香 港 大 學

THE UNIVERSITY OF HONG KONG

Bachelor of Engineering

Department of Electrical & Electronic Engineering

ELEC3249: Pattern Recognition and Machine Intelligence

2020-2021 Semester 2

Online Examination

Date: ___15 May 2021_____ Time: ___9:30am-12:30pm_____

* We have a total of 6 problems and each problem contains several sub-questions. Each problem is shown in one separate page.
* Open book and notes exam. Candidates are permitted to refer to any electronic/printed/handwritten materials in the examination. Internet searching and crowdsourcing from group messages, online forums or social media, etc. are strictly forbidden.
* The maximum possible score on this exam is 100. You have 3 hours.
* Do not spend too much time on any problems. If you get stuck on any of the problems, move on to another one and come back to that problem if you have time.
* One question may contain contents from multiple lectures.
* There may not exist a standard solution for some problems and write down your understanding.
* Be concise. Do not write too much for a question. The key points are the most important.
* Do not overthink. The problems are not very difficult. Good Luck!

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."

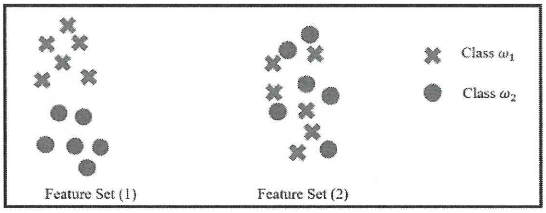
(Use a fresh page for each problem)

Problem 1: Short Questions (16 marks)

Please answer the following short questions.

(a) **Please** summarize what feature extraction is in **1-2** sentences and list **two** properties of good features for classification. In the following, **which** set of features are good for classification? And explain why.

(5 marks)



(b) **Please** describe what model generalization is in **one** sentence and list **two** strategies to improve model generalization. **Does** a lower error on the training set always imply a higher accuracy on the test set? And explain why.

(5 marks)

(c) **Please** summarize the differences between generative models and discriminative models for pattern classification in **2-3** sentences. Please give **an example** of a classifier belonging to generative models and describe how to get the classification results given a new pattern and the trained model. And, similarly, **give an example** of a discriminative model and describe how to get the corresponding classification results given a new pattern and the trained model.

(6 marks)

Problem 2: Maximum Likelihood Estimation and Bayesian Estimation (18 marks)

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

$$p(x|\theta) = \frac{1}{2\theta}\exp(-\frac{|x|}{\theta})$$

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.

(8 marks)

(b) Assume that the parameter θ has a priori probability density function given by $p(\theta) = 0.5[\delta(\theta - 1) + \delta(\theta - 3)]$, where δ is the ideal unit impulse function informally defined as:

$$\delta(y-k) = \begin{cases} \infty & y = k \\ 0 & y \neq k \end{cases} \qquad and \qquad \int_{-\infty}^{+\infty} \delta(y-k) = 1$$

Suppose data samples $D = \{x_1, x_2, \ldots., x_n\}$ are given. **Please** calculate the posterior distribution over θ, i.e., $p(\theta|D)$ in Bayesian estimation. Hint: you can keep the impulse function in your results.

(5 marks)

(c) **Please** compare maximum likelihood estimation and Bayesian estimation. Hint: include how they obtain the probability density and how they compute the likelihood given a new sample x_{new}.

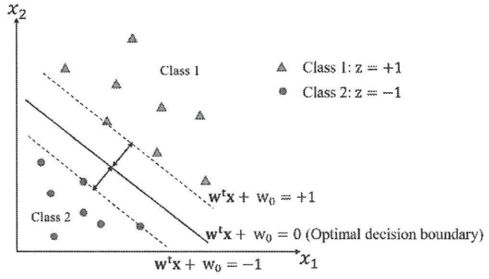
(5 marks)

Problem 3: Linear Discriminant Functions and Support Machines (20 marks)

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).


x_2
Class 1
Class 1: z = +1
Class 2: z = −1
$\mathbf{w}^t\mathbf{x} + w_0 = +1$
Class 2
$\mathbf{w}^t\mathbf{x} + w_0 = 0$ (Optimal decision boundary)
$\mathbf{w}^t\mathbf{x} + w_0 = -1$
x_1

(a) **How** many support vectors are shown in the Figure? **What** is the margin of separation (in one sentence)? **Prove** that the margin here is given by $\frac{1}{||\mathbf{w}||}$.

