香 港 大 學

THE UNIVERSITY OF HONG KONG

Bachelor of Engineering

Department of Electrical & Electronic Engineering

ELEC 3249 Pattern Recognition and Machine Intelligence

2022-2023 Semester 2
Examination

Date: ___15 May 2023_____ Time: ___9:30am-12:30pm_____

This is an open book examination. Candidates may bring to their examination any printed/written materials.

You should include your explanation for multiple-choice questions if required. Otherwise, you can only get half of the total marks even if you are right. In fill-in-the-blank questions, points can be awarded if the provided answer conveys the correct meaning.

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: Pattern Classification and Generative Methods (33 marks)

A1. Fill in the following blanks.

(a) In pattern recognition, we have a process to map the sensed pattern into a vector of number to represent the pattern. This process is called: _____.

(b) The ability of a classifier to produce correct prediction results on novel test patterns is called _____. Please list two ways to improve the above ability of a classifier_____ and _____.

(c) Does a complicated model always improve the ability of a classifier in producing correct prediction results on novel test samples? (True or False) Please also explain your choice.

(8 marks)

A2. Which of the following statements is **correct**? And explain why the other answers are incorrect. (Select one answer)

a) Logistic regression belongs to generative models while SVM belongs to discriminative models.

b) Suppose the input feature vector is $\mathbf{x}$ and label is y, discriminative methods will model the joint distribution of $p(\mathbf{x}, y)$

c) Supervised learning is to train a model guided by annotated labeled data. That is each sample is annotated with a corresponding category label.

d) More features will always improve classification performance.

e) Discriminative models will use maximum likelihood estimation to estimate the distribution of the training data $p(\mathbf{x}|y)$.

(3 marks)

A3. A probability density function with a positive parameter θ is given by:

$$p(x|\theta) = \frac{\theta^x e^{-\theta}}{x!}, \qquad x = 0,1,2, \ldots.$$

Note ! is known as the factorial operator where $x! = 1 \times 2 \times 3 \ldots \ldots \times (x-1) \times x$

Please answer the following questions:

(a) Given data samples $D = \{x_1, x_2, \ldots, x_n\}$ with each sample independently drawn according to probability density $p(x|\theta)$, please derive the maximum likelihood estimate of θ and list the key steps.

(8 marks)

(b) Suppose the prior of θ follows a gaussian distribution,

$$p(\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}}$$

Please calculate the posterior distribution $p(\theta|D)$ over θ. You can use a constant α to represent $P(D)$ if needed.

(3 marks)

(c) Given (b), please estimate the parameter θ based on maximum a posteriori.

(3 marks)

(d) Please compare maximum likelihood estimation and maximum a posteriori estimation.

(4 marks)

(e) Given the estimated parameter $\hat{\theta}$ for two classes: $p(x|\hat{\theta}_1)$ for class ω_1 and $p(x|\hat{\theta}_2)$ for class ω_2 and the prior probability for class ω_1 and ω_2 as $P(\omega_1)$ and $P(\omega_2)$, please answer how we would classify a new sample x_{new} using bayesian classifier.

(4 marks)

Section B: Logistic Regression and Support Vector Machine (38 marks)

B1. Please fill in the following blanks.

(a) Logistic regression uses a logistic function to represent_____. The decision boundary found by a logistic regression model is _____.

(b) Please list the three regularization techniques that can be used in logistic regression_____, _____ and _____.

(c) The goal of SVM is to find a separating hyperplane that can _____.

(d) For gradient descent, the parameter controls the search step is also called _____. Please discuss its effects on gradient algorithm if it is too large or too small (You can draw figure to illustrate your solution if needed).

(12 marks)

B2. Which of the following about logistic regression and SVM is **incorrect** and explain your choice? (Select one answer)

a) Regularization is used in logistic regression to avoid model overfitting.

b) Kernel methods are used in SVM to handle data that are not linearly separable and can reduce the computation complexity.

c) L_1 (lasso) regularization can enforce the learned parameter (weights) to be small and L_2 (ridge) regularization can enforce them to be sparse.

d) The training objective of logistic regression is to find a set of parameters θ that can maximize the probability of correctly classifying the training samples.

