Solution sheet for Final Exam CS2109S AY 24/25 Semester 1 (Version 2)

# of Questions: 39, **Total Exam Points:** 90.00

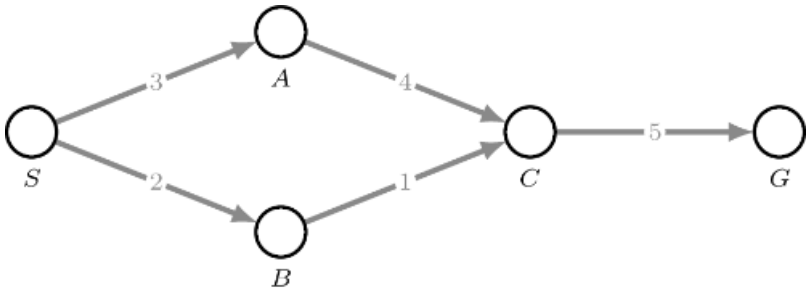
Answer sheet:

Q1	C		Q21	D B&D
Q2	D		Q22	B
Q3	B&C		Q23	B&D
Q4	B E		Q24	A
Q5	A		Q25	A&B
Q6	B (False)		Q26	B&D
Q7	A&C		Q27	L K
Q8	B&D		Q28	A&B&C&E
Q9	D&E		Q29	B
Q10	B		Q30	B
Q11	E		Q31	E
Q12	A&D		Q32	F
Q13	C D		Q33	B
Q14	[-1,1,-6]		Q34	B&D
Q15	G		Q35	B C D E
Q16	C		Q36	B&E B&D&E B&C&E B&C&D&E
Q17	5, 0, 0.5, 0.5 0.5, 0.5, 5, 0 (Partial)		Q37	A&E E
Q18	D A&D		Q38	D
Q19	B		Q39	C
Q20	B			

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.

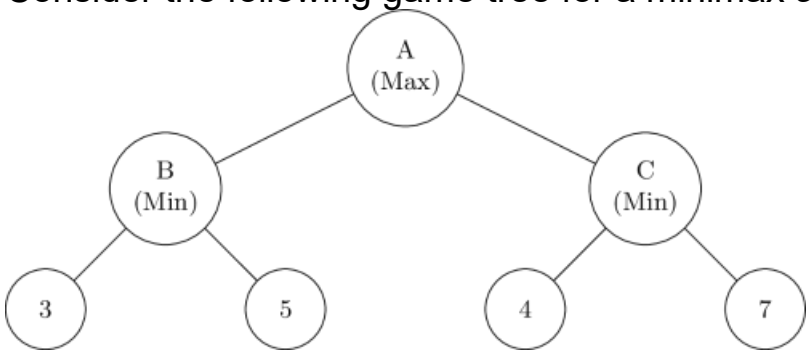
Item Weight: 2.0

Solution: C.

Explanation: Normally in CS2109s, we see undirected graphs but for this question we test the directed case – the algorithm works the same (majority got this correct). The trick to doing this question fast is to notice that both A* with constant h1 and h2 heuristics would expand the nodes like UCS, finding the optimal path. Students took more than 5 mins, which indicates that most students did not have this insight. Note that we do not allow reasoning based on the admissibility of h1.

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

Solution: D.

Explanation: The trick to doing this question is to notice that alpha beta minimally needs to visit some nodes before pruning can happen – which means for left to right, we need to visit up to leaf 4, and for right to left visit up to leaf 5. From this, we can eliminate option B. We can work out more for each case:

- Left to right: Min player on the left branch chooses 3 and alpha is set to 3. When the right branch visits 4, it is that alpha < 4, hence the next node will not be pruned. We can eliminate options A and C
- Right to left: Exploring the right branch sets alpha=4. When arriving at node 5 it holds that 5 > alpha hence no pruning.

The answer would be D no pruning for both directions.