(5 marks)

(b) Given the training data $D = \{(\mathbf{x}_1, z_1), (\mathbf{x}_2, z_2), \ldots, (\mathbf{x}_n, z_n)\}$ and

$$z_k = \begin{cases} +1 & \text{if } \mathbf{x}_\mathrm{k} \text{ is in class 1} \\ -1 & \text{if } \mathbf{x}_\mathrm{k} \text{ is in class 2} \end{cases}$$

SVM solves an optimization problem (notes page 27-31) with constraints using Lagrangian multipliers. The Lagrangian function for solving the support vector machine problem is given by:

$$L(\mathbf{w}, w_0, \lambda) = \frac{1}{2} \| \mathbf{w} \|^2 - \sum_{k=1}^{n} \lambda_k [z_k(\mathbf{w}^t\mathbf{x}_k + w_0) - 1], \quad \lambda_k \geq 0$$

Please explain the physical significance of the part $\frac{1}{2}||\mathbf{w}||^2$ and $\sum_{k=1}^{n} \lambda_k \, [z_k(\mathbf{w}^t\mathbf{x}_k + w_0) - 1]$.

(5 marks)

(c) Following the formulation in (b), in solving the problem, we minimize $L(\mathbf{w}, w_0, \lambda)$ with respect to $\mathbf{w}$ and w_0 , and maximize $L(\mathbf{w}, w_0, \lambda)$ with respect to $\lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$. **Please** explain the intuition behind this strategy. Hint: discuss the effects of minimization and maximization separately.

(5 marks)

(d) For non-support vectors $\mathbf{x}_k$, **what** is the value of λ_k and explain why using your own words. Hint: consider the maximization part in (c).

(5 marks)

Problem 4: k-means and Gaussian Mixture Model (17 marks)

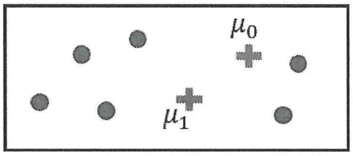
Please answer the following questions related to k-means and gaussian mixture models (GMM).

(a) **Write down** the key steps of the k-means algorithm and the EM procedure to solve GMM and **show** their relationships and differences.

(8 marks)

(b) Consider applying EM to train a Gaussian Mixture Model (GMM) to cluster the data below into two clusters. The '+' points indicate the current means μ_0 and μ_1 of the two gaussian mixture components. **Draw** the directions in which μ_0 and μ_1 will move during the next M-step and explain why (hint: consider the E-step assignment).

(5 marks)



(c) If the number of clusters (or components) and the initializations for the centers are the same for k-means clustering and GMM trained with EM, **will** they converge to the same centers (i.e., means)? Explain why.

(4 marks)

Problem 5: Principal Component Analysis and Linear Discriminant Analysis (14 marks)

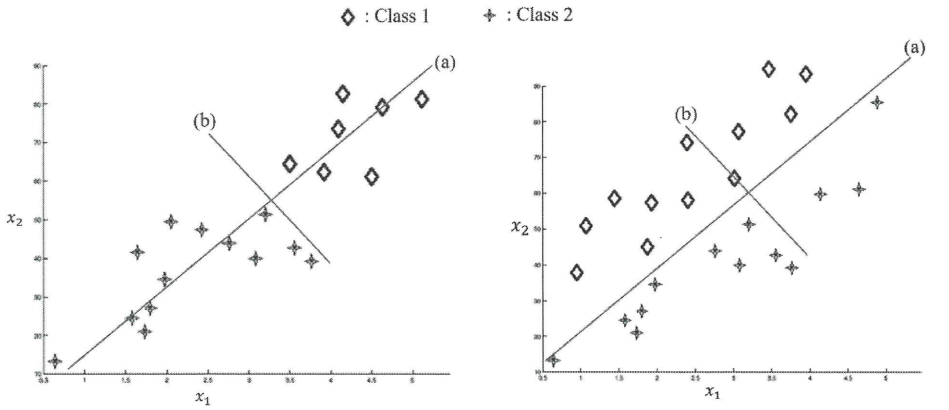
Please answer the following questions related to principal component analysis (PCA) and linear discriminant analysis (LDA).

