請勿攜去
Not to be taken away

香 港 中 文 大 學
The Chinese University of Hong Kong

Course Examination, 1st Term, 2022-23

科目編號及名稱
Course Code & Title : CSCI3230 Fundamentals of Artificial Intelligence

時間 Time allowed : 2 小時 hours : 0 分鐘 minutes

學號 Student I.D. No. : ____________ 座號 Seat No. : ____________

Answer **ALL** Questions. Full Score is 100%.

- This is a **close-book** examination. You can only bring one A4 page of notes as reference.
- You are **not allowed** to communicate with anyone directly or indirectly during the examination.

Part I: Answer the following multiple choice questions (30%)

1. (3%) Which of the following statements is true:

A. Ridge regression penalizes L1 norm of the model parameters to alleviate overfitting.
B. Shrinkage methods sacrifice some variance to reduce the bias of the model.
C. When the model complexity is very low, overfitting problem will become very severe.
D. If the input training data matrix X (with bias term absorbed) is invertible, the objective function (RSS) of linear regression on the training data can be minimized to zero.

2. (3%) Which of the following statements on dimensionality reduction or PCA is true:

A. To reduce the dimensionality of a data matrix $\boldsymbol{X}$ via PCA, we need to perform singular value decomposition (SVD) on $\boldsymbol{X}\boldsymbol{X}^T$.
B. PCA is applicable to any non-zero high-dimensional data matrix $\boldsymbol{X}$ for dimensionality reduction without regard to the rank of $\boldsymbol{X}$.
C. $U = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 0 & 1 \\ 1 & -1 & 1 \end{bmatrix}$ could be the optimal solution to some PCA problem.
D. Suppose U is the optimal solution obtained by running PCA on a data matrix $\boldsymbol{X}$. Then $UU^T\boldsymbol{X}$ is the low-dimensional representation of $\boldsymbol{X}$.

3. (3%) For classification methods, which of the following statement is true?

A. Once a linear SVM classifier is trained, you only rely on support vectors to make predictions for new samples.
B. Kernel SVM aims to deal with outlier samples in training data.
C. Hinge loss is a common loss function used for logistic regression.
D. Neural networks use backpropagation to train classification models on RSS loss.

4. (3%) Which of the following statement is correct about clustering methods?

A. For hierarchical clustering, we must set the number of output clusters beforehand.
B. Agglomerative clustering is a top-down approach.
C. If use DBSCAN for the set of six points $\{(4,4), (6,4), (2,2), (6,2), (3,1), (1,1)\}$, suppose ϵ is 1 and MinPts is 2, then there is no core point.
D. The advantages of DBSCAN include being able to handle clusters of various shapes and density.

5. (3%) Which of the following statement is true?

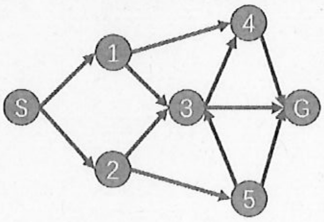
A. The output mean value for Sigmoid activation function is centred at 0.
B. Rectified Linear Unit (ReLU) function has slower convergence speed than Tanh function.
C. In CNNs, neurons in adjacent layers are only sparsely connected.
D. For neural network training, a good initialization of the model weights is more important than designing the model architecture itself.

6. (3%) In a classification problem, the false negative value is 5 and the true positive value is 20, what is the evaluation metric of recall?

A. 0.25
B. 0.6
C. 0.8
D. 1.0

7. (3%) Assuming for every edge, the cost is at least ε, with ε> 0. If we use searching algorithms to find the optimal path from node S to node G. Which of the following statement is not correct?

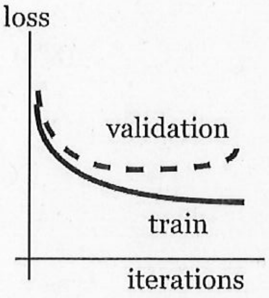
A: Search strategy determines the order in which nodes are expanded from the fringe.
B: Depth-first search cannot guarantee to return an optimal solution.
C: Breadth-first search can guarantee to return an optimal solution because costs are non-negative values.
D: For breadth-first graph search, node 3 must be visited.



