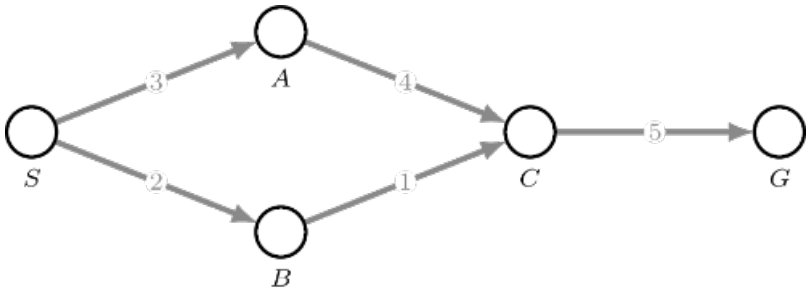
Student number: _____

# of Questions: 39

Total Exam Points: 90.00

Question #: 1

Consider the following graph where nodes represent states, edges represent the cost to move from one state to another, and heuristic values are provided for each node.



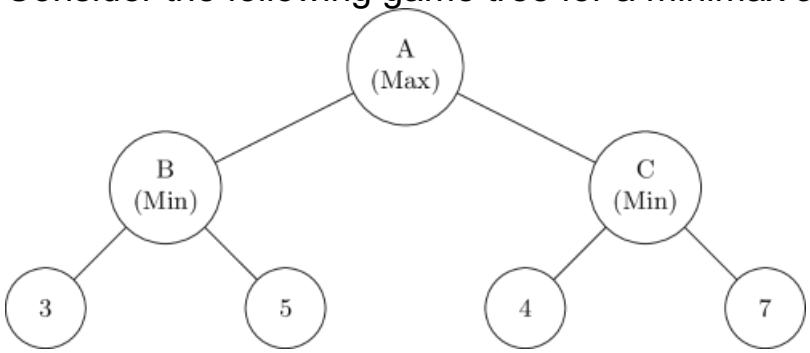
You are tasked with finding the optimal path from S to G using the A* search (tree) algorithm with two different heuristic functions: $h_1(n) = 2$ and $h_2(n) = 0$ for all nodes n. Which of the following statement(s) about A* search and these heuristics is/are correct?

A. A* search with h_1 would not be able to find the optimal path.
B. Only A* search with h_2 finds the optimal path.
C. Both A* search with h_1 and A* search with h_2 will find the optimal path.
D. None of the above.

Item Weight: 2.0

Question #: 2

Consider the following game tree for a minimax algorithm with alpha-beta pruning:



Which of the following statements is correct?

A. For left-to-right, node 7 is pruned. For right-to-left, node 3 is pruned.
B. For left-to-right, node C is pruned. For right-to-left, node B is pruned.
C. For left-to-right, node 7 is pruned. For right-to-left, no pruning.
D. No pruning for both directions.
E. None of the above.

Item Weight: 2.0

Question #: 3

Which of the following statements about Lasso, Ridge, Linear and Logistic Regression is/are correct?

A. Both Lasso and Ridge Regression are equally effective at removing irrelevant features.
B. Compared to Lasso and Ridge Regression, Linear Regression is more prone to overfitting when there are many irrelevant features.
C. Logistic Regression can be regularized using either Lasso or Ridge techniques to prevent overfitting.
D. The final output for Logistic Regression is continuous, similar to Linear Regression, but regularized differently.
E. None of the above.

Item Weight: 2.0

Question #: 4

You are minimizing the cost function $J(w) = 1/2\ w^2 + 4w$ using gradient descent, what is the **largest** learning rate α that can be used so that it always finds the optimal value regardless of the initial value?

A. α=0.5
B. α=1
C. α=2
D. α=4
E. None of the above

Item Weight: 4.0

Question #: 5

Consider a logistic regression model for multi-class classification with three classes: Pizza, Burger, and Sushi. The weight vectors for our multi-class (One vs One) classifiers where the $h_{A/B}(x)$ represents the probability of the class A. The weight vectors for each classifier include the bias term as the first element in each weight vector (2 is the bias for $w_{Pizza/Burger}$).

