香 港 大 學

THE UNIVERSITY OF HONG KONG

Bachelor of Engineering

Department of Electrical & Electronic Engineering

ELEC 3249 Pattern Recognition and Machine Intelligence

2021-2022 Semester 2
Online Examination

Date: ___May 13, 2022_____ Time: ___9:30am-12:30pm_____

For all the questions that require you to give an explanation, if you don't provide an explanation, you can only get half of the full marks even though your answer is correct.

Candidates are permitted to refer to the following electronic/printed/handwritten materials in the examination: textbook, lecture slides, assignment handout, and self-made notes. Internet searching and crowdsourcing from group messages, online forums or social media, etc. are strictly forbidden.

Questions are not ordered in terms of difficulties.

Use of Electronic Calculators:

"Only approved calculators as announced by the Examinations Secretary can be used in this examination. It is candidates' responsibility to ensure that their calculator operates satisfactorily, and candidates must record the name and type of the calculator used on the front page of the examination script."

Section A: Basic Concepts and Generative Methods (24 marks)

A1. Which of the following statement is **incorrect**? And explain why.

a) Feature extraction in pattern classification generally refers to the process of mapping a pattern into a feature vector

b) More features don't always yield good model performance

c) For a model, a good classification result on the training data always implies a good generalization ability of that model

d) Supervised learning is to train a model guided by labeled data, i.e., each sample is annotated with a corresponding category label

e) None of the above

(4 marks)

A2: Which of the following statement of generative/discriminative models is **incorrect**? And explain why

a) SVM is a discriminative model

b) Discriminative models will directly estimate the decision boundary

c) Bayesian classifier is a generative model

d) Maximum a posterior is used to directly estimate the weights that defines a linear discriminant function

e) None of the above

(4 marks)

A3. Suppose we collected a dataset for a two-category classification problem: class ω_1 $D_1 = \{x_1^1, x_2^1, \ldots, x_m^1\}$ (a total of m samples for class ω_1); and class ω_2 $D_2 = \{x_1^2, x_2^2, \ldots, x_n^2\}$ (a total of n samples for class ω_2), you are asked to obtain a generative model using maximum likelihood estimation to separate the two classes.

(1) Step 1: Calculate the prior probability for each class $P(\omega_1)$ and $P(\omega_2)$.

(4 marks)

(2) Step 2: Use maximum likelihood estimation method to obtain the likelihood model (that is the class conditional probability density) for each class $p(x|\omega_1)$ and $p(x|\omega_2)$. Here we assume $p(x|\omega_1) \sim p(x|\theta_1)$ and $p(x|\omega_1) \sim p(x|\theta_2)$. That is the parameters θ_1 and θ_2 fully determine the likelihood model of ω_1 and ω_2. Please write down the key steps to obtain $\hat{\theta}_1$using dataset $D_1$and $\hat{\theta}_2$ using dataset D_2 with maximum likelihood estimation.

(6 marks)

(3) Step 3: given the prior probability obtained in problem (1) and likelihood model obtained in problem (2), please design discriminate functions $g_1(x)$ and $g_2(x)$ to classify the data into two categories using what you have learned in the course and describe how would you classify a new sample x_{new}.

(6 marks)

Section B: Linear Discriminant Functions and Support Vector Machine (29 marks)

B1. Which of the following about how to obtain a linear classifier is **true**?

a) You need to calculate the sample mean for each class

b) You need to calculate the sample variance for each class

c) You need to use iterative methods, such as gradient descent, to minimize an error function over the training data

d) For a C category classification problem, you would get more than C decision regions using a linear machine.

e) None of the above

(4 marks)

B2. Which of the following statement about SVM is **incorrect**? And explain why

a) SVM aims to find a separating hyperplane that maximizes the margin of separation

b) If some data samples are removed from the training set, the result obtained by SVM will always change

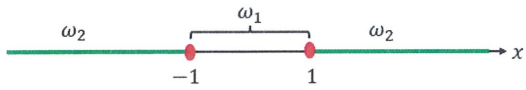
c) Kernel method is used to map a low dimensional feature into a high dimensional one

d) The margin measures the distance of the nearest sample from the decision boundary

e) None of the above

(4 marks)

B3. For a one-dimensional feature space x, the decision region (Class $\omega_1: -1 < x < 1$, otherwise ω_2) is given as below which is not linearly separable. However, we want to design a linear classifier to separate the data. Which of the following approaches can help us develop a linear classifier with a 100% accuracy (that's to make the data linearly separable)? And explain why.



a) Add a feature x^2

b) Add a feature 1

c) Add a feature x^3

d) Add a feature $1 - x$

e) None of the above

(4 marks)

B4. Given a training dataset for two categories, to obtain a linear separating hyperplane, we can use gradient descent to minimize the empirical risk or train a support vector machine. Please describe the differences of the linear separating hyperplane found by the two approaches.

