UNIVERSITY COLLEGE LONDON

EXAMINATION FOR INTERNAL STUDENTS

MODULE CODE : **COMP0078**

ASSESSMENT PATTERN : **COMP0078A7PD**

MODULE NAME : **COMP0078 - Supervised Learning**

LEVEL: : **Postgraduate**

DATE : **25-Apr-2022**

TIME : **14:30**

Controlled Condition Exam: 3 Hours exam

You cannot submit your work after the date and time shown on AssessmentUCL – you must ensure to allow sufficient time to upload and hand in your work

This paper is suitable for candidates who attended classes for this module in the following academic year(s):

Year
2021/22

Additional material	N/A
Special instructions	N/A
Exam paper word count	N/A

TURN OVER

Supervised Learning, COMP0078 (–)

Main Summer Exam Period, 21/22

Suitable for Cohorts: 21/22

Answer ALL EIGHT questions.

Notation: Let $[m] := \{1, \ldots, m\}$. We also overload notation so that

$$[\texttt{pred}] := \begin{cases} 1 & \texttt{pred} \text{ is true} \\ 0 & \texttt{pred} \text{ is false} \end{cases} .$$

Recommended answer lengths are only for informational purposes, no points will be deducted for shorter or longer answers.

Marks for each part of each question are indicated in square brackets.

Standard calculators are permitted.

1. Consider a dataset $(x_1, y_1), \ldots, (x_m, y_m) \in R \times R$, i.e., both x_i and y_i are real-valued scalars for all i. Suppose we wish to fit a linear+offset model,

$$\hat{y} = ax + b\,.$$

   a. Suppose we have a software library that performs least squares (without offset). Explain how we can use that software library to fit the linear+offset model.

[5 marks]

[Question 1 cont. over page]

[Question 1 cont.]

b. Suppose we use the least squares criterion to fit a linear model to these data, by solving the following optimisation problem:

$$(a^*, b^*) = \text{argmin}_{a,b} \sum_{i=1}^{m} (y_i - ax_i + b)^2$$

Assume that the solution is unique. Which of the following statements are necessarily true? (If a statement is false you do not need to disprove it - but you need to prove true statements).

i. $\sum_{i=1}^{m} (y_i - a^* x_i + b^*) y_i = 0$

ii. $\sum_{i=1}^{m} (y_i - a^* x_i + b^*) x_i^2 = 0$

iii. $\sum_{i=1}^{m} (y_i - a^* x_i + b^*) x_i = 0$

iv. $\sum_{i=1}^{m} (y_i - a^* x_i + b^*)^2 = 0$

[5 marks]

2. a. i. State the ridge regression objective function.

ii. Give both the primal and dual solutions (no derivation required).

iii. Discuss why depending on the setting, why one might select to use the primal solution or the dual solution.

[5 marks]

b. Suppose $K : X \times X \rightarrow R$ is a kernel function. Then consider the inequality,

$$K(x_1, x_2)^2 \leq K(x_1, x_1) \times K(x_2, x_2)\,.$$

i. Is this inequality true for all $x_1, x_2 \in X$?

ii. Support your answer to "i." by providing an argument.

[5 marks]

3. Given the data set,

$$((1,1),+1),((2,2),+1),((2,0),+1),((0,0),-1),((1,0),-1),((0,1),-1)$$

a. Plot the dataset as well as the maximal margin hyperplane. Give the equation of the maximum margin hyperplane. (Note: this may be done by inspection rather than by derivation and proof). Finally, identify the support vector(s)

[5 marks]

b.

Suppose we remove a datapoint from the above training set. How will the maximum margin change?

i. Strictly increases.

ii. Strictly decreases.

iii. Stays the same.

iv. Depends on the removed data point. If you choose this then explain how it depends on the data point.

[5 marks]

[Question 5 cont. on next page]

[Question 5 cont.]

4. a. Given the data set with five examples,

$$((1,1),+1),((1,-1),+1),((-1,1),+1),((-1,-1),-1),((0,0),-1)$$

Plot the dataset. Consider training a classifier with Adaboost using decision stumps. Indicate which example(s) have their weights increased after a single iteration of boosting. Explain why.

[5 marks]

b. Suppose AdaBoost is run on m training examples, and suppose on every round that the weighted training error ϵ_t of the tth weak hypothesis is at most $1/2-\gamma$, for some $\gamma > 0$. What is an upper bound on the number of iterations, T (if such a number exists), until the combined hypothesis H is *always* consistent with the m training examples, i.e., achieves zero (unweighted) training error? Your answer should only be expressed in terms of m and γ. Justify your answer.