$w_{Pizza/Burger}$ =[2, -0.5, 0.3]

$w_{Sushi/Pizza}$ =[-1, 0.2, -0.4]

$w_{Burger/Sushi}$ =[0, 0.4, 0.1]

Given the input [3, 2], determine which class the model predicts:

A. Pizza
B. Sushi
C. Burger
D. All classes have equal probability
E. None of the above

Item Weight: 4.0

Question #: 6

The support vector machine maximizes the decision boundary.

A. True
B. False

Item Weight: 2.0

Question #: 7

For SVMs, which of the following statements is/are true?

A. Given linearly separated data, hard-margin SVMs handle noise in the data better than logistic classifiers.
B. All training points become support vectors.
C. The SVM allows for the w parameters to be represented as a linear combination of the training data.
D. They cannot be applied to multi-class classification.
E. None of the above.

Item Weight: 2.0

Question #: 8

Which of the following statements about the primal formulation of Support Vector Machines (SVMs) is/are correct?

A. It is a non-convex optimization problem.
B. It is an optimization problem with inequality constraints.
C. The primal formulation does not obtain the offset b.
D. Both primal and dual formulation are valid formulations for SVMs.
E. None of the above.

Item Weight: 2.0

Question #: 9

Which of the following statements about the dual formulation of Support Vector Machines (SVMs) is/are correct?

A. From the solution of the dual formulation, we cannot compute the solution of the primal formulation.
B. The dual formulation obtains a different decision boundary.
C. The dual formulation is trained with mini-batch descent.
D. The dual formulation allows the use of kernel functions for classification of non-linear data.
E. The dual formulation involves dot products between all the training points.
F. None of the above.

Item Weight: 2.0

Question #: 10

Recall the decision rule for SVMs for deciding if a point x is classified as +or -. Let

$$w = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$$

$$b = -10$$

The point

$$x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

is classified as:

A. +
B. -
C. Cannot tell.

Item Weight: 2.0

Question #: 11

Given the point

$$x = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

and the decision boundary defined by

$$w = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, b = 4$$

What is the Euclidean distance between the point and the decision boundary?

A. 0
B. 1
C. 2
D. $\sqrt{2}$
E. 3
F. $\sqrt{3}$
G. 5
H. $\sqrt{5}$
I. None of the above.

Item Weight: 4.0

Question #: 12

Consider the Perceptron as discussed in the lecture. Select all statement(s) that correctly describe the Perceptron.

A. The input to the activation function of the perceptron is a linear combination of features.
B. The Perceptron is inspired by the transformer architecture in LLMs.
C. The Perceptron has a probabilistic output.
D. The Perceptron can be stacked into multi-layer architectures.
E. None of the above.

Item Weight: 2.0

Question #: 13

Select all the correct statement(s) regarding the Perceptron.

A. The Perceptron cannot be used with transformed features.
B. The activation function of the Perceptron is used to transform every input feature individually.
C. The original Perceptron has an output in the real-number interval [-1,1].
D. None of the above.

Item Weight: 2.0

Question #: 14

You are given training data as in the following table and are training a Perceptron.

x_1	x_2	y
-1	-1	1
-2	1	1
-0.5	1	-1

.

Perform a single step of the Perceptron update rule starting from the weights $w = [3, -1, -2]^T$

.

What are the weights obtained with learning rate 2?
Put your weights in correct ordering into the blanks: [__1__,__2__,__3__]^T

1. _____
2. _____
3. _____

Item Weight: 4.0

Question #: 15

You decide to replace the hidden layer activation function with a ReLU activation function.
Select the mathematical expression that describes the forward propagation to compute
$\hat{y}$
in this case.

