CITY UNIVERSITY OF HONG KONG

Course code & title	:	CS5489 Machine Learning: Algorithms & Applications
Session	:	Semester B 2020
Time allowed	:	Two hours (May 18, 6:30pm-8:30pm)

This question paper has 4 pages (including this cover page).

---

1. This paper consists of 13 questions.

2. Answer ALL questions in the accompanying **"CS5489-final-2020-answersheet.docx"**.

---

*This is a **open-book** examination.*

Instructions:

- Answer all questions in the accompanying Word document **"CS5489-final-2020-answersheet.docx"**.
- The following resources are **allowed** during the final:
  - Panopto lecture videos (or the same videos on OneDrive),
  - any material on the CS5489 Canvas page, including lecture notes, tutorials, etc.
  - course textbooks
  - videos on Zoom taken during the lecture discussions.
- Any other resources are **not allowed**, for example
  - internet searches
  - classmates
  - other textbooks
- 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 18, submit the completed quiz to the "Final Exam" 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.*

Q1. Which statements about Principal Component Analysis (PCA) are correct? (select all that apply)
A) Classification accuracy using the PCA coefficient space will always be the same or better than that in the original input space.
B) The PCA objective of minimizing reconstruction error is the same as maximizing the preserved variance.
C) PCA assumes that the data density is a multivariate Gaussian.
D) PCA is slow because all the eigenvectors of the covariance matrix need to be computed.
E) When increasing the number of principal components from p to (p+1), only one principal component will be different and the rest are the same as before.

Q2. Which statements are true about GMM clustering? (select all that apply)
A) If we set the covariance matrices to be identity matrices and the component weights to be uniform, then it is the same as the k-means algorithm.
B) The E and M steps correspond to computing the cluster membership and computing the cluster shape.
C) Assuming the covariance matrix is a diagonal matrix makes the clusters into circle shapes.
D) Running EM will always give the global optimal solution.
E) The number of clusters can be automatically selected using Bayesian methods.

Q3. Which statements about Multi-Layer Perceptrons (MLPs) are correct? (select all that apply)
A) The gradient cannot be computed exactly, so approximation methods are required.
B) The choice of activation function does not affect the gradient.
C) Each layer extracts features from all other layers.
D) An MLP with no hidden layers, with output nodes using the soft-max activation function, and trained with cross-entropy loss is equivalent to multi-class logistic regression.
E) An MLP with a single hidden layer and non-linear activation is sufficient to approximate any continuous function.

Q4. Which statements about convolutional layers are correct? (select all that apply)
A) the receptive field size increases as more convolutional layers are used.
B) they introduce translation invariance to the classifier.
C) the convolution operation is equivalent to multiplication in the frequency domain.
D) the number of parameters is equal to the input size times the output size.
E) they are equivalent to a fully-connected layer with a weight matrix that has specific form and uses shared values.

Q5. Which statements about Stochastic Gradient Descent (SGD) optimization are correct? (select all that apply)
A) One advantage of SGD is that it can jump away from local minima.
B) When the computed mini-batch gradient is 0, a local minimum of the cost function is reached and training is finished.
C) The variance of the update step depends on the learning rate.
D) Using momentum is one way to smoothen the gradient.
E) Reducing the learning rate during the epochs will prevent SGD from converging.

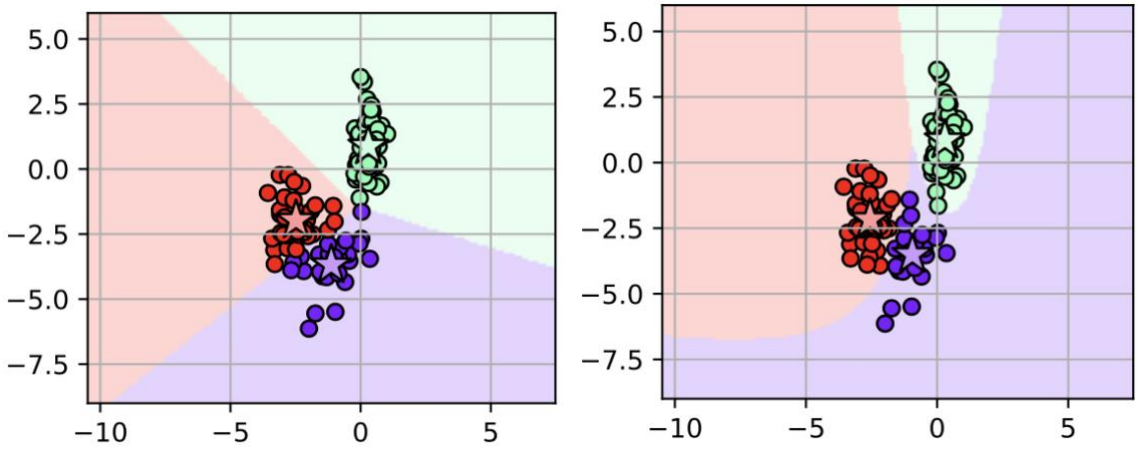
Q6. Which statements are true about GAN? (select all that apply)
A) GAN uses the reparameterization trick to generate samples from the target distribution.
B) One advantage of GAN is that it can compute the likelihood value of a sample easily.
C) One disadvantage of GAN is that it needs to compute the partition function.
D) In practice, training the generator to minimize $-\mathbb{E}_z[\log D(G(z))]$ is more stable than minimizing $\mathbb{E}_z[\log(1 - D(G(z)))]$.
E) Mode collapse problem means that the generator fails to fool the discriminator.