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
Solution: B&C.
Explanation: Notice that Lasso and Ridge Regression are Linear Regression with L1 and L2 regularization, discussed in tutorials.
A. Incorrect. The Lasso method is better at removing irrelevant features (feature selection) due to its ability to set coefficients to zero.
B. Correct. Regularization is a technique to combat overfitting – any regularization (Lasso and Ridge) is better than none (Linear Regression). Lasso and Ridge regression was defined in the tutorials. Overfitting arises from large weights given to less relevant features (for example, high-degree monomial terms).
C. Correct. This answer tests the subtle distinction between the concepts of Lasso/Rigde regression, and a possible view of them as general techniques that can be applied in various scenarios (beyond MSE for example). The Lasso or Ridge techniques are regularizations using L1 and L2 norms, respectively.
D. Incorrect. The statement is partially correct, probability is continuous, but it is thresholded to finally produce a class label.

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

Solution: B or E

Explanation: The first thing is to compute the derivative $\frac{dJ}{dw} = w + 4$. The equation is a convex parabola – you can do this by observation or remembering $a = \frac{1}{2} > 0$ or $\frac{d^2J}{dw^2} > 0$. Following which, by solving the first derivative for the minimum (first-order condition), you will get optimal value $w^{\cdot} = -4$. This means that the gradient descent has to converge to –4 for whatever values for the value of alpha we are looking for. Since we are interested in the largest value, we can consider from the largest to the smallest, plugging in the equation of gradient descent - $w_{i+1} = w_i - \alpha \times \frac{dJ}{dw}$ which can be expanded to $w_{i+1} = w_i - \alpha \times (w_i + 4)$ which can be expressed as $w_{i+1} = w_i(1 - \alpha) - 4\alpha$.

- $\alpha = 4$, starting from $w = 0 \rightarrow -16 \rightarrow 32 \rightarrow -112 \rightarrow 320 \cdots$ which oscillates between +ve and –ve numbers and having divergent behavior.
- $\alpha = 2$, starting from $w = 0 \rightarrow -8 \rightarrow 0 \rightarrow -8 \rightarrow 0 \cdots$ which oscillates between 0 and –8, which does not converge.
- $\alpha = 1$, starting from $w = 0 \rightarrow -4 \rightarrow -4 \rightarrow -4 \rightarrow \cdots$ which converges in the first iteration.

Hence, $\alpha = 1$ may be the largest learning rate. Remembering that the question specifies regardless of the initial value, we perform more analysis. Substituting $\alpha = 1$, $w_{i+1} = w_i(1 - \alpha) - 4\alpha \Rightarrow w_{i+1} = -4$. Hence, option B would converge regardless of the starting value of w.

Technically, since we are asking for the largest α, we can stop at $\alpha = 1$. However, you may continue for $\alpha = 0.5$ in the same manner, after substituting, $w_{i+1} = w_i(1 - \alpha) - 4\alpha \Rightarrow w_{i+1} = \frac{1}{2}w_i - 2$. In general, $w_N = \frac{1}{2}w_{N-1} - 2 = \frac{1}{2}\left(\frac{1}{2}w_{N-2} - 2\right) - 2 = \frac{1}{2^2}w_{N-2} - 1 - 2 = \frac{1}{2^N}w_0 - \sum_{n=1}^{N} 2\left(\frac{1}{2}\right)^{n-1}$. Taking limits, $\lim_{N\to\infty}\left\{\frac{1}{2^N}w_0 - \sum_{n=1}^{N} 2\left(\frac{1}{2}\right)^{n-1}\right\} = \lim_{N\to\infty}\left(\frac{1}{2^N}\right)w_o - \frac{2}{1-\frac{1}{2}} = 0 - 4 = -4$.

As the largest value for convergence is $\alpha < 2$, we also accept E.

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

Solution: A
Explanation: Remember to extend the input with the bias term, to be [1,3,2]. We compute the probabilities similar as to tutorials:

- wPizza/Burger: 1/(1+exp(-(2 – 1.5 + 0.6))) = 0.7502601055951177
- wSushi/Pizza: 0.23147521650098238
- wBurger/Sushi: 0.8021838885585818

Remember we are doing One vs One here, so

- Pizza/Burger classifies as Pizza
- Sushi/Pizza classifies as Pizza
- Burger/Sushi classifies as Burger

We obtain Pizza twice hence conclude Pizza, which is option A.

