتك مسؤوليتك كمسلم فلا تتخاذل! "
تعلم العلم فلا تدرى متى تحتاج لك الأمة!

علمك = ثغرك = مُقاومة ♡

لا تنسونى وأهلى وأهل غزة والمسلمين المُستضعفين
فى الأرض من دعواتكم ولكم بالمثل. ♡

* نرحل ويبقى الأثر
* أفعل الخير ولا تستصغر، فلا تدرى اى حسنة
تُدخلك الجنة.



بصمجة لإمتحان

1- Quadratic function has Numerical & Analytical solutions.
2- The derivative at the max & min points is zero.
3- Gradient descent idea is to take repeated steps aganist the gradient direction to find the min. of the fun.
4- Linear Regression $\Rightarrow y = WX + b$ or $y_K = \sum_{i=1}^{d} W_{Ki} X_i + b_K$
5- Objective function of Linear regression
Mean Square error "MSE" $= \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

or $\frac{1}{N} \sum_{n=1}^{N} E^n$ or $\sum_{n=1}^{N} \sum_{K=1}^{K} (y_K^n - t_K^n)^2$ or $\sum_{n=1}^{N} \sum_{K=1}^{K} \left(\sum_{i=1}^{d} W_{Ki} X_i + b_K - t_K^n \right)^2$

6- Activation function of Linear Regression is Linear Function.
7- The gradient of MSE loss function with respect to
Weight $\Rightarrow (y_r^n - t_r^n) X_s^n$ or $\frac{1}{N} \sum_{n=1}^{N} (y_r^n - t_r^n) X_s^n$
bias $\Rightarrow (y_r^n - t_r^n)$ or $\frac{1}{N} \sum_{n=1}^{N} (y_r^n - t_r^n)$
8- Update for weights & bias Vector

$$W_{rs}^{t+1} = W_{rs}^{t} - \eta \frac{\partial E(w,b)}{\partial W_{rs}} \text{ or } W_{rs}^{t+1} = W_{rs}^{t} - \eta \frac{1}{N} \sum_{n=1}^{N} (y_r^n - t_r^n) X_s^n$$

$$b_r^{t+1} = b_r^t - \eta \frac{\partial E(w,b)}{\partial b_r} \text{ or } b_r^{t+1} = b_r^t - \eta \frac{1}{N} \sum_{n=1}^{N} (y_r^n - t_r^n)$$

9- Objective function of Linear regression is Scalar. "MSE"
10- Predict house price $\Rightarrow$ Application of Linear Regression.
11- Binary activation function isn't Continuous and not differentiable at zero.
12- Sigmoid is Continuous & differentiable.
13- Sigmoid $\Rightarrow F(z) = y = \frac{1}{1 + e^{-z}}$

14- Objective function of binary Classification is Cross entropy or Log maximum Likelihood

15- Binary Cross entropy (BCE) Loss function

16- Objective function of binary Classification ⇒

$$E^n = -\left[t^n \log y^n + (1-t^n) \log (1-y^n)\right]$$

17- Range of Sigmoid function is $[0, 1]$.

18- Output of Sigmoid function when input is zero is 0.5

19- Sigmoid maximum score is 1

20- Gradient of loss function with respect to weights $= (y^n - t^n)X_i$ & with respect to bias $= (y^n - t^n)$

21- Value of y is between 0 & 1, So to make decision we use Threhold (0.5).

22- Threhold for decision → $y \begin{cases} 1 & \text{if } y \geq 0.5 \\ 0 & \text{if } y < 0.5 \end{cases}$

23- Binary Classification must have only one output "scalar"

24- Sigmoid function has a Probabilistic interpretation.

25- Minimum Value of Cross entropy objective function is Zero.

26- Spam email detection is application of binary Classification.

27- Activation Function of Multiclass Classification is Softmax.