A. $\mathrm{Relu}((\mathbf{W}^{[2]})^T(\sigma(((\mathbf{W}^{[1]})^T\mathbf{x}+\mathbf{x})))$

B. $\mathrm{Relu}((\mathbf{W}^{[2]})^T(\sigma(((\mathbf{W}^{[1]})^T\mathbf{x})+\mathbf{x}))$

C. $\sigma((\mathbf{W}^{[2]})^T(\mathrm{Relu}(((\mathbf{W}^{[1]})^T\mathbf{x}+\mathbf{x})))$

D. $\mathrm{Relu}((\mathbf{W}^{[1]})^T(\sigma(((\mathbf{W}^{[2]})^T\mathbf{x}+\mathbf{x})))$

E. $\mathrm{Relu}((\mathbf{W}^{[1]})^T(\sigma(((\mathbf{W}^{[2]})^T\mathbf{x})+\mathbf{x}))$

F. $\sigma((\mathbf{W}^{[1]})^T(\mathrm{Relu}(((\mathbf{W}^{[2]})^T\mathbf{x}+\mathbf{x})))$

G. $\sigma((\mathbf{W}^{[2]})^T(\mathrm{Relu}(((\mathbf{W}^{[1]})^T\mathbf{x})+\mathbf{x}))$

H. $\sigma((\mathbf{W}^{[1]})^T(\mathrm{Relu}(((\mathbf{W}^{[2]})^T\mathbf{x})+\mathbf{x}))$

I. None of the above.

Item Weight: 3.0

Question #: 16

Using the definition

$$\delta^{[2]} := \frac{\partial \hat{y}}{\partial \boldsymbol{f}^{[2]}}$$

,

select the correct expression for

$$\frac{\partial \hat{y}}{\partial \boldsymbol{W}^{[2]}}$$

.

A. $g^{[1]}(\boldsymbol{f}^{[1]})\delta^{[2]} + \boldsymbol{x}$

B. $g^{[1]}(\boldsymbol{f}^{[1]})\delta^{[2]} + g^{[2]}(\boldsymbol{x})$

C. $(g^{[1]}(\boldsymbol{f}^{[1]}) + \boldsymbol{x})\delta^{[2]}$

D. $(\boldsymbol{f}^{[1]} + g^{[2]}(\boldsymbol{x}))\delta^{[2]}$

E. $(\boldsymbol{f}^{[1]} + \boldsymbol{x})\delta^{[2]}$

F. $\boldsymbol{f}^{[1]} + g^{[2]}(\boldsymbol{x})\delta^{[2]}$

G. $\boldsymbol{f}^{[1]}\delta^{[2]} + \boldsymbol{x}$

H. $(g^{[1]}(\boldsymbol{f}^{[1]}) + g^{[2]}(\boldsymbol{x}))\delta^{[2]}$

I. None of the above.

Item Weight: 3.0

Question #: 17

Consider the following dataset consisting of three points in 2D space:
A =(5, 0)
B =(1, 0)
C =(0, 1)
We initialized with two cluster centers by randomly choosing 2 points from the data points A,B,C. We run the K-means algorithm, using Euclidean distance, until convergence. Which of the following represents the final locations of the cluster centers after the algorithm converges?

Fill in the blanks with the coordinates of the two final cluster centers: (__1__,__2__) and (__3__,__4__). Arrange the cluster centers by their Euclidean norm, placing the coordinates with the larger norm in the first pair of blanks and the coordinates with the smaller norm in the second pair. Use decimal format for fractions (e.g., 0.25).

1. _____
2. _____
3. _____
4. _____

Item Weight: 4.0

Question #: 18

Which of the following statements is/are true with respect to the K-means algorithm?

A. The K-means algorithm always converges.
B. The K in K-means is learned within the K-means algorithm.
C. The K-means algorithm always converges to the global minimum.
D. The K-means algorithm can be configured to use Manhattan Distance.
E. None of the above.

Item Weight: 2.0

Question #: 19

Does the K-means algorithm sometimes converge to different clusterings on the same dataset? Why?

A. K-means is a non-deterministic algorithm because it randomly changes the cluster centroids during iterations.
B. The final clusters depend on the initial placement of centroids, which can be randomized.
C. K-Means is a deterministic algorithm and will always converge to the same final clusters regardless of the initial centroids given.
D. The final clusters may change because the number of clusters k is automatically adjusted during training.
E. None of the above.