Question #: 6

The support vector machine maximizes the decision boundary.

A. True
B. False

Item Weight: 2.0
Solution: B
Explanation: The SVM maximizes the margin. Maximizing the decision boundary does not have a meaning.

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

Solution: A&C.

Explanation:

A: Correct. Robustness property is mentioned on Slide 58 of Lecture 8. Noise robustness follows from having a maximum margin (out of all possible margins).
B: Incorrect as the support vectors are only the vectors that are on the margin. Generally, not all vectors will be on the margin.
C: Correct, see lecture slides.
D: Incorrect. We have discussed using logistic regression for multi-class classification. The same techniques can be used for the SVM.

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
Solution: B & D.

Explanation:

A: Incorrect. The objective of the primal formulation is a convex function. The constraints are linear. Hence the convex function is optimized over a convex set.
B: Correct. The constraints are inequality constraints.
C: Incorrect. The primal formulation obtains the offset b (Lecture 7 Slide 44)
D: Correct, as discussed.

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

Solution: D & E

Explanation:

A: Incorrect. The solutions of the dual obtain the alpha's from which we can compute the w and b. (L7 P44)
B: Incorrect. The decision boundary is defined by w and b, which are obtained from the alpha's.
C: Incorrect. Mini-batch descent is typically not used to train the SVM.
D: Correct. As discussed in L7.
E. Correct. As discussed in L7.

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

Solution: B

Explanation: Use the decision rule w.x + b to compute 4+5-10 = -1, hence it is classified as -.

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

Solution: E

Explanation: Find a point that is on the decision boundary by solving w.x + b = 0. For example, z = [0,-1] satisfies this. Now use the margin computation from L7 P42. Imagine a parallel decision boundary going through point x. From the margin computation, we have (x-z).w/||w|| = (11 + 4)/sqrt(9 + 16) = 15 / 5 = 3.

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

Solution: A & D

Explanation:

A: Correct. We input the linear combination w.x. The bias is included via a dummy feature set to 1.
B: Incorrect. Perceptron is a historic idea inspired from biological neurons and is a precursor to modern deep learning and transformers.
C. Incorrect. It has a deterministic output.
D. Correct: Multilayer Perceptrons (MLPs).

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

Solution: C or D.

Explanation:

A: Incorrect. We can transform input features similarly to other models.
B: Incorrect. The activation function acts on a linear combination of features, not each feature individually.
C. Correct. The original Perceptron's output is {-1,1} and is a subset of [-1,1]. However, the language can be misinterpreted, hence choices of correct and incorrect are allowed here.
D. Is a valid choice because of the ambivalence of option C.

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

Solution: [-1, 1, -6]^T.

Explanation: Given the initial weights with the bias weight 3, the first misclassified example is the third example. This is because 3*1+0.5*1 – 2 = 1.5 leads to +1.
The update rule: w^next = w + 2 (-1 –1) [1 –0.5 1]^T = [3, -1, -2]^T - 4 [1, -0.5, 1]^T = [-1, 1, -6]^T.

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
Note: As corrected during the exam, the three parenthesis "(((" should be replaced by "((" for all options.
Solution: G.

Explanation: The input is multiplied by the weight matrix W[1], then the ReLU is applied, then the x is added, then the weight matrix W[2] is applied, then the sigmoid is applied.

Question #: 16

Using the definition $\delta^{[2]} := \frac{\partial \hat{y}}{\partial f^{[2]}}$,select the correct expression for $\frac{\partial \hat{y}}{\partial W^{[2]}}$.

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

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

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

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

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

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

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

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

I. None of the above.

Item Weight: 3.0

Solution: C.

Explanation: dy/dW[2] = dy/df[2] df[2]/dW[2] = delta[2] a[1] = delta[2] (g[1](f[1])+x)

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

Solution: 1. 5; 2. 0; 3. 0.5; 4. 0.5. (We also accepted equivalent strings, eg. 5.0 for 5. We give partial credit for switching the order of the two final cluster centers.)
Explanation: The cluster centers are (5,0) and (0.5,0.5). Regardless of the choice of initial points, k-means will converge to the same cluster centers for this problem.