(4 marks)

B5. Let's consider a two-category classification (ω_1 and ω_2) problem with a linear discriminant function $\mathbf{g(x)} = \mathbf{w^T x} + w_0$ where the weight vector $\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ and a

two-dimensional feature vector $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. That is $\mathrm{g}(\mathbf{x}) = w_1 x_1 + w_2 x_2 + w_0$. A sample will be assigned to category ω_1 if $\mathrm{g}(\mathbf{x}) > 0$ otherwise assigned to ω_2.

We collect a training dataset $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ with corresponding labels $\{z_1, z_2, \ldots, z_n\}$. Here, $\mathbf{x}_\mathrm{i} = \begin{bmatrix} x_1^i \\ x_2^i \end{bmatrix}$ represents the two-dimensional feature vector for the i-th sample. Besides, $z_i = 1$ if $\mathbf{x}_i$ belongs to category ω_1 and $z_i = -1$ if $\mathbf{x}_i$ belongs to category ω_2.

To obtain $\{w_1, w_2, w_0\}$, we design the following error function.

$$J(w_1, w_2, w_0) = \sum_{i=1}^{n} \ln\left(1 + \exp(-z_i g(\mathbf{x}_i))\right)$$

where $g(\mathbf{x}_i) = w_1 x_1^i + w_1 x_2^i + w_0$, "ln" denotes the natural logarithm and "exp" denotes the exponential function.

(1) Suppose we use gradient descent to minimize the above function and to obtain w_1, w_2, w_0, please write down **the key steps for gradient descent**. Note you need to calculate $\frac{\partial J}{\partial w_1}, \frac{\partial J}{\partial w_2}, \frac{\partial J}{\partial w_0}$.

(8 marks)

(2) Please explain why minimizing the above error function will help obtain the decision boundary that can separate the training data, that is to make the training data be correctly classified.

(5 marks)

Section C: Unsupervised Learning (17 marks)

C1. Which of the following statement about K-means and gaussian mixture model is **incorrect**? And explain why

a) Gaussian mixture model assigns a sample to each cluster with a probability

b) Gaussian mixture model can be viewed as a soft-version K-means algorithm

c) For a gaussian mixture model, a cluster can be fully represented by its centroid

d) The result of both K-means and Gaussian mixture model will be influenced by the initialization status

e) None of the above

(4 marks)

C2. Consider the following dataset $D = \{0,1,5,10,14\}$ with 5 samples. Please derive the result of hierarchical clustering (agglomerative approach) using Euclidian distance. For measuring the distance between two clusters, single linkage is adopted.

(5 marks)

C3. Given the result obtained in C2, if we would like to have two clusters from the hierarchical results, what should be done? And please write down the result.

(3 marks)

C4. How about we utilize K-means algorithm to cluster the above data $D = \{0,1,5,10,14\}$ into 2 clusters. The two cluster centroids are initialized as $c_1 = 8$ and $c_2 = 10$. Please derive the process for k-means clustering and write down the results.

(5 marks)

Section D: PCA and LDA (13 marks)

D1. Which of the following statement about PCA and LDA is incorrect? And explain why

a) PCA is unsupervised while LDA is supervised

b) LDA aims to find a projection axis that can make the projected data be better discriminated

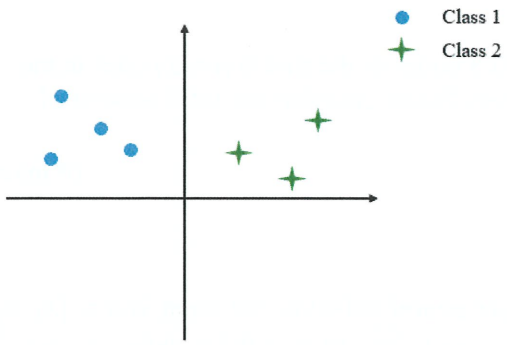
c) PCA can minimize the information loss after dimensionality reduction

d) LDA uses the eigenvectors of the data covariance matrix as its projection axis.