Item Weight: 2.0

Question #: 20

Consider the following distance matrix representing the distances between five data points A, B, C, D, E:

	A	B	C	D	E
A	0	5	3	7	6
B	5	0	5	8	7
C	3	5	0	6	5
D	7	8	6	0	4
E	6	7	5	4	0

Remember that in Single Linkage hierarchical clustering, we are using the distance between the closest elements in the clusters. Using Single Linkage hierarchical clustering, we first merge A and C into a new cluster {A, C}. Compute the updated distance matrix if needed. Based on this matrix, identify which two points/clusters should be merged next?

A. {A, C} and B
B. D and E
C. {A, C} and D
D. {A, C} and E
E. None of the above.

Item Weight: 2.0

Question #: 21

In which of the following scenarios is the curse of dimensionality likely to impact model performance? Select all that apply.

A. When using a dataset with many samples relative to the number of features.
B. When all features in the dataset are highly correlated.
C. When using a simple linear model on low-dimensional data.
D. When using a neural network on a dataset with thousands of dimensions but limited samples.
E. None of the above.

Item Weight: 2.0

Question #: 22

Which of the following is the primary goal of Principal Component Analysis (PCA) in data processing?

A. To eliminate all correlations between features in a dataset.
B. To reduce the dimensionality of the data by finding a new set of orthogonal axes that capture the maximum variance.
C. To increase the number of features by creating new, independent features from the original dataset.
D. To standardize the data by ensuring all features have a mean of zero and a variance of one.
E. None of the above.

Item Weight: 2.0

Question #: 23

Which of the following task(s) can the wine dataset be used for?

A. Classification task, predicting the geographic origin
B. Classification task, predicting wine quality levels
C. Regression task, predicting wine yield
D. Regression task, predicting wine quality
E. None of the above

Item Weight: 2.0

Question #: 24

Which of the following preprocessing steps would be the most suitable to address missing values in this dataset before applying linear regression?

A. Remove all samples with missing values
B. Fill the missing values with zeros
C. Remove all attributes (columns) with missing values
D. Ignore the missing values as they will not impact model performance
E. None of the above

Item Weight: 2.0

Question #: 25

When applying distance-based models (models that rely on distance calculations, for example, using euclidean distance), which feature transformations is/are important to improve model performance?

A. Mean normalization
B. Min-max scaling
C. No feature transformations are required
D. Binning to reduce feature value diversity
E. None of the above

Item Weight: 2.0

Question #: 26

Your friend suggests using linear regression (using the normal equation) to predict wine quality as a numerical score. What advice(s) would you give them about preparing the data and key model considerations?

A. Ensure that features are scaled appropriately, as linear regression is sensitive to the scale of input features.
B. Check for and handle highly correlated features, as they can affect model stability and interpretability.
C. Recommend logistic regression, as it is better suited for predicting numerical scores.
D. Address potential outliers, as they may disproportionately influence the model's parameters.
E. None of the above.

Item Weight: 2.0

Question #: 27

Considering linear regression model is applied to the first 2 features ('fixed acidity' and 'volatile acidity') as X_1,X_2 respectively and θ_i are the respective parameters, which of the following represent the *best* linear regression model to predict Y 'quality' (where *best* refers to a valid model with the most appropriate transformation of the input features)?
Hint: Consider how the transformations exp and log impact outliers.