Choose A and B as initial cluster centroids c1 and c2. Assignment step: C is assigned to cluster c2. Centroid computation: c2 becomes (0.5,0.5), c1 does not change. Assignment step: B and C assigned to c2, A to c1. Converged. Norm of c1 (Euclidean distance, so L2 norm) = 5 which is larger than norm of c2 = 0.7071.

You can use other initial points from A,B,C and you would arrive at the same answer.

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
Solution: D or A&D.
Explanation:
A: Both options accepted. K-means always converge only when the tie breaking strategy is deterministic, which was not specified in this question. We accept both answers here, considering that you may or may not assume deterministic tie-breaking.
B: Incorrect. K is a user-chosen hyperparameter.
C: Incorrect. K means converges to a local optimum.
D: Correct. Instead of Euclidean distance, Manhattan distance can be chosen.

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

Solution: B.

Explanation:

A: Incorrect reason. K-means algorithm does not randomly change cluster centroids.

B. Correct. Different initial centroids may lead to different clustering.

C. Incorrect. Different initial centroids may lead to different clustering.

D. Incorrect. Number of clusters is a hyperparameter and it is not learnt.

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

Solution: B.

Explanation: After the first merging the single link distances of this cluster to all other points are B:5, D:6, E:5. All of these distances are greater than distance between D and E, 4, hence D and E will be merged next.

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
Solution: D or B&D

Explanation: The curse of dimensionality in the context of the lecture is that, for learning a hypothesis class of functions of many variables, one may require a number of samples exponential in the number of variables to learn a hypothesis from this class.
A. Incorrect. The curse of dimensionality discussed in lecture does not negatively impact model performance in this case, as we have many samples compared to the number of features.
B. Partially correct. If features are highly correlated, the effective dimensionality of the dataset is reduced because the correlated features contain redundant information, which can be reduced using dimensional reduction techniques like PCA, therefore mitigating the effects of the curse of dimensionality. We allow both options for B.
C. Incorrect. Low dimensional data means small number of features.
D. Correct. We may not have enough samples for this complex model.

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
Solution: B
Explanation:
A. Incorrect. While PCA does transform features into a new set of uncorrelated (orthogonal) axes, it is not the primary goal of PCA and we cannot expect to eliminate all correlations.
B. Correct. This was explained in lecture 11, and PS5.
C. Incorrect. PCA reduces the number of features, not increases them.
D. Incorrect. Standardization is often a preprocessing step before applying PCA but is not the goal of PCA itself.

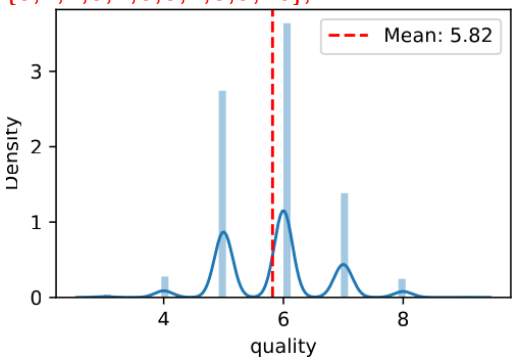
Question #: 23

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

Solution: B&D
Explanation:
A. Incorrect. Geographic origin is not part of the data.
B. Correct. From the graphs in the context document, the wine quality levels are clearly binned to {0,1,2,3,4,5,6,7,8,9,10}; hence we can use multiclass classification to classify wines.



C. Incorrect. Yield is not part of the data.
D. Correct. Even though within the context document the wine quality is shown to have integer values, wine quality is inherently a continuous variable that is binned. You can infer this by considering that a wine with quality 10 is better than a wine with quality 0 – essentially the various quality levels have some ordering.