Discussion Questions (70 marks)
10 marks each question.

Q7. You are training a binary classifier on a dataset with 2,000 samples. You will test two approaches: A) run kernel PCA on the training set first to reduce the dimension, then train a linear SVM on the KPCA coefficients; 2) train a kernel SVM directly on the data. Will you get the same result? Why or why not? If not, which approach will yield a better classifier?

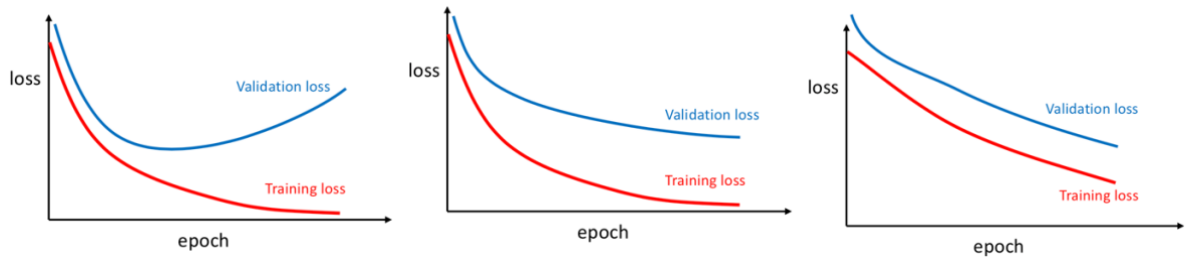
Q8. Consider the two clustering results below. Which result is from K-means clustering, and which is from GMM clustering? Explain your reasoning, and why this happens.

(a) (b)



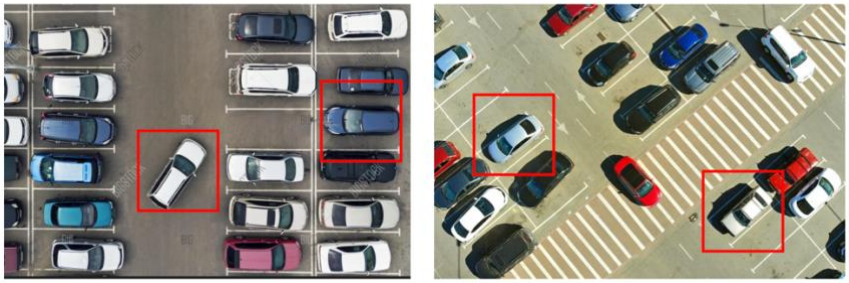
Q9. Consider the following training/validation loss curves obtained by training a deep neural network. For each curve, what is the problem exhibited by the curve? Give the reason why you think this. How would you fix the problem? Describe the solution in detail.

(a) (b) (c)



Q10. Explain the advantages and disadvantages of a "shallow" MLP with 1 hidden layer versus a "deep" MLP with 10 hidden layers.

Q11. You are training a CNN to classify patches in overhead aerial images as either "*car*" or "*not car*", as in the below image. What kind of data augmentation would you use during training and why? What is the purpose of data augmentation?



Q12. What are the similarities and differences between an autoencoder (AE) and a variational autoencoder (VAE)? Why are these differences important for the VAE?

Q13. Consider the "Leaky" ReLU activation function: $\sigma(x) = \begin{cases} x, & if\ x \geq 0 \\ \frac{x}{a}, & if\ x < 0 \end{cases}$, where typically a=1000. What are the advantages and disadvantages of this activation function?

--- END ---