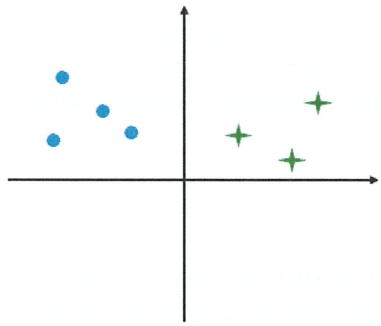
e) None of the above

(4 marks)

D2. Consider the following dataset, if we would project the data onto a one-dimensional space using PCA and LDA. Please draw the axis created by each approach and explain why.



(a) Draw PCA



(b) Draw LDA

(5 marks)

D3. If Bob wants to have a classification model to automatically classify dog w_1 and cat w_2, he can only collect a small number of images for cat and dog. Each image has a size of 1080×2040. Please use what you have learned in this course except neural networks to help Bob. You need to write down the key steps.

(4 marks)

Section E: Neural Networks and Deep Learning (17 marks)

E1. Which of the following statement about neural networks is **incorrect**? And explain why.

a) Data augmentation is a good strategy to prevent model overfitting

b) Sigmoid activation function in the hidden layer often leads to gradient diminishing and causes difficulties for neural network training

c) Initialize all the weights to zero is not a good idea for training deep learning models

d) To obtain the weights of deep neural networks (more than four layers) with nonlinear activation functions, we need to calculate the gradient using backward propagation and set the gradient to zero to obtain the weight parameters.

e) One major advantage of deep learning is that the feature representations are learned from data

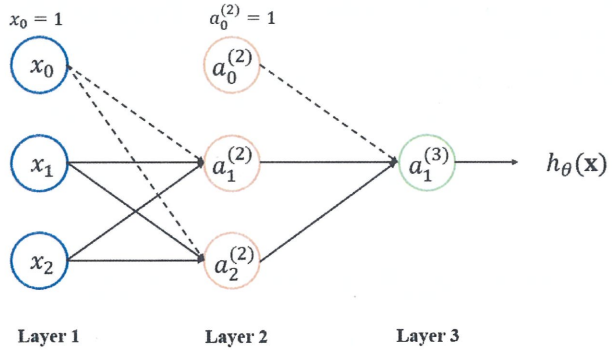
(4 marks)

E2. For a three-layer neural network with 5 nodes in the first layer, 6 nodes in the second layer, and 3 nodes in the third layer. Please calculate the total number of weights considering the bias term.

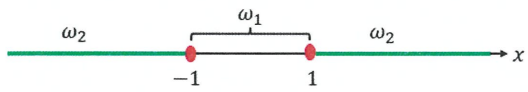
(4 marks)

E3. Considering the following three layer neural network, the input is $\mathbf{x} = [x_1, x_2]^T$. The associated weights are 0 for "node x_1 and $a_1^{(2)}$" (that is $\theta_{11}^{(1)}$), "bias x_0 and $a_2^{(2)}$" (that is $\theta_{20}^{(1)}$), and "node x_2 and $a_2^{(2)}$" (that is $\theta_{22}^{(1)}$). All the other weights are 1. The activation function of the 2nd layer $g_2(z)$ is a ReLU function $g_2(z) = \begin{cases} z \; if & z > 0 \\ 0 \; if & z \leq 0 \end{cases}$ and the activation function of the 3rd layer $g_3(z)$ is a sigmoid function $g_3(z) = \frac{1}{1+\exp(-z)}$. Please calculate the output value $h_\theta(\mathbf{x})$ given $\mathbf{x} = [-0.5, 0.5]^T$.

(4 marks)



E4. Please design a multi-layer neural network with **one output unit** for a two-category classification problem. The input contains only a one-dimensional feature x. The sample is assigned to category ω_1 if $-1 < x < 1$. Otherwise, the sample is assigned to category ω_2. Please design a neural network with its weight parameters and activation functions. And explain how the network output can be used to categorize the data into two categories. (Note that a three-layer neural network is already enough for this problem)



(5 marks)

*** END OF PAPER ***