CITY UNIVERSITY OF HONG KONG

Course code & title	:	CS5489 Machine Learning: Algorithms & Applications
Session	:	Final, Semester B 2021
Time allowed	:	Two hours (May 14, 6:30pm-8:30pm)

This question paper has 5 pages (including this cover page).

1. This paper consists of 13 questions.
2. Answer ALL questions.
3. Write your answers in the accompanying **“CS5489-final-2021-answersheet.docx”**.

This is an ***open-book*** *final exam, see the allowed resources below.*

Instructions:

- Answer all questions in the accompanying Word document **“CS5489-final-2021-answersheet.docx”**.
- The following resources are **allowed** during the final:
  - Videos of CS5489 lectures on Zoom,
  - any **unaltered** material on the CS5489 Canvas page, including lecture notes, tutorials, etc.
  - **unaltered** course textbooks
- Any other resources are **not allowed**, for example
  - internet searches.
  - Classmates.
  - other textbooks.
  - any text/notes copied into your lecture notes or textbook.
  - translation software.
- You should stay on Zoom during the entire exam time in case there are any announcements.
  - If you have any questions, please use the private chat function of Zoom to message Antoni.
- By 8:30pm May 14, submit the completed final to the “Final” Assignment on Canvas.
  - If you have trouble accessing Canvas, then you can send the completed docx via email to Antoni (abchan@cityu.edu.hk).

CS Departmental Hotline (phone, whatsapp, wechat): +852 6375 3293

Multiple Choice/Selection Questions (30 points)
5 marks each. *For a multiple selection question, an incorrect answer will be penalized 5/K marks, where K is the number of correct answers. If more incorrect answers are given than correct answers, the marks will be 0.*

Q1. Which statements about support vector machines (SVM) are correct? (select all that apply)
A) Maximizing the margin in SVM can allow most "wiggle" room for the hyperplane w, while keeping points correctly classified.
B) Adding the "slack" variable can help SVM handle non-separable data.
C) In kernel SVM, all the training data is needed to make a prediction on new data.
D) In kernel SVM, the possible shapes of the decision boundary are determined by the kernel.
E) In kernel SVM, using RBF kernel will obtain better classification performance than polynomial kernel.

Q2. Which statements are correct about linear dimensionality reduction methods? (select all that apply)
A) Applying linear dimensionality reduction will always preserve the class separation.
B) Linear dimensionality reduction methods approximate the data as a weighted sums of basis vectors.
C) The coefficients are computed in closed form using a linear transformation.
D) LSA and PCA both minimize the squared reconstruction error.
E) Reducing the dimension can extract semantically meaningful features.

Q3. Which statements are true about K-means clustering? (select all that apply)
A) The K-means algorithm is sensitive to the initialization.
B) K-means clustering is the same as GMM clustering when the covariance matrix is the identity matrix.
C) In some cases, the K-means algorithm will not converge.
D) Assigning points to cluster centers will partition the space with linear surfaces.
E) The number of clusters can be determined by minimizing the sum-squared difference between points and their centers.

Q4. Which of the following conditions will make a Multi-layer perceptrons (MLPs) equivalent to multi-class logistic regression? (select all that are required)
A) There are no hidden layers.
B) Threshold activation function and cross-entropy loss are used.
C) Soft-max activation function and cross-entropy loss are used.
D) Linear activation function and mean-square error loss are used.
E) There is 1 hidden layer, which is sufficient to model any continuous function.

Q5. Suppose we apply L2 regularization on a deep neural network. Which statements are correct? (select all that apply)

A) L2 regularization is more effective when applied only to the later layers.
B) L2 regularization is more effective when applied only to the early layers.
C) L2 regularization is more effective when applied to all layers.
D) L2 regularization directly prevents the activations from becoming too large.
E) L2 regularization is equivalent to assuming a Gaussian prior on the weights and using MAP estimation.

Q6. Which statements are true about variational auto-encoders (VAEs)? (select all that apply)

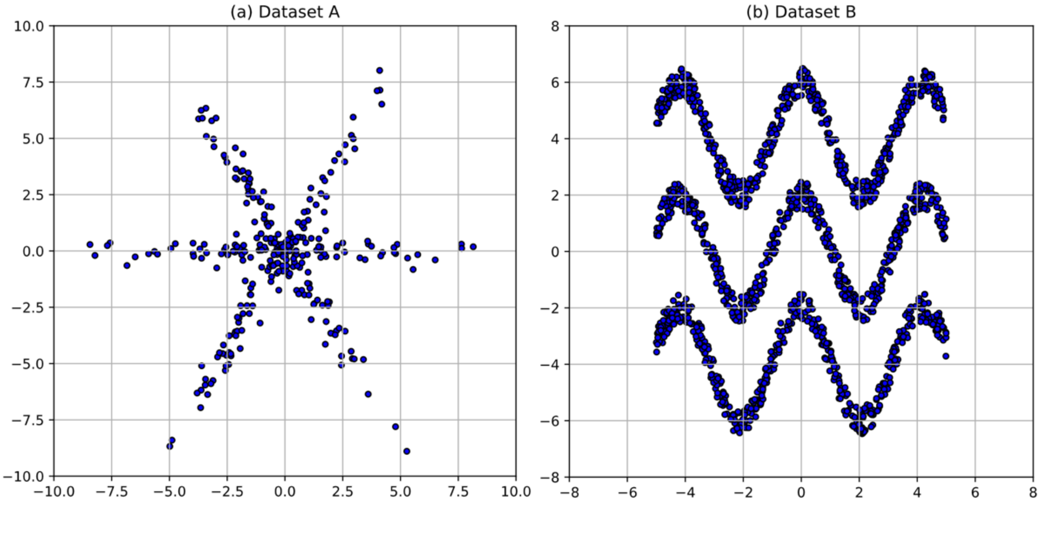
A) Noise is added to the latent vector to make the decoder more robust.
B) The decoder does not need to share weights with the encoder.
C) Given an input, the VAE estimates a posterior distribution of the latent vector.
D) The encoder and decoder are decoupled for easier training.
E) Training the VAE requires both data samples and class labels.

