City University of Hong Kong

Course code & title	:	CS5487 Machine Learning
Session	:	Semester B 2021/22
Time allowed	:	Two hours
Format	:	Online

1. The final exam has 6 pages including this page, consisting of 4 questions.
2. The following resources are allowed on the exam:
   - You are allowed a cheat sheet that is **one** A4 page (**double-sided**) handwritten with pen or pencil.
3. All other resources are not allowed, e.g., internet searches, classmates, other textbooks.
4. Answer the questions on physical paper using pen or pencil.
   - Answer **ALL** questions.
   - Remember to write your **name, EID, and student number** at the top of each answer paper.
5. You should stay on Zoom during the entire exam time.
   - If you have any questions, use the private chat function in Zoom to message Antoni.
6. Final submission
   - Take pictures of your answer paper and submit it to the "Final Exam" Canvas assignment. You may submit it as jpg/png/pdf.
   - *It is the student's responsibility to make sure that the captured images are legible. Illegible images will be graded as is, similar to illegible handwriting.*
   - If you have problems submitting to Canvas, then email your answer paper to Antoni (abchan@cityu.edu.hk).
7. **CS Departmental Hotline (phone, whatsapp, wechat):** +852 6375 3293

Question	**1**					**2**				**3**				**4**					**total**
Max Marks	**25**					**25**				**25**				**25**					**100**
CILO Question Weights (% of exam)																			
	(a)	(b)	(c)	(d)	(e)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(e)	
CILO 1	5	5				5				5	5			5	5				35
CILO 2																			0
CILO 3					5			5					5			5			20
CILO 4			5	5			5		10			10					5	5	45

Statement of Academic Honesty

Below is a **Statement of Academic Honesty**. Please read it.

I pledge that the answers in this exam are my own and that I will not seek or obtain an unfair advantage in producing these answers. Specifically,

- I will not plagiarize (copy without citation) from any source;
- I will not communicate or attempt to communicate with any other person during the exam; neither will I give or attempt to give assistance to another student taking the exam; and
- I will use only approved devices (e.g., calculators) and/or approved device models.
- I understand that any act of academic dishonesty can lead to disciplinary action.

I pledge to follow the Rules on Academic Honesty and understand that violations may led to severe penalties.

Name:
EID:
Student ID:
Signature:

(a) If you have not already, *copy the entire above statement of academic honesty to your answer sheet*. Fill in your name, EID, and student ID, and sign your signature to show that you agree with the statement and will follow its terms.

Problem 1 EM for MAP estimation [25 marks]

Let X be the observed data, Z the corresponding hidden values, and θ the parameters. We will use the EM algorithm to find the MAP solution of θ, i.e., the maximum of the posterior distribution over parameters $p(\theta|X)$. In the E-step, we obtain the MAP Q function by taking the expectation of the posterior $\log p(\theta|X,Z)$,

$$Q_{MAP}(\theta; \hat{\theta}^{\text{old}}) = \mathbb{E}_{Z|X,\hat{\theta}^{\text{old}}}[\log p(\theta|X,Z)]. \tag{1}$$

In the M-step, $Q_{MAP}(\theta; \hat{\theta}^{\text{old}})$ is maximized with respect to θ.

(a) [5 marks]: Show that the E- and M-steps of the MAP-EM algorithm can be written as

$$\begin{aligned} \mathrm{E-step}: \quad Q(\theta; \hat{\theta}^{\text{old}}) &= \mathbb{E}_{Z|X,\hat{\theta}^{\text{old}}}[\log p(X,Z|\theta)], \\ \mathrm{M-step}: \quad \hat{\theta}^{\text{new}} &= \operatorname*{argmax}_{\theta} Q(\theta; \hat{\theta}^{\text{old}}) + \log p(\theta). \end{aligned} \tag{2}$$

How is this related to the ordinary EM algorithm?

Now consider a univariate GMM with 2 components,

$$p(x) = \pi_1 \mathcal{N}(x|\mu_1, \sigma_1^2) + (1-\pi_1)\mathcal{N}(x|\mu_2, \sigma_2^2), \tag{3}$$

where $\theta = \{\pi_1, \mu_1, \mu_2\}$ are the parameters and the variances σ_j^2 are *known*. The prior distribution is $p(\theta) = p(\pi_1)p(\mu_1)p(\mu_2)$ where

$$p(\pi_1) = 1, \quad 0 \le \pi_1 \le 1, \tag{4}$$

$$p(\mu_1) = \mathcal{N}(\mu_1|\mu_0, \sigma_0^2), \tag{5}$$

$$p(\mu_2) = \mathcal{N}(\mu_2|\mu_0, \sigma_0^2). \tag{6}$$

(b) [5 marks] Write down the complete data log-likelihood, $\log p(X, Z|\theta)$. (For convenience, you can define $\pi_2 = 1 - \pi_1$.)