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
Solution: A
Explanation:
A. Correct. Removing training examples will train on a smaller training set without missing values and with the same number of attributes – this is the best choice.
B. Incorrect. Setting to zero will lead to fitting to these zeros as well. The true values could be far from zero – this would be the third worse choice.
C. Incorrect. Removing the attribute will lead to removing 7 attributes out of 12, which is more than half of the attributes for linear regression to use – this would be the second worse choice.
D. Incorrect. Ignoring attributes for some training examples is impossible as linear regression considers all of the selected attributes – this would be the worse choice.

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

Solution: A&B.

Explanation: Distance-based models (e.g., k-Nearest Neighbors, k-Means Clustering, etc...) rely on calculating distances (e.g., Euclidean distance) between data points. When features have different scales, features with larger ranges can dominate the distance calculations, leading to biased model performance. Hence, anything that reduces the dominance of any feature within the distance calculation would be important to consider.

A: Correct. Based on class and forum discussion, there are two possible interpretations:

- "Mean normalization" (aka. Standardization) can as shown in L5 Slide 45 as $x_j = \frac{x_j - \mu_j}{\sigma_j}$.
- Mean normalization as $\frac{x - \overline{x}}{\max(x) - \min(x)}$.

Both interpretations subtract the mean of each feature and divide by a value related to the scale of the features. This dividing by the scale of the feature helps ensure that all features contribute equally to the distance computation. Having features of similar scales is important to model performance.

B: Correct. Min-max scaling scales each feature to a fixed range, say [0, 1], ensuring no feature disproportionately affects the distance.

C. Incorrect. Distance-based models are sensitive to the scale of features, and transformations are often essential. (For example, a log transformation similar to Q27 could be applied).

D. Incorrect. Binning (converting continuous data into categories) reduces information and may lead to poorer model performance for distance-based methods.

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

Solution:B&D
Explanation:
A. Incorrect. The normal equation for linear regression is indeed not as sensitive to the scale of input features, as noted in lecture slides L5. This is because the normal equation analytically computes the solution.
B. Correct. Highly correlated features may lead to dependent columns or rows of the data matrix. This may lead to the problems with invertibility in the normal equation.
C. Incorrect. Logistic regression is intended for classification problems, not for predicting scores in the range of [0,10].
D. Correct. Outliers can have a significant impact on the linear regression model, especially since it minimizes the squared error.

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

Solution: L or K
Explanation: We would need valid linear regression models (D-I are invalid), and from that we are left with options A,B,C,J,K,L,M,N,O. Then, we can consider the appropriate transformations. The logarithmic function suppresses large values by growing slowly as input increases, making it useful for normalizing or compressing data. In contrast, the exponential function amplifies large values by growing rapidly. Studying the fixed/volatile acidity graphs in the Context, both show a distribution where the mode is different and to the left of the mean, and where there is a considerable amount of probability density on the right-hand side of the graph (tail of the distribution / outliers) - applying log for both features would compress large values into a smaller range while preserving the relative order of the data, reducing the skew of the data distribution. Because the graph of volatile acidity shows the potential for very small positive values, and a log transform could induce a corresponding skew towards the left-hand side of the resulting graph, we also accept answer K (log transform only on the first feature.)

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
Solution: A&B&C&E
Explanation:
A. Correct – possible to try: Consider length-1 RNNs, which are just DNNs and can be used as a classifier.
B. Correct – possible to try: It was shown that K-means can also be used for classification in PS5, you may process using techniques similar to PS5.
C. Correct – possible to try: This is the most obvious choice.
D. Incorrect. Linear Regression performs the regression task, we are performing classification.
E. Correct – possible to try: Although the context mentions “that red and white wine samples share a similar distribution of physicochemical properties and cannot be easily distinguished”, we may apply kernel SVM methods with an appropriate transformation that causes the data to be linearly separable.

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

Solution: B.

Explanation:

A. Incorrect. Increasing the margin alone does not change the decision boundary. Hence does not change the classification of test examples, hence does not change the test accuracy.

B. Correct. Random classification without any further information implies that we are classifying with equal probability. From the pricides information, we can additionally take that information into account: note that the data samples are skewed with red (1,599 samples) and white (4,898 samples). Hence, another reasonable assumption we accept for the random classifier is that the classifier classifies with probabilities P(Red)=0.2461 and P(White)=0.7538.

