City University of Hong Kong

Course code & title	:	CS5487 Machine Learning
Session	:	Semester A 2020/21
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).

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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)	(e)	(a)	(b)	(c)	
CILO 1	5	5				5	5			5	5				5			35
CILO 2																		0
CILO 3					5				5					5			5	20
CILO 4			5	5				10				5	5			15		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:

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

(a) Copy the 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 GMMs with shared mean [25 marks]

Let $x \in \mathbb{R}$ be distributed as a (univariate) Gaussian mixture model with K components. Assume that the Gaussian components have a *shared* mean μ and different variances σ_j^2,

$$p(x|\theta) = \sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu, \sigma_j^2), \tag{1}$$

where $\theta = \{\{\pi_j, \sigma_j^2\}_{j=1}^K, \mu\}$ are the parameters. This is called a *Gaussian scale mixture*. Let $X = \{x_1, \cdots, x_n\}$ be a set of observed samples, and $Z = \{z_1, \cdots, z_n\}$ the set of corresponding hidden values.

(a) [5 marks] Plot an example of the Gaussian scale mixture with parameters

$$K = 2, \quad \pi_1 = \pi_2 = \tfrac{1}{2}, \quad \mu = 0, \quad \sigma_1^2 = 1, \sigma_2^2 = 9. \tag{2}$$

Intuitively, what type of data could be well modeled with a Gaussian scale mixture?

(b) [5 marks] For the general case (not assuming the parameters in (a)), write down the complete data log-likelihood, $\log p(X, Z|\theta)$,

(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)? Note any differences or similarities with the EM algorithm for standard GMMs (without a shared mean).

.....

试用水印

Problem 2 BDR and Naive Bayes for discrete variables [25 marks]

For high-dimensional observation spaces, it might be difficult to learn a joint density over the space (e.g., if not enough data is available). One common assumption is to use a "Naive Bayes" model, where we assume that the individual features (dimensions) are conditionally independent given the class,

$$p(x|y=j) = \prod_{i=1}^{d} p(x_i|y=j), \tag{3}$$

where $x = [x_1, \cdots, x_d]^T$ is the observation vector, and x_i is the individual feature. While the features are conditionally independent given the class, the features are still dependent in the overall distribution of observations $p(x)$ (similar to a GMM with diagonal covariance matrices).

Let the vector x be a collection of d binary-valued features, i.e.

$$x = [x_1, \cdots, x_d]^T, \quad x_i \in \{0, 1\}. \tag{4}$$

The individual features x_i are binary, e.g., each x_i is an indicator variable representing the presence/absence of a specific property in the data sample. Assume there are 2 classes, with class variable $y \in \{0, 2\}$ and prior distribution $p(y=1) = \pi$, $p(y=2) = 1-\pi$. Now define

$$p_i = p(x_i = 1|y = 1), \ \forall i \tag{5}$$
$$q_i = p(x_i = 1|y = 2), \ \forall i. \tag{6}$$

The goal is to recover the class y given a measurement x.

Assignment Project Exam Help
Add WeChat powcoder

(a) [5 marks] Interpret in words the meaning of p_i, q_i, and $\log \frac{p_i}{q_i}$.

(b) [5 marks] Write down the class-conditional densities $p(x|y=1)$ and $p(x|y=2)$ in terms of p_i, q_i, and x_i.

(c) [10 marks] Using the 0-1 loss function, derive the Bayes decision rule (BDR) for class y for a given x. State the derived BDR in the following form:

$$y = \begin{cases} 1, & g(x) > 0, \\ 2, & \text{otherwise}, \end{cases} \tag{7}$$

for some decision function $g(x)$ (which you derive).

(d) [5 marks] What is the intuitive interpretation of the decision function $g(x)$? In particular, what is the role of feature x_i, its associated p_i and q_i, and the prior π? What is the shape of the decision surface of this BDR classifier?