(c) [5 marks] Derive the E-step, i.e., the Q function, $Q(\theta; \hat{\theta}^{\text{old}})$.

(d) [5 marks] Derive the M-step, i.e., the parameter updates of θ.

(e) [5 marks] What is the intuitive explanation of the E- and M-steps in (c) and (d)?

....

Problem 2 BDR with unbalanced loss function [25 marks]

Consider a two-class problem with $y \in \{0, 1\}$ and measurement x, with associated prior distribution $p(y)$ and class-conditional densities $p(x|y)$. In this problem, assume that the loss-function is:

$$L(g(x), y) = \begin{cases} 0, & g(x) = y \\ \ell_0, & y = 0 \text{ and } g(x) = 1 \\ \ell_1, & y = 1 \text{ and } g(x) = 0, \end{cases} \tag{7}$$

where $g(x)$ is the classifier prediction for x. In other words, the loss for misclassification is different for each class.

(a) [5 marks] When might this type of loss function be useful? Can you give a real-world example?

(b) [5 marks] Derive the Bayes decision rule (BDR) for y. Write the BDR as a log-likelihood ratio test. What is the threshold?

(c) [5 marks] Explain how the loss values ℓ_0 and ℓ_1 influence the threshold.

(d) [10 marks] Derive the BDR for the specific case when the class-conditional densities are Gaussians,

$$p(x|y = 0) = \mathcal{N}(x|\mu_0, \sigma^2), \quad p(x|y = 1) = \mathcal{N}(x|\mu_1, \sigma^2). \tag{8}$$

and the prior distribution is uniform $p(y = 0) = p(y = 1) = 0.5$. Write down the rule for selecting $y = 0$ and $y = 1$, given x.

. . . .

Problem 3 Soft-margin SVM with 2-norm penalty [25 marks]

The soft-margin SVM primal problem typically uses a 1-norm penalty on the slack variables (i.e., $C\sum_i \xi_i$). Consider the soft-margin SVM primal problem using a **2-norm** penalty on the slack variables, where $X = [x_1, \cdots, x_n]$ are the input features with $x_i \in \mathbb{R}^d$, and $y = [y_1, \cdots, y_n]^T$ are the class values with $y_i \in \{+1, -1\}$:

$$\begin{aligned} \min_{w,b,\xi} \; & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i^2 \\ \text{s.t. } & y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \forall i \\ & \xi_i \geq 0, \quad \forall i. \end{aligned} \tag{9}$$

ξ_i is the slack variable that allows the ith point to violate the margin, and C the hyperparameter.

(a) [5 marks] Show that the non-negative constraint $\xi_i \geq 0$ is redundant, and hence can be dropped.

(b) [5 marks] Let α_i be the Lagrange multiplier for the i-th inequality constraint. Write down the Lagrangian $L(w, b, \xi, \alpha)$ for the problem. Derive conditions for the minimum of $L(w, b, \xi, \alpha)$ w.r.t. $\{w, b, \xi\}$.

(c) [10 marks] Derive the dual function $L(\alpha) = \min_{w,b,\xi} L(w, b, \xi, \alpha)$, and write down the dual problem for SVM with 2-norm.

(d) [5 marks] Comment on the similarity and differences between the dual problems for the SVM with 2-norm penalty and the original SVM with 1-norm penalty. What is the interpretation of any differences?

....

Problem 4 Kernel perceptron [25 marks]

For a training set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{+1, -1\}$, the Perceptron algorithm is as follows:

Perceptron algorithm

---

set $w = 0, b = 0, R = \max_i ||x_i||$
repeat
    for $i = 1, \ldots, n$ **do**
        if $y_i(w^T x_i + b) \leq 0$ **then**
            set $w \leftarrow w + \eta y_i x_i$
            set $b \leftarrow b + \eta y_i R^2$
        end if
    end for
until there are no classification errors

For an x_* input, the classifier is $y_* = \text{sign}(w^T x_* + b)$.

(a) [5 marks] Show that the learning rate η is not relevant in the Perceptron algorithm.

(b) [5 marks] Using (a), we let $\eta = 1$ without loss of generality. Show that w learned by the Perceptron algorithm must take the form $w = \sum_{i=1}^{n} \alpha_i y_i x_i$, where $\alpha_i \geq 0$, $\forall i$.

(c) [5 marks] What is the interpretation to the parameters α_i?

(d) [5 marks] Using (b) derive an equivalent Perceptron algorithm (the dual perceptron).

(e) [5 marks] Apply the kernel trick the dual perceptron algorithm to obtain the *kernel perceptron algorithm*. What is the kernelized decision function?

. . . .