1. Assume random classifier outputs Red or White with equal probability.
    a. The probability of having a correct classification would be P(Red)*P(Predict Red) + P(White)*P(Predict White). (Note that these 2 events are independent)
    b. P(Red)*P(Predict Red) + P(White)*P(Predict White)
    = P(Red)*0.5 + P(White)*0.5 = 0.5 (P(Red) + P(White)) = 0.5
    c. Hence, the accuracy of the random classifier would be 0.5
2. Assume random classification considering data distribution (0.7538-White vs 0.2461-Red)
    a. Same - P(Red)*P(Predict Red) + P(White)*P(Predict White)
    b. P(Red)*P(Predict Red) + P(White)*P(Predict White)
    = 0.2461^2 + 0.7538^2 = 0.628
    c. Hence, the accuracy of the random classifier would be 0.628

Hence the model of 70% performs better than the random classifier in either case. There is not enough information given to us that we can assume that the random classifier has P(White)=0.9 or other highly skewed examples.

C. Incorrect. Feature scaling may improve the classifier and hence the test accuracy.

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

Solution: B.

Explanation: We are asking for the best choice here, so let's rank the options for evaluating a model's performance. From the options, C,D are not appropriate options as evaluation metric. Hence, we are left only with A or B. Given the skewed nature of the data, the majority class (white) dominates. A classifier that predicts "white" for all instances would achieve an accuracy of 0.7538, but this high accuracy could be misleading and fail to reveal very poor performance on the minority class (red). In contrast, Precision and Recall provide a more nuanced analysis by evaluating both false positives and false negatives, evaluating its performance across all classes. Hence, Precision and Recall.

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

Solution: E

Explanation: The set {A,T,G,C} has four elements, hence we need two bits (Boolean variables) to describe which elements in the set we mean. Also, any two-element subset can be associated with a Boolean variable {0,1}. We have two features Position 1 denoted as x_1 and Position 2 denoted as x_2, both are one of {A,T,G,C}. We see that in the data

AA followed by G
AG followed by A
GA followed by A
GG followed by G

Recall the XNOR discussed in class (or the XOR discussed in tutorial). XNOR truth table is:
00 -> 1
10 -> 0
01 -> 0
11 -> 1
Hence with A=0, G =1 we can use the given MLP h(x_1,x_2) to predict Position 3.

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

Solution: F

Explanation: The answer of Q31 is {A,G} hence the distinct subset is {C,T}. From the data:

CC followed by C
TC followed by C
CT followed by C
TT followed by T.

This is an AND relationship when we associate C= 0 and T = 1. Constraining to only this subset:

The Perceptron Step(x_1+x_2 – 1.5) models this relationship, as Step(0+0-1.5) = 0, Step(1+0-1.5) = 0, Step(0+1-1.5) = 0, and Step(1+1-1.5) = 1.

To check a false option: The Perceptron Step(-x_1-x_2 +1.5) we have Step(0+0+1.5) = 1, Step(-1+0+1.5) = 1, Step(0-1+1.5) = 1, and Step(-1-1+1.5) = 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
Solution: B

Explanation:
B. Correct as softmax can output the probabilities for each group.
A. Incorrect, one-versus-all classification is not part of the DNN itself.
C. Incorrect. One-versus-one is not done via a layer of nodes inside the DNN.
D. Incorrect, a single logistic function is not able to give the probability for each group.