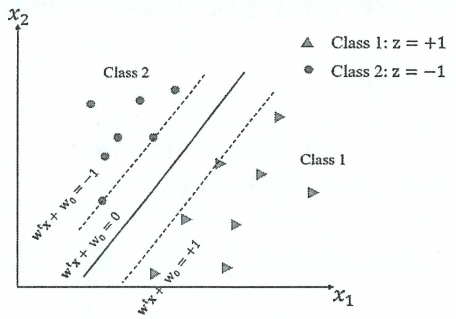
e) Iterative optimization methods such as gradient descent are used to train a logistic regression model.

(3 marks)

B3. Let the discriminant function for a support vector machine be (refer to Figure below):

$$g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0 \begin{cases} \geq +1 & \text{for a sample } \mathbf{x} \text{ belonging to category 1: } z = +1 \\ \leq -1 & \text{for a sample } \mathbf{x} \text{ belonging to category 2: } z = -1 \end{cases}$$

Please answer the following short questions (refer to the Figure).



(a) What are support vectors? How many support vectors are shown in the figure? What is the margin of separation? Please indicate the margin in the figure.

(5 marks)

(b) Please verify that the margin in the above case is $\frac{1}{||\mathbf{W}||}$.

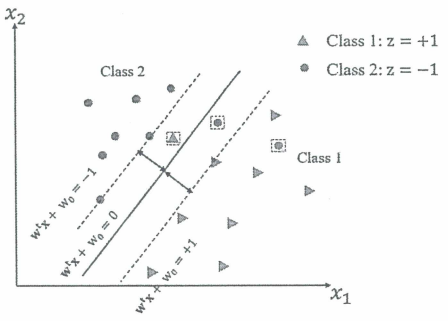
(3 marks)

(c) If we remove some non-support vectors from the training data and re-train the SVM classifier, does the solution change? And explain why.

(3 marks)

(d) In practice, as shown below, the training data often contain some outliers (samples with rectangular box below) making the data not-linearly separable, please show the solution that helps the model avoid being influenced by such outlier samples and explain how it addresses this problem based on your understanding.

(3 marks)



B4. Let's consider a two-category classification (ω_1 and ω_2) problem with a linear discriminant function $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$ 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 $g(\mathbf{x}) = w_1x_1 + w_2x_2 + w_0$. A sample will be assigned to category ω_1 if $g(\mathbf{x}) > 0$ 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, $\mathbf{x}_i = \begin{bmatrix} x_1^i \\ x_2^i \end{bmatrix}$ 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 and minimize the function.

$$J(w_1, w_2, w_0) = \sum_{i=1}^{n} \exp\big(-z_i g(\mathbf{x}_i)\big) = \sum_{i=1}^{n} \exp(-z_i(w_1x_1^i + w_2x_2^i + w_0))$$

Note that "exp" denotes the exponential function which is equivalent to $e^{-zg(\mathbf{x}_i)}$

(a) 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 $\frac{\partial J}{\partial w_1}, \frac{\partial J}{\partial w_2}, \frac{\partial J}{\partial w_0}$.

(6 marks)

(b) Please explain why minimizing the above error function will help obtain the decision boundary that works well on classifying the training data.

(3 marks)

Section C: Unsupervised Learning (12 marks)

C1. Please fill in the following blanks.

(a) The output of _____ is a tree-like structure that represents the hierarchical organization of the clusters. (Fill in an algorithm you learned in this course).
(b) For k-means algorithm, _____ are used to represent clusters.
(c) For gaussian mixture model, _____are used to represent clusters.