A. $Y = \theta_0 + \theta_1 X_1$
B. $Y = \theta_0 + \theta_2 X_2$
C. $Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2$
D. $Y = \exp(\theta_0 + \theta_1 X_1)$
E. $Y = \exp(\theta_0 + \theta_2 X_2)$
F. $Y = \exp(\theta_0 + \theta_1 X_1 + \theta_2 X_2)$
G. $Y = \log(\theta_0 + \theta_1 X_1)$
H. $Y = \log(\theta_0 + \theta_2 X_2)$
I. $Y = \log(\theta_0 + \theta_1 X_1 + \theta_2 X_2)$
J. $Y = \theta_0 + \theta_1 X_1 + \theta_2 \log(X_2)$
K. $Y = \theta_0 + \theta_1 \log(X_1) + \theta_2 X_2$
L. $Y = \theta_0 + \theta_1 \log(X_1) + \theta_2 \log(X_2)$
M. $Y = \theta_0 + \theta_1 X_1 + \theta_2 \exp(X_2)$
N. $Y = \theta_0 + \theta_1 \exp(X_1) + \theta_2 X_2$
O. None of the above

Item Weight: 2.0

Question #: 28

In modeling a classification task to predict the type of wine variety (red/white), which model(s) might you try after properly preprocessing the data?

A. Recurrent neural network
B. K-means
C. Decision tree
D. Linear regression
E. Hard-margin support vector machine
F. None of the above

Item Weight: 2.0

Question #: 29

After much discussion, your friend decided to implement a custom soft-margin support vector machine to predict the type of wine variety (red/white) and managed to achieve 70% test accuracy. Which statements is/are true about the test accuracy of their model?

A. The test accuracy will improve if the margin is increased.
B. The model performs better than random classification.
C. Test accuracy will remain unaffected by feature scaling.
D. None of the statements are correct.
E. None of the above.

Item Weight: 2.0

Question #: 30

When predicting wine variety (red/white), which evaluation metric is the most appropriate to evaluate the model's performance?

A. Accuracy
B. Precision and recall
C. Mean squared error
D. Weighted binary cross entropy loss
E. None of the above

Item Weight: 2.0

Question #: 31

Let $h(x_1, x_2)$ be the multi-layer Perceptron architecture for the XNOR(x_1, x_2). Based on the data, we can associate a subset of {A, T, G, C} with a Boolean variable for which it holds
that Position 3 is predicted by $h(x_1, x_2)$. What is the subset?

A. {A,T}
B. {T,G}
C. {T,C}
D. {A,C}
E. {A,G}
F. {G,C}
G. None of the above

Item Weight: 2.0

Question #: 32

Define the following step function for the Perceptron:

$$Step(x) = \begin{cases} 1 \text{ if } x \geq 0, \\ 0 \text{ otherwise}. \end{cases}$$

There is the other subset of {A,T,G,C} distinct from the answer to Question 31. You notice that the relationship of Position 1 and 2 with Position 3 within this subset can also be described by a Boolean function. This relationship is best described by which of the following Perceptrons?

A. $\text{Step}(x_1 + x_2).$

B. $\text{Step}(x_1 + x_2 + 1.5).$

C. $\text{Step}(-x_1 + x_2).$

D. $\text{Step}(x_1 - x_2 - 1.5).$

E. $\text{Step}(x_1 - x_2 + 1.5).$

F. $\text{Step}(x_1 + x_2 - 1.5).$

G. $\text{Step}(-x_1 - x_2 - 1.5).$

H. $\text{Step}(-x_1 - x_2).$

I. $\text{Step}(-x_1 - x_2 + 1.5).$

J. None of the above.

Item Weight: 2.0

Question #: 33

You discard the previous small subset as too simplistic and move on to longer sequences. Your colleagues have given you a deep neural-network classifier called ERNIE that was trained on 6-letter sequences and that is able to classify 6-letter sequences into 10 groups W_1, ..., W_{10}.
For the output part of the ERNIE model, you expect to have one of the following:

A. Several one-versus-all multi-class layers.
B. A softmax layer that gives probabilities p_i for group i.
C. A layer of nodes for one-versus-one classification.
D. A single logistic function that gives the probability for each group i.
E. None of the above.