Question #: 34

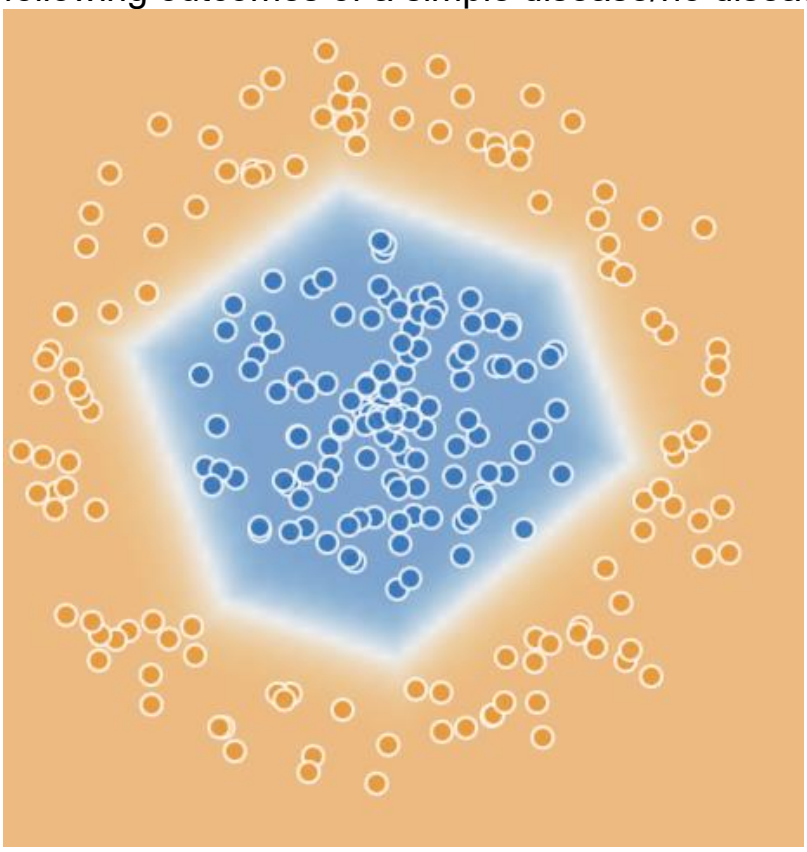
By looking at the groups produced by ERNIE, you note that group W_5 contains all 6-letter sequences that show up in data relevant for Parkinson's disease. You take one of the human sequences
GATACCTTCA...ACGGATTATT...TTAACCATCTC
and use the classifier on the marked part of the sequence and obtain the 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
Solution: B & D.
Explanation: A. Not a flaw. ERNIE can process 6 letter inputs. B. A flaw. In the absence of other information, the six-letter sequence may overlap with another subsequence (of any length, including length 6 subsequences). C. Not a flaw. The classifier was designed for 6 letter sequences of letters A,G,T,C.
D. A flaw. Context is important for sequences (think of language).

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

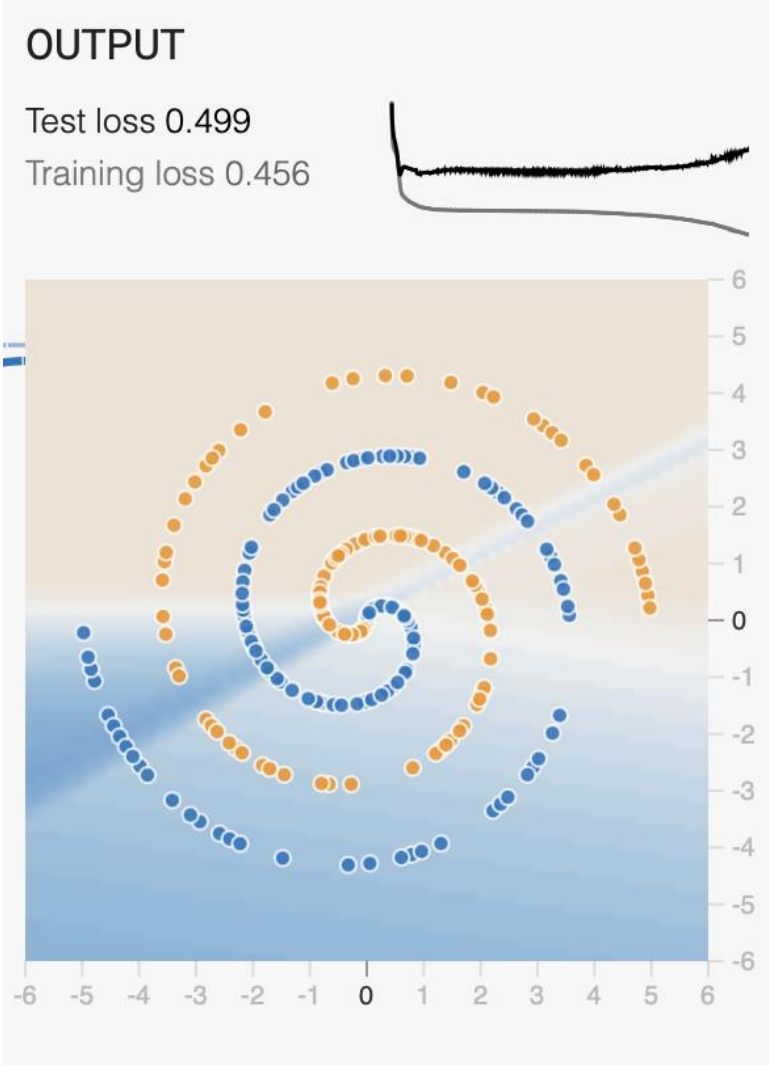
Solution: B or C or D or E - All options B-E are correct (and selecting one of them will lead to full marks).