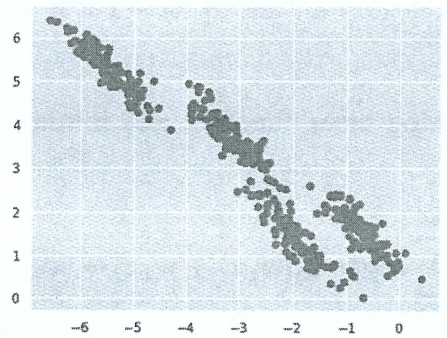
(4 marks)

C2. K-means algorithm and EM algorithm iterates between assigning data to clusters (E-step in GMM) and updating cluster parameters (M-step in GMM). Please list their differences in the above two steps.

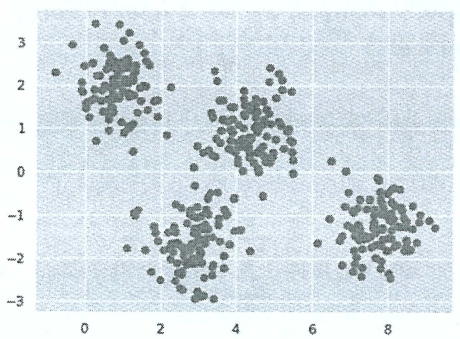
(4 marks)

C2. For the two datasets shown in the following figure, we would apply K-means or gaussian mixture model to group the data into four clusters. Please choose the best algorithm for the two datasets considering both complexities and accuracy. And explain your solution.

(4 marks)



(a)



(b)

Section D: PCA (11 marks)

D1. Which of the following statements about PCA is incorrect? And explain why

a) PCA is an unsupervised dimensionality reduction method.

b) PCA 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) PCA uses the eigenvectors of the data covariance matrix as its projection axis.

e) None of the above

(3 marks)

D2. Given the following dataset: $\begin{bmatrix}4\\1\end{bmatrix}, \begin{bmatrix}2\\3\end{bmatrix}, \begin{bmatrix}5\\4\end{bmatrix}, \begin{bmatrix}1\\0\end{bmatrix}$, we want to use principal component analysis (PCA) to reduce the dimension to 1 with minimized reconstruction error. Please show the new data with one-dimension of features.

(Note you need first apply PCA and reduce the data dimension by projection. Before applying projection, we should subtract the mean from the original data)

(8 marks)

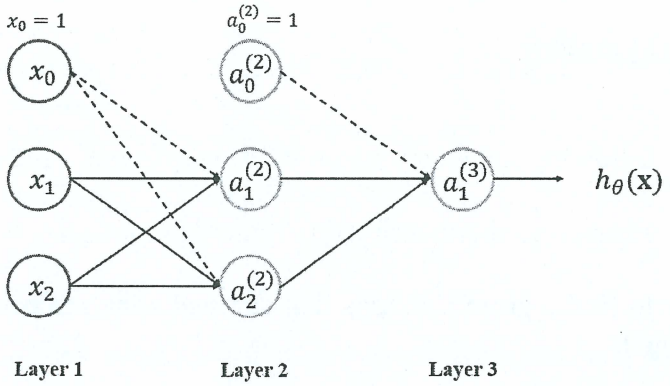
Section E: Neural Networks and Deep Learning (6 marks)

E2. For a three-layer neural network with 6 nodes in the first layer, 7 nodes in the second layer, and 4 nodes in the third layer. Please show the total number of weights considering the bias term.

(3 marks)

E3. Considering the following three-layer neural network, the input is $\mathbf{x} = [x_1, x_2]^T$. All the other weights are initialized as 0. 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 $\mathbf{x} = [1, 1]^T$.

(3 marks)


$x_0 = 1$
$a_0^{(2)} = 1$
x_0
x_1
x_2
$a_0^{(2)}$
$a_1^{(2)}$
$a_2^{(2)}$
$a_1^{(3)}$
$h_\theta(\mathbf{x})$
Layer 1
Layer 2
Layer 3

*** END OF PAPER ***