Discussion Questions (70 marks)

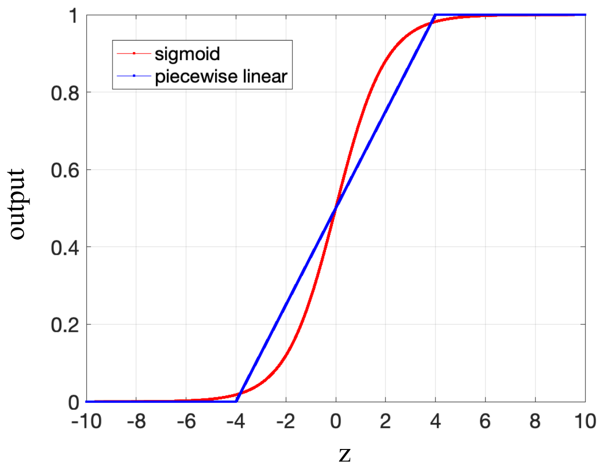
10 marks each question.

Q7. Suppose we want to apply PCA to a set of N=1,000 images, each with resolution of 1000x1000 pixels. In this case each vectorized image is a vector of length D=1,000,000. What kind of numerical computation problems might we encounter when applying PCA? How to deal with these problems? Explain why your solution will help.

Q8. Consider the following two datasets. Which clustering algorithm(s) would you use on each dataset? Explain why.



Q9. Typically the sigmoid activation function is used for binary classification with neural networks. Suppose we replace the sigmoid activation with a piece-wise linear function $\phi(z)$ as shown below:

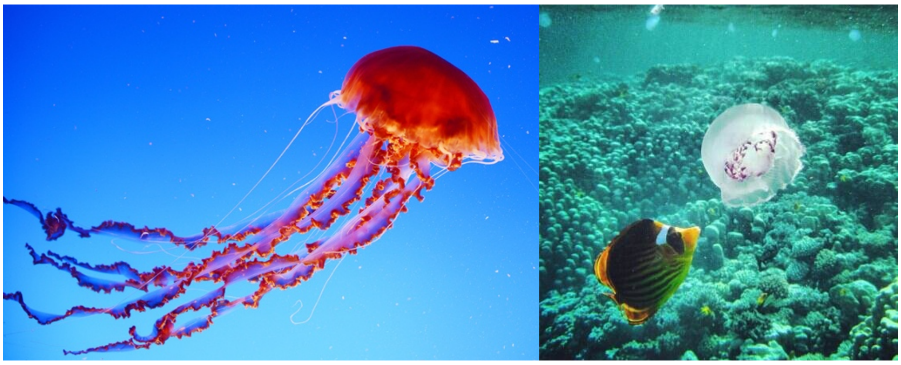


Let $f(\mathbf{x})$ be the output of the last hidden layer (penultimate layer) of the neural network, and thus $\phi(f(\mathbf{x}))$ is the output of the classifier. The loss function using binary cross-entropy is then:

$$L(y, f(\mathbf{x})) = -y \log \phi(f(\mathbf{x})) - (1-y)\log(1-\phi(f(\mathbf{x})))$$

What are the advantages and disadvantages of using this activation function and loss?

Q10. You are working with a marine biologist on the task of classifying underwater images as containing jellyfish or not (a binary classification task).

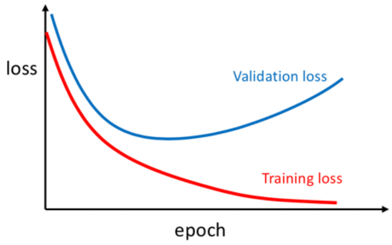


Consider the following CNN for this binary classification task on 128x128 color images.

Layer Name	**Description**
Input Layer	128 x 128 RGB input image
Conv Layer 1	7x7 kernel, 16 output channels, stride 1, “valid” mode
Conv Layer 2	9x9 kernel, 8 output channels, stride 1, “valid” mode
Conv Layer 3	3x3 kernel, 4 output channels, stride 1, “same” mode
Flatten Layer	
Dense Layer	128 hidden nodes, ReLU activation
Dense Layer	2 output nodes, soft-max activation

Calculate the size of the feature maps produced by each of the three convolution layers, and the size of the receptive fields for each of the three convolution layers. Show your calculations.

Q11. After training the CNN in Q10 on the dataset, you obtain the following training and validation loss curves.



What problem is exhibited by the curve? How would you change the CNN **architecture** in Q10 to fix the problem? You may add, remove, or change layers. Explain why these changes would help.

Q12. What kind of data augmentation would you use for the image classification problem in Q10? Explain why. What is the purpose of data augmentation?

Q13. GANs use the reparametrization trick to sample from a complex probability distribution by learning a transformation, $\mathbf{y} = f(\mathbf{x})$, $\mathbf{x} \sim N(0,\mathbf{I})$, where f is the transformation function modelled by a neural network. Explain why we generally cannot calculate the likelihood function of the complex probability distribution. Write down conditions that make it possible to calculate the likelihood function.

--- END ---