28- Softmax ⇒ $y_k = \frac{e^z_k}{\sum_{j=1}^{k} e^z_j}$, $0 < y_k < 1$ $\sum_{k=1}^{k} y_k = 1$

29- Compute the Jacobian matrix &-

if $K = 1$ ⇒ $y_i(1-y_k)$ or $y_i - y_i y_k$

if $K \neq 1$ ⇒ $-y_i y_k$

30- Derivative of Softmax is Matrix (Jacobian matrix).

31- Objective function of Multiclass Classification is Categorical Cross entropy ⇒ $E^n = -\sum_{k=1}^{k} t^n_k \log y^n_k$

32- The Summation of Softmax Vector elements equal to 1
33- Softmax Function is Vector, & it's derivative is Matrix.
34- Categorical Cross entropy objective function is Scalar.
35- In One-hot encoding Only a Single element Can have a Value of 1 . ex: [1,0,0] , [0,1,0,0]
36- Cross entropy is an objective function for binary & Multiclass Classifications. " Scalar "
37- Recognition of handwritten digits (0-9) is an application of Multiclass Classification.
38- Objective Function of MultiLabel lassification is K binary Cross entropy $\Rightarrow E = -t \log y - (1-t) \log (1-y)$
39- Multi-hot Vector more than one element can have a Value of One ex: [0,1,1,1] , [0,0,0,0] , [1,1,1,1]
40- Output of Linear regression is Scalar. $K \geq 1$
41- Output of Binary Classification is Scalar. $K = 1$
42- Output of Multiclass Classification is Vector $K \geq 3$
43- Output of MultiLabel Classification is Vector $K \geq 2$
44- Range of tanh activation function is $[-1, 1]$
45- Output of tanh function when input is Zero = 0 , any +Ve no $(\infty) = 1$, any -Ve no $(-\infty) = -1$
46- Relu Function = $F(X) = \max(0, X)$
47- Non ~~activ~~ Linear activation Functions help the network to avoid Saturation or Vanishing gradients.
48- the role of hidden Layer(s) is to perform Computations & extract features from the data.
49- Information Flow Forward Only , from input to output in multiLayer network.
50- The main benefit of using multiLayer network is the ability to Learn more Complex relationships in data.

ال Linear regression ممكن تكون vector برضو ودا علي حسب التاسك.

في التاسك اللي اخذناه بتاع توقع سعر البيت كانت scalar

51- Convolution operation is $\Rightarrow S(t) = \int X(a)\, W(t-a)\, da$
$= X(t) * W(t) = W(t) * X(t)$ "Commutative"

52- Discrete Convolution : $S(n) = \sum_{a=-\infty}^{\infty} X(a)\, W(n-a)$

53- Two dimentional Convolution :
$S(i,j) = (X*W)(i,j) = \sum_{K=0}^{m-1} \sum_{L=0}^{n-1} X_{K,L}\, W_{i-k, j-L}$

54- Cross Correlation $\Rightarrow$
$S(i,j) = \sum_{K=0}^{m-1} \sum_{L=0}^{n-1} X_{i+k}, W_{K,L} = (W \otimes X)(i,j)$

55- Equation of the output size $= (W-F+2P)/S + 1$

$\frac{W-F+2P}{S} + 1 \Rightarrow \frac{\text{input size} - \text{Filter size} + 2\,\text{Padding}}{\text{Stride}} + 1$

56- Dense Layers in CNN may be Linear regression, or Binary Classification or Multiclass Classification.

57- Main Purpose of Convolutional Layer in CNN is to extract Features from the input image.

58- The role of Pooling Layer in CNN is to downsample the Feature maps and reduce Computational Cost.

59- Padding helps to preserve Spatial information.

60- Purpose of dropout layers in CNN is to prevent overfitting.

61- Stride is the step size taken by the filter when moving across the input data.

62- The most Suitable activation function for hidden Layer is Relu.

63- The input of CNN is matrix, 2D matrix for grayscale image & 3D matrix for Color image (RGB channels).

64- Output of CNN is ⇒ على حسب انا هستخدم ايه في ال Dense Layer.