Explanation: The visualization shows data points being correctly separated into two groups. The decision boundaries appear to be piecewise linear, and the shape of the blue region is a polygon with rather sharp edges. So, the activation function most likely was the ReLU function (and in fact it was), where architectures often lead to these piecewise linear boundaries. However, your colleagues may have trained a sigmoid/tanh network with weights so large that piecewise linear boundaries could arise as well. Hence Options B-E are valid options and lead to full marks. We rule out the activation function being a

linear function, as we would not be able to make such a classifier with only linear functions.

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

Solution: B&E or B&C&E or B&D&E or B&C&D&E.
Explanation: The meaning of the plot is similar to Q36.
A. Incorrect. The training data is the spiral, and it's likely that the test samples also come from the same distribution. Since the classifier is problematic, i.e., it does not classify properly yet, we expect that adding more test examples will not improve the test accuracy, as test examples do not contribute to model learning.

B. Correct. The training error appears to still decrease significantly. At the same time, the classifier is not very good. The test error going up could be an indication of overfitting. However, it does not look like we are overfitting, because the decision boundary is quite simple. We might be overfitting to only a few of the training examples because SGD did not have the chance to properly sample and update the weights with respect to all training points yet. Hence, it is reasonable to train for more epochs.
In fact, running for more epochs decreases again the test error as well.

C. In general, changing the activation function may speed up convergence, hence it may be something to try. But in this case, given that tanh is used, and tanh is like sigmoid,

$$\tanh(x) = 2\sigma(2x) - 1$$

it may not be the best option to try. You receive the same points whether you chose this option or not.

D. We do not know about the model complexity; a less complex model may improve convergence. However, a more complex model may be able to better fit the data. You receive the same points whether you chose this option or not.

E. Correct. Using transformed features can help with non-linear data. (Selecting only E gives partial credit.)

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
Solution: A & E or only E
Explanation:
A. Both selecting this option and not selecting is correct, as we did not specify which RNN architectures we are comparing to. It is a true statement about the transformer. The bidirectional RNN considers relationships between all parts of the sequence as well (however with the memory bottleneck mentioned in option E).
B. Incorrect. It requires training as well (Where do the weight matrices discussed in the lecture come from?).
C. Incorrect. While we can use gradient descent for training, the problem is not a convex optimization problem. We have the softmax in the attention layer which is a non-convex function.
D. Incorrect. Not correct in general.
E. Correct. Memory is the main limiting factor of the RNNs.

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
Solution: D
Explanation: The encoding discussed in class is the one-hot encoding. We associate a specific token with a basis vector with a 1 in the position dedicated to it. We consider sequences of letters of length 10 with 4 options per letter, i.e., we have 4^10 = 2^20 different tokens. Hence, we need 2^20 unique encoding vectors, and the vector dimension is 2^20.

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

Solution: C.

Explanation: From the Context the sequence length is 10^6. After tokenization, we obtain sequence

length of 10^5 tokens, as each token is of length 10 and non-overlapping. The attention score matrix computes the score between all the tokens; the matrix is square and of dimension 10^5 times 10^5.

End of Paper