[5 marks]

[Question 5 cont. over page]

[Question 5 cont.]

5. a. Consider the expert setting with absolute loss in the range $[0, C]$ where the upper bound C is know in advance to the algorithm,

Protocol:

$L_A = 0$

For $t = 1$ To m Do

Get expert predictions	$\mathbf{x}_t \in [0, C]^n$
Predict	$\hat{y}_t \in [0, C]$
Get label	$y_t \in [0, C]$

$L_A = L_A + |y_t - \hat{y}_t|$

i. Give your algorithm.

ii. Give the regret bound for your algorithm.

iii. Sketch the proof of your regret bound.

iv. Explain how the regret bound depends on C.

[10 marks]

[Question 5 cont. on next page]

[Question 5 cont.]

b. Consider the following algorithm for the expert setting with absolute loss,

Algorithm:

$L_A = 0$

$\mathbf{w}_1 = (1/n, \ldots, 1/n)$

For $t = 1$ To m Do

Receive expert predictions $\mathbf{x}_t \in [0,1]^n$

Predict $\hat{y}_t = \mathbf{w}_t \cdot \mathbf{x}_t$

Receive label $y_t \in [0,1]$

Update weights $\mathbf{w}_{t+1} = \textbf{update}(\mathbf{w}_t, \mathbf{x}_t, y_t)$

% Comment: assume $\mathbf{w}_{t+1} \in \{\mathbf{w} \in [0,1]^n : \sum_{i=1}^n w_i = 1\}$

$L_A = L_A + |y_t - \hat{y}_t|$

Suppose for this algorithm we have the following regret bound,

$$L_A - \min_{1 \le i \le n} \sum_{t=1}^{m} |y_t - x_{t,i}| \le B\,.$$

Explain in detail how we can use this algorithm to obtain the following expected regret bound for the allocation (Hedge) setting

$$\mathbb{E}[\sum_{t=1}^{m} \ell_{t,\hat{y}_t}] - \min_{1 \le i \le n} \sum_{t=1}^{m} \ell_{t,i} \le B\,.$$

Recalling that the allocation (Hedge) protocol is,

Protocol:

Nature selects $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots, \boldsymbol{\ell}_m \in [0,1]^n$.

For $t = 1$ To m Do

Predict $\hat{y}_t \in [n]$

Receive loss vector $\boldsymbol{\ell}_t \in [0,1]^n$

[10 marks]

6. Consider the following questions in the context of the model of *PAC learning*.

   i. Give an example of an infinite hypothesis class with **finite** sample complexity.

   ii. Give an example of an infinite hypothesis class with **infinite** sample complexity.

   iii. For a finite hypothesis class what does the cardinality of the hypothesis class reveal about the sample complexity. [Recommended answer length 2-4 sentences]

   iv. For the infinite hypothesis class of "part ii" sketch an argument and provide intuitions on why the sample complexity is infinite. [Recommended answer length 4-8 sentences]

[10 marks]

7. This question concerns spectral clustering. Let L denote the graph Laplacian of an n-vertex undirected graph. Let μ_1, μ_2 denote the first and second eigenvector of L respectively. The (normalised) first eigenvector is $\mu_1 = (1/\sqrt{n}, \ldots, 1/\sqrt{n})$. The second eigenvector μ_2 is orthogonal to the first eigenvector and also has eigenvalue equal to zero. The third eigenvector has eigenvalue $c > 0$ every other eigenvector has an eigenvalue larger than c.

   a. Give a brief explanation why the first eigenvector and first eigenvalue have the values given above.

   [5 marks]

   b. What can be inferred about the structure of the graph corresponding to L? Give the mathematical reasoning behind your inference.

   [5 marks]

   c. Suppose that we are given that the value at first index of the (normalised) second eigenvector is $\mu_{2,1} = a$. What can we infer about the values of μ_2 at other indices?

   [5 marks]

   d. Using the results of (b) and (c) above give an argument that supports the use of the second eigenvector for spectral clustering.

   [5 marks]

8. a. Recall the Hedge and Exp3 settings. Explain the similarities and differences between the settings. (A good answer should typically use no more than 4 sentences).

   [5 marks]

   b. We only proved an upper bound on the expected regret for the Exp3 algorithm. Would it be possible to improve the upper bound of Exp3 so that it matches that of the Hedge algorithm? What if we could use a different algorithm for this bandit model? Explain your reasoning.

   [5 marks]