67- RNNs are nonLinear models.

68- $h_t = f_w(h_{t-1}, X_t)$

69- Hidden State activation ⇒ $Z_t^h = W_h h_{t-1} + W_x X_t + b_h$

$h_t = \tanh(Z_t^h)$

70- Output activation ⇒ $Z_t^y = W_y h_t + b_y$

$y_t = \text{Softmax}(Z_t^y)$

71- Norms of Jacobians :

- $\left\| \frac{\partial h_{j+1}}{\partial h_j} \right\| \leq \|W_h\| \cdot \|\text{diag}(\phi_h'(h_j))\|$ "الكتاب"

ϕ_h = tanh activation function & ϕ_h' is its derivative

- $\left\| \prod_{i=K+1}^{t} \frac{\partial h_i}{\partial h_{i-1}} \right\| \leq \prod_{i=K+1}^{t} \|W\| \cdot \|\text{diag}(f'(h_{t-1}))\| \leq (\gamma_w \gamma_h)^{t-K}$

نوتس الدكتور

- $\left\| \prod_{j=K}^{t} \frac{\partial h_{j+1}}{\partial h_j} \right\| \leq (\gamma_w \gamma_h)^{t-K}$ "الكتاب"

الاتنين نفس الحاجة بس عشان تكون عارف لما تيجي في ال mcq.

72- if the distance t-K is Large ⇒ $\gamma < 1$ gradient will Vanish, if $\gamma > 1$ gradient will explode.

73- - input gate ⇒ $i_t = \sigma(W_i h_{t-1} + U_i X_t)$

- Forget gate ⇒ $f_t = \sigma(W_f h_{t-1} + U_f X_t)$
- Output gate ⇒ $O_t = \sigma(W_o h_{t-1} + U_o X_t)$

memory cell ⇒ $C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$

$\tilde{C}_t = \tanh(W_c h_{t-1} + U_c X_t)$

$C_t = f_t \odot C_{t-1} + i_t \odot \tanh(W_c h_{t-1} + U_c X_t)$

$h_t = O_t \odot \tanh(C_t)$ ⇒ hidden State

74- What function is typically used to activate the forget gate, input gate, and output gate in LSTMs?
Sigmoid for gates, tanh for cell state.

75- What's the main advantage of LSTMs compared to standard RNNs? They can learn long-term dependencies in data.

76- What kind of neural network architecture is LSTM a part of?
Recurrent Neural Network (RNN)

77- Compared to a standard classification network with softmax activation, How does the output layer of multi-label classification network differ? It uses a Sigmoid activation function on each node.

78- What is Backpropagation Through Time (BPTT)?
An extension of backpropagation that accounts for the recurrent nature of the network.

79- What is the loss function used in RNN training?
Categorical Cross-Entropy $\Rightarrow E^n = -\sum_{k=1}^{K} t_k^n \log y_k^n$

80- What is the formula for the overall loss E?

① $E = \sum_{t=1}^{T} E_t(l_t, y_t)$

② $E = -\sum_{t=1}^{T} l_t \log y_t$

③ $E = -\sum_{t=1}^{T} l_t \log[\text{softmax}(y_t)]$

قول معايا: سبحان الله، الحمد لله لا إله إلا الله، الله أكبر لا حول ولا قوة إلا بالله ♡ اللهم صلِّ على سيدنا محمد صلى الله عليه وسلم.

81- Seq 2 Seq Translation, Input: Sequence of word vectors from the source language (Matrix)
Output: Sequence of word vectors in the target language (Matrix)

82- Which Problem does the attention mechanism address in Seq2Seq model? - Difficulty in processing Long Sequences.

83- How does the attention mechanism improve the performance of Seq2Seq models?
By assigning different weights to different parts of the input sequence.

84- What's the role of the Context Vector in a Seq2Seq model with attention?
Carry the weighted sum of the encoder hidden states.