(a) Please compare PCA and LDA.

(6 marks)

(b) The following Figure shows two datasets: dataset 1 and dataset 2). For dataset 1 and dataset 2, **which line** is closer to the first principal component (or basis) created by PCA, and **which line** is closer to the first axis created by LDA? Please write down your answer for **different** scenarios (i.e., datasets and PCA or LDA).

(4 marks)



Dataset 1

Dataset 2

(c) **Can** we correctly classify the two datasets by using a threshold function after projecting onto the first axis chosen by PCA and LDA? Please write your answer for **different** scenarios (datasets and PCA or LDA).

(4 marks)

Problem 6: Neural Networks and Deep Learning (15 marks)

Please answer the following questions related to neural network and deep learning.

(a) **Please** write down the major advantage of deep learning in one sentence and list two driven factors make it work so well in recent years.

(3 marks)

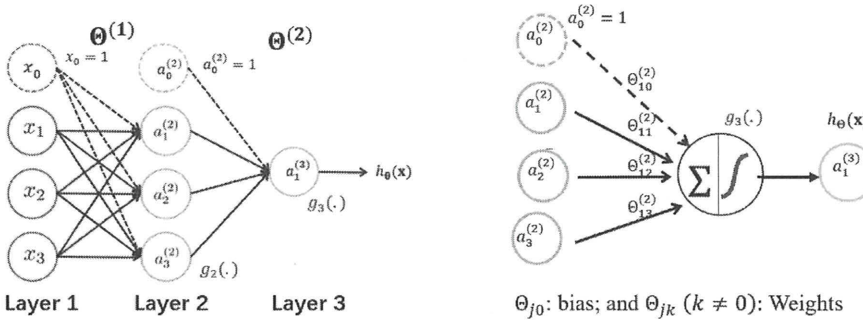
(b) **Please** write down **two** strategies in deep learning to improve model generalization and avoid overfitting.

(2 marks)

(c) The following Figure shows a three-layer neural network designed for two category classification:

The input $\mathbf{x} = [x_1, x_2, x_3]^T$ and the network parameters are $\mathbf{\Theta}^{(1)} = \begin{bmatrix} \Theta_{10}^{(1)} & \Theta_{11}^{(1)} & \Theta_{12}^{(1)} & \Theta_{13}^{(1)} \\ \Theta_{20}^{(1)} & \Theta_{21}^{(1)} & \Theta_{22}^{(1)} & \Theta_{23}^{(1)} \\ \Theta_{30}^{(1)} & \Theta_{31}^{(1)} & \Theta_{32}^{(1)} & \Theta_{33}^{(1)} \end{bmatrix}$

and $\mathbf{\Theta}^{(2)} = [\Theta_{10}^{(2)}\ \Theta_{11}^{(2)}\ \Theta_{12}^{(2)}\ \Theta_{13}^{(2)}]$. The activation function g_2 is ReLU function $g_2(z) = \begin{cases} z\ if & z > 0 \\ 0\ if & z \le 0 \end{cases}$ and g_3 is sigmoid function: $g_3(z) = \frac{1}{1 + \exp(-z)}$



(i) If the network parameters $\mathbf{\Theta}^{(1)}$and $\mathbf{\Theta}^{(2)}$are all initialized with zeros, please calculate the output value $h_\Theta(\mathbf{x})$ given an input $\mathbf{x} = [x_1, x_2, x_3]^T$.

(2 marks)

(ii) Following the initialization in (i), to train the network, we adopt the binary cross-entropy loss:

$$L(\mathbf{\Theta}) = -[\,y \log(h_\Theta(\mathbf{x})) + (1 - y) \log(1 - h_\Theta(\mathbf{x}))]$$

please calculate the gradient of $L(\mathbf{\Theta})$ with respect to all the parameters $\mathbf{\Theta}^{(1)}$ and $\mathbf{\Theta}^{(2)}$, e.g. $\frac{\partial L(\mathbf{\Theta})}{\partial \Theta_{ij}^{(l)}}$ for different layer l given a pair of labeled data $(\mathbf{x}, y)$.

(5 marks)

(iii) Is initializing all the neural network parameters with zeros a good idea? Explain why.

(3 marks)