Item Weight: 2.0

Question #: 34

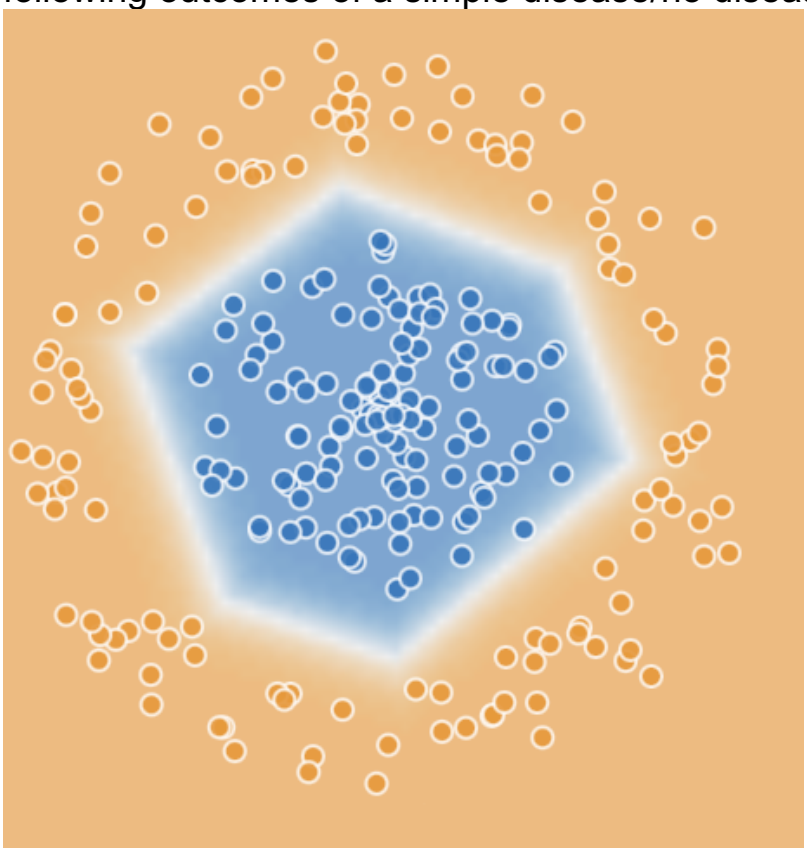
By looking at the groups produced by ERNIE, you note that group W_5 contains all 6-letter sequences that show up in data relevant for Parkinson's disease. You take one of the human sequences
GATACCTTCA...ACGGATTATT...TTAACCATCTC
and use the classifier on the marked part of the sequence and obtain the highest probability for group W_5. Based on this output, we deduce that this person is at risk for the disease. Select some flaw(s) in your procedure.

A. The marked sequence does not fit the input of the classifier.
B. The sequence GGATTA may be part of another subsequence.
C. We should apply the classifier to the letters A,G,T,C only.
D. The meaning of marked part may depend on the preceding and subsequent parts of the sequence.
E. None of the above.

Item Weight: 2.0

Question #: 35

You decide to talk more to the deep-learning colleagues and they show the following outcomes of a simple disease/no disease classifier.