85- What is the key difference between a traditional RNN & LSTM used in Seq2Seq models?
LSTMs have gates to control the flow of information.

86- What is the benefit of using bidirectional RNNs in the encoder part of a Seq2Seq model?
They provide Context from both past & Future tokens in the sequence.

87- What Problem can arise if Seq2Seq models without attention mechanisms are used for long sequences?
Vanishing gradients.

88- What is the main advantage of using an LSTM or GRU over a traditional RNN in Seq2Seq model?
They mitigate the vanishing gradient problem.

89- What is a significant benefit of multi-head attention in transformer models?
It allows the model to focus on different parts of the input simultaneously.

قال رسول الله صلى الله عليه وسلم: إنّ الله يُحبُّ إذا عمل أحدُكم عملاً أن يُتقنه.

90- What does the 'alignment model' in an attention based Seq2Seq System do?
It Calculates the attention weights.

91- What type of function is Commonly Used to Calculate the alignment Scores in attention mechanisms?
Softmax Function.

92- What is the main function of the "query", "key" & "value" Vectors in the attention mechanism?
They are used to Compute the alignment scores for attention.

93- How does the transformer model handle the Vanishing gradient problem that is Common in RNNs?
By employing residual Connections.

* Transformer model uses residual Connections to mitigate the Vanishing gradient Problem, allowing gradients to flow more easily through the network during backpropagation.

94- Which mechanism allows transformers to better Capture long dependencies Compared to RNNs?
Self-attention

95- What is the Purpose of Layer normalization in Transf.
Stabilize & accelerates training.

96- What is the purpose of the Softmax function applied to the alignment scores?
To Convert the scores into Probabilities.

97- In transformer model, what type of input is provided to the encoder & decoder? Matrices.

98- What is Purpose of key matrix K?
To Compute the dot product with query matrix.

99- In multi-head attention mechanism, what type of output is produced by each individual attention head?

Sequence of Vectors (Matrix)

100- Which equation describes the scaling factor applied in the Scaled dot-product attention?

$$\text{Scale} = \sqrt{d_k}$$

101- Equation represents the attention scores before scaling $\Rightarrow$ $\text{Attention}(Q, K, V) = \text{Softmax}(QK^T)V$

102- Equation represents the scaling step in the Scaled dot-product attention mechanism?

$$\text{Scaled Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

103- Equation represents the calculation of the attention output in the multi head attention mechanism after concatenation?

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W_0$$

104- Equation represents the transformation applied to each attention head in the multi-head mechanism?

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

105- Equation represents the output of the feed-forward neural network (FFN) for a given input X?

$$FFN(x) = W_2(\text{ReLU}(W_1X + b_1)) + b_2$$

عارف استمرارية السعي هتخدلك ايه؟

هتخدلك إنك مش هتندم حتى لو ماوصلتش في الآخر...

106- RNNs are Sequential models, operate in a recurrent manaar.

107- Transformers are non-Sequential models that process data in parallel.

108- Transformers rely on Self-attention mechanisms. They don't have recurrent Connections or hidden states.

109- Transformers Can handle both short & long sequences due to their parallel processing nature.

110- Model architecture is best suited for image Classification task? CNN

"في زحام الأيام لا تنس الغاية التي خُلقت لأجلها، انت هنا وديعة ستُرد إلى ربها يوماً فأحسن"

لو اشتغلت بإخلاص في اي مكان، حتى لو محدش سمع بيك أو شافك، هتلاقي الرزق بييجي عليك، ربنا بيحب عباده المخلصين، واللي ربنا صرح بحبه ميتخافش عليه ♡

ليس كل مُجتهد مُصيب بالضرورة، ولكن مُجتهد مأجور فالحمدُ لله الذي يجازينا على السعي والمحاولات، لا على النتيجة والوصول والكمال.

أذكر نفسي وإياكم بقول الحبيب صلى الله عليه وسلم "من سلك طريقاً يلتمس فيه علماً، سهل الله له طريقاً إلى الجنة ♡"