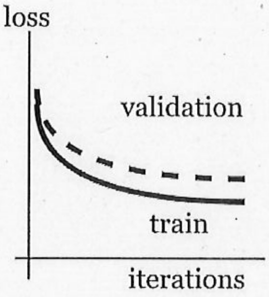
8. (3%) Which of the following statement about CNN is not true?

A. Data augmentation and dropout are common tricks to alleviate overfitting.
B. The number of kernels equals to the number of output activation maps.
C. Zero-padding of images would typically affect the prediction accuracy of CNN models for classification task.
D. The first convolutional neural network was proposed as early as 1990s.

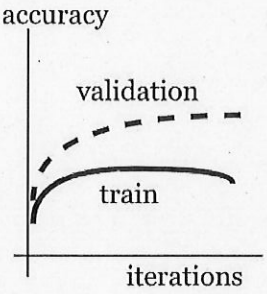
9. (3%) Which of the following learning curves demonstrates overfitting?



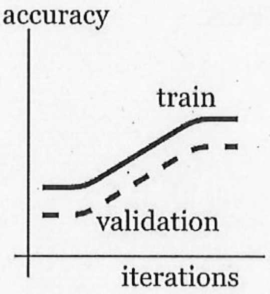
A



B



C



D

10. (3%) Consider the sigmoid function $(x) = \frac{1}{1+e^{-x}}$. The derivative $f'(x)$ is:

A. $f(x)\log f(x) + (1 - f(x))\log(1 - f(x))$
B. $f(x)(1 - f(x))$
C. $f(x)\log f(x)$
D. $f(x)(1 + f(x))$

Part II: Answer the short answer questions (38%)

1. In which era was the Dartmouth conference (considered as the founding event of AI) held? What is the major difference between supervised learning and unsupervised learning? (2%)

2. What function is often used as the last activation function in the prediction layer of a multi-class classification neural network? Write down its mathematical formulation and explain how it works. (4%)

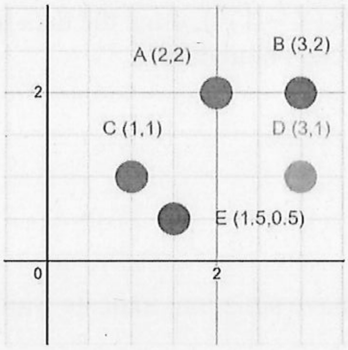
3. A single-state search problem consists of which components? List all the components and explain each of them. (4%)

4. What is the full name of PCA? Describe the statistical view of PCA. Write down the full algorithm of PCA. (8%)

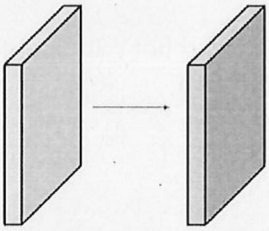
5. Why CNNs can obtain higher accuracy than traditional methods for image recognition? Would deeper models always give better performance? Why? How to solve it? (6%)

6. Answer the following questions about SVM.
a) What optimization method is used to transform SVM primal problem to a dual problem? (2%)
b) State two popular SVM kernels. (2%)
c) What is the use of penalty term (such as hinge loss) for soft-margin SVM? (2%)

7. Use K-means to cluster the five data points: A (2,2), B (3,2), C (1,1), D (3,1), E (1.5,0.5). If the initial centroids are (2,2) and (1,1). What are the two clusters after the first iteration? (2%)



8. For a convolutional layer, if we use 64 filters with kernel size 5 × 5, stride 3 and padding 2. Suppose the input volume size is 112 × 112 × 3, what is the output feature map size and the number of trainable parameters in this layer (no need to consider bias offset)? (2%)

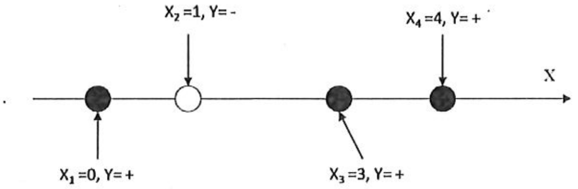


9. Explain why fully-connected layer is usually needed for a convolutional neural network. (2%)

10. State two application fields of AI and give a concrete example for each field. (2%)

Part III: Answer the following analysis questions (32%)