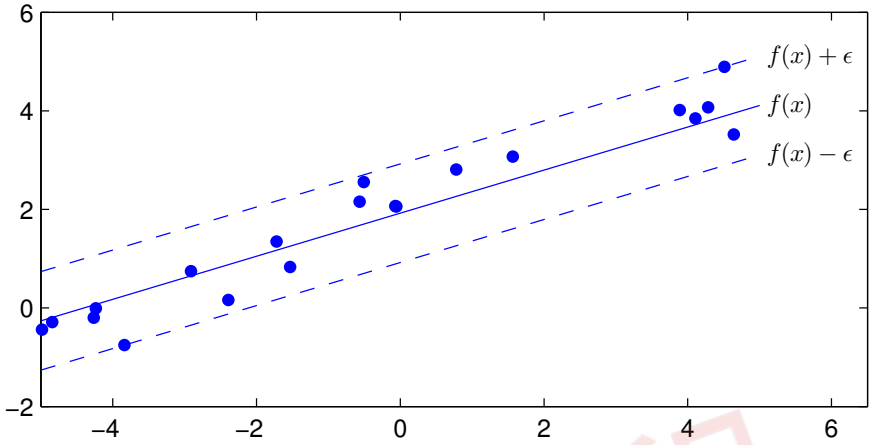
.....

Problem 3 Support vector regression [25 marks]

In this problem, we will consider support vector regression (SVR), which applies margin principles from SVMs to regression. The goal is to learn a linear function,

$$f(x) = w^T x + b, \tag{8}$$

which fits a given dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Suppose we form a "band" or "tube" around $f(x)$ with width ϵ (see figure below).



We can consider any training pair (x_i, y_i) that falls inside of the tube as correctly regressed, while points falling outside of the tube are errors. Assuming that ϵ is known, the SVR primal problem is

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad & y_i - (w^T x_i + b) \le \epsilon \\ & (w^T x_i + b) - y_i \le \epsilon, \quad \forall i \end{aligned} \tag{9}$$

试用水印 WeCha p w

(a) [5 marks] Explain the role of the objective function and the inequality constraints in the SVR primal problem. When are the equality constraints "active" or "inactive"?

(b) [5 marks] Let $\{\alpha_i\}$ and $\{\hat{\alpha}_i\}$ be the Langrange multipliers for the first and second sets of inequality constraints, respectively. Show that the Lagrangian $L(w, b, \alpha, \hat{\alpha})$ for the SVR primal problem is:

$$L(w, b, \alpha, \hat{\alpha}) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^n \alpha_i(\epsilon - y_i + (w^T x_i + b)) - \sum_{i=1}^n \hat{\alpha}_i(\epsilon + y_i - (w^T x_i + b)). \tag{10}$$

(c) [5 marks] Use the Lagrangian to derive conditions for the minimum of the SVR primal problem.

(d) [5 marks] Derive the SVR dual problem.

(e) [5 marks] Using the KKT conditions, at the optimum of the SVR dual problem, what do the Lagrange multipliers $\alpha_i, \hat{\alpha}_i$ indicate about the data point (x_i, y_i)?

.....

Problem 4 Kernel logistic regression [25 marks]

Consider the two-class logistic regression problem with a prior on the weight vector w, where $x \in \mathbb{R}^d$ is the input vector and $y \in \{0, 1\}$ is the class label. The training set is $\mathcal{D} = \{X, y\}$, where $X = [x_1, \cdots, x_n]$ are the input vectors, and $y = [y_1, \cdots, y_n]^T$ are the class labels. The conditional probability of the output class is

$$p(y_i|x_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, \tag{11}$$

where $\pi_i = \sigma(w^T x_i)$ is the conditional probability that x_i belongs to class 1, and $\sigma(a) = \frac{1}{1+e^{-a}}$ is the logistic sigmoid function. The prior distribution on w is a zero-mean Gaussian with known precision matrix Γ (i.e., inverse of the covariance matrix), $p(w) = \mathcal{N}(w|0, \Gamma^{-1})$.

The MAP estimate can be obtained using the Newton-Raphson iterations

$$w^{(new)} = (XRX^T + \Gamma)^{-1} XRz, \tag{12}$$

where $\{R, z, \pi\}$ are calculated from the previous $w^{(\text{old})}$,

$$\pi_i = \sigma(x_i^T w^{(\text{old})}), \quad \pi = [\pi_1, \cdots, \pi_n]^T, \tag{13}$$

$$R = \text{diag}(\pi_1(1 - \pi_1), \cdots, \pi_n(1 - \pi_n)), \tag{14}$$

$$z = X^T w^{(old)} - R^{-1}(\pi - y) \tag{15}$$

(a) [5 marks] Consider the case when the precision matrix is $\Gamma = \lambda I$. How does Γ help to regularize the estimate of w in (12)?

(b) [15 marks] Derive *kernel* logistic regression, i.e., apply the *kernel trick* to calculate probability $\pi_* = \sigma(w^T x_*)$, and to MAP estimation using the Newton-Raphson iterations.

(c) [5 marks] Discuss the role of the prior precision Γ in kernel logistic regression.

.....

ps po