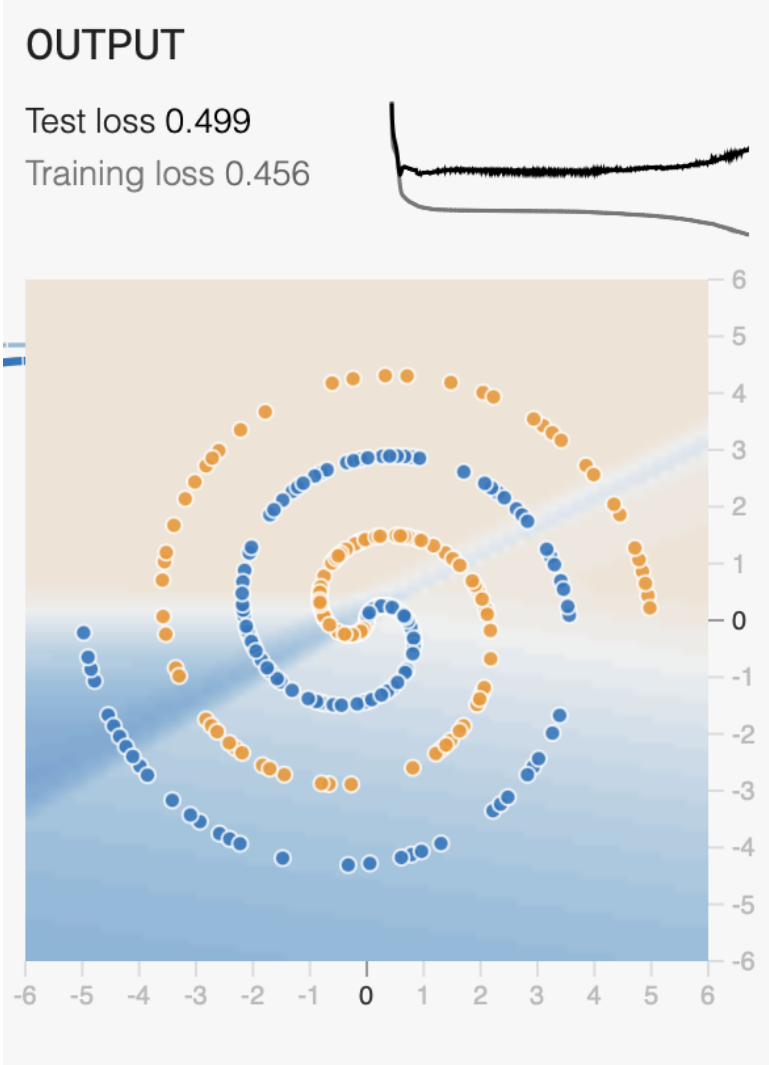
What activation function did they use in their neural network?

A. Linear function.
B. Sigmoid function.
C. ReLU function.
D. Tanh function.
E. Cannot tell.

Item Weight: 2.0

Question #: 36

You are shown the following result, which seems problematic. The result is after 1300 epochs of stochastic gradient descent with step size 0.03, with a model that has tanh activation functions.



Select from the following options which advice(s) you give to your colleagues.

A. The test loss is too high, so they should use more test examples.
B. Keep running for more epochs.
C. Use sigmoid activation functions.
D. Use a less complex model.
E. Use transformed features.
F. None of the above.

Item Weight: 2.0

Question #: 37

You decide to use modern transformer architectures to learn about the complete dataset from Duke-NUS.
What is/are the advantages of using the Transformer architecture with the self-attention mechanism over RNN architectures.

A. The transformer takes into account the relationships between all the parts of the sequences.
B. The transformer is pre-trained and hence does not require training.
C. It can be trained by convex optimization.
D. The transformer architecture in general has less parameters.
E. The transformer does not have a sequential memory bottleneck.
F. None of the above.

Item Weight: 2.0

Question #: 38

For using the transformer with DNA sequences, you have to come up with an encoding. You decide to use an encoding analogous to the word-encoding shown for RNNs in class and define all possible DNA sequences of length 10 as all possible words. These words we can also call **tokens**. What is the dimension of your encoding vectors using this encoding?

A. 4
B. 2^{10}
C. 10^4
D. 2^{20}
E. None of the above.

Item Weight: 2.0

Question #: 39

As in Question 38, define all possible words/tokens as all possible sequences of length 10. Using this definition, for a given sequence of DNA, we can tokenize the sequence by splitting the sequence into sequences of length being the token length. As an example, the sequence ATAATAAACTCGCGTATGCG...

is tokenized as |ATAATAAACT|CGCGTATGCG|...

with tokens |token 1|token 2| ...

The transformer includes a self-attention part with a set of matrices for query, key, and value parts. The attention scores can be summarized in a matrix that depends on query and key matrix. Based on the complete DNA sequences collected by Duke-NUS and our tokenization, the matrix dimension of the attention score matrix is given by the following.

A. $10^4 \times 10^4$

B. $10^4 \times 10^5$

C. $10^5 \times 10^5$

D. $10^5 \times 10^6$

E. $10^6 \times 10^6$

F. $10^6 \times 10^7$

G. $10^7 \times 10^7$

H. None of the above.

Item Weight: 2.0

End of Paper