1. (8%) A robot has collected data with its sensor. We want to use the data to build a classifier using support vector machine. Currently, the feature space is a one dimensional space $X \in R$. The desired classification output is $Y = \{+, -\}$, as shown in the figure below. The training set contains three positive examples, $x_1 = 0, x_3 = 3, x_4 = 4$, and one negative example $x_2 = 1$.



a) Currently, the data points are not linearly separable. We want to define a transformation that maps the data into a projected 2D space in R^2. If we consider the feature mapping function as $\Phi(X) = (X, (X-1)^2)$, draw the data points after the transformation to the 2D space, and draw the line of decision boundary. (4%)

b) In the above situation, indicate which examples out of x_1, x_2, x_3, x_4 are support vectors? (2%)

c) If the robot got one more negative data point $x_5 = 1.5$, would it affect the margin? Please justify the reason. (2%)

2. (8%) This question is about clustering. Given five data points with the following matrix of Euclidean Distance, use agglomerative hierarchical clustering to cluster them.

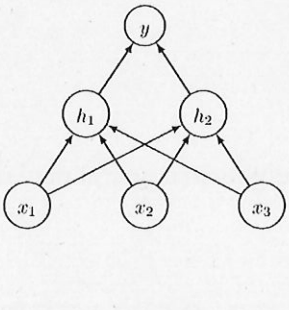
$$\begin{array}{c|ccccc}
 & 1 & 2 & 3 & 4 & 5 \\
\hline
1 & 0.0 & & & & \\
2 & 2.0 & 0.0 & & & \\
3 & 6.0 & 5.0 & 0.0 & & \\
4 & 10.0 & 9.0 & 4.0 & 0.0 & \\
5 & 9.0 & 8.0 & 5.0 & 3.0 & 0.0
\end{array}$$

a) If use Single Linkage, write down the clustering results in a dendrogram. Give the updated matrix of distance for each step. (4%)

b) State one advantage and one disadvantage of hierarchical clustering compared to K-means. (2%)

c) Besides Euclidean distance (L2-norm), there are other similarity measurements such as Manhattan Distance (L1-norm) and Minkowski Distance (p-norm). Write down the mathematical formula of p-norm, and explain why it is a general description of L1-norm and L2-norm. (2%)

3. (16%) This question is about forward and backward propagation of a basic neural network. Specifically, the following graph shows the architecture of a neural network with a single hidden layer. The input layer has three neurons $x = (x_1, x_2, x_3)$. The hidden layer has two neurons $h = (h_1, h_2)$. The output layer has one unit y.



a) We would like to use ReLU activation function for the hidden layer and the output layer. Please write down the mathematical formula $\sigma(z)$ for ReLU function. (z just generally denotes a variable). State two advantages of ReLU activation function compared to logistic sigmoid activation. (3%)

b) We use ReLU as the activation function in this network. Moreover, denote by W as the weight matrix connecting input and hidden layer, and V as the weight matrix connecting hidden layer and output unit. Write out the symbolic function of the mapping $x \rightarrow y$ using σ, W, V. (1%)

c) We further design the loss function as $\ell(\mathrm{y}, \mathrm{t}) = \frac{1}{2}(\mathrm{y} - \mathrm{t})^2$ where t is the ground-truth value for the output unit y. Assume the network parameters are initialized as follows.

$$\mathrm{W} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix} \text{ and } \mathrm{V} = \begin{bmatrix} 0 & 1 \end{bmatrix}$$

Assume that we have one training sample (x, t) with $x = [1,2,1]$ and $t = 1$. Compute the numerical value of hidden neurons h and output unit y. (2%)

d) Compute the gradient of the loss function with respect to the weights. In particular, compute the following terms symbolically: (8%)

- The gradient relative to V, i.e., $\frac{\partial \ell}{\partial V}$
- The gradient relative to W, i.e., $\frac{\partial \ell}{\partial W}$
- Compute the values numerically for the values of W, V, x, y as given above.

e) Does it make sense to initialize all weights in a neural network to 0? Why? (2%)

<END OF PAPER>