UNIVERSITY COLLEGE LONDON

EXAMINATION FOR INTERNAL STUDENTS

MODULE CODE	:	**COMP0078**
ASSESSMENT PATTERN	:	**COMP0078A7UA; COMP0078A7PA**
MODULE NAME	:	**Supervised Learning**
LEVEL:	:	**Undergraduate Masters; Postgraduate**
DATE	:	**23 April 2019**
TIME	:	**14:30**
TIME ALLOWED	:	**2 hrs 30 mins**

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

Year
2018-19

EXAMINATION PAPER CANNOT BE REMOVED FROM THE EXAM HALL. PLACE EXAM PAPER AND ALL COMPLETED SCRIPTS INSIDE THE EXAMINATION ENVELOPE

Hall Instructions	.
Standard Calculators	Y
Non-Standard Calculators	N

TURN OVER

Supervised Learning, COMP0078 (A7U, A7P)

Main Summer Examination Period

There are TEN questions.

Answer ALL TEN questions.

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

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

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

Standard calculators are permitted.

1. a. Consider k-NN. We consider the following problem with two variants of the same underlying data distribution. Both distributions are based on adding noise to the labels of the same underlying function $f^* : X \to \{0, 1\}$ so that

$$p_c(y|x) := \begin{cases} 1 - c & f^*(x) = y \\ c & f^*(x) \neq y \end{cases}$$

For variant 'a' let $c = 0.1$ and for variant 'b' let $c = 0.4$. Denote the value of k to be used for variant 'a' as k_a and as k_b for variant 'b'.

Now suppose the sample size is very large, i.e., $m >> |X|$. What relationship should we expect between k_a and k_b if we want to choose them so as to minimise generalisation error.

i. $k_a = k_b$

ii. $k_a < k_b$

iii. $k_a > k_b$

Explain your reasoning.

[5 marks]

[Question 1 cont. over page]

b. Let $p(x,y)$ denote a probability mass function over (X,Y) where $X = [n]$ and $Y = [k]$ thus $\sum_{x\in X}\sum_{y\in Y} p(x,y) = 1$. For this subquestion let $\boldsymbol{C} \in [0,\infty)^{k\times k}$ be a $k \times k$ matrix of costs. Define $L_{\boldsymbol{C}} : [k] \times [k] \rightarrow [0,\infty)$ as,

$$L_{\boldsymbol{C}}(y,\hat{y}) := [y \neq \hat{y}]C_{y,\hat{y}}$$

as the class-sensitive loss function, i.e., if we incorrectly predict $\hat{y}$ and the correct outcome was y we suffer $C_{y,\hat{y}}$ loss. Derive the Bayes estimator.

Now suppose $X = \{1,2,3,4\}$ and $Y = \{1,2\}$ and

$$\mathbb{P}[X = i] = i/10$$
$$\mathbb{P}[Y = 1|X = i] = 3/10 + i/20$$
$$\mathbb{P}[Y = 2|X = i] = 7/10 - i/20$$

and

$$C = \begin{pmatrix} 0 & 1 \\ 3 & 0 \end{pmatrix}$$

Give the Bayes estimator for this particular problem after performing algebraic simplification.

[5 marks]

2. a. Suppose we apply linear regression and ridge regression with regularisation coefficient $\lambda = 1$ to the dataset $\{((1,2),4),((3,3),9)\}$. Will the solutions be the same or different. Explain why.

[5 marks]

b. Let X be a finite set define $K : X \times X \rightarrow \mathbb{R}$ as

$$K(A,B) := 2^{|A \cap B|}$$

where $A, B \subseteq X$. Is K a kernel? Explain why or why not.

[5 marks]

3. a. Please briefly describe the bagging algorithm. Explain a possible motivation for using bagging in terms of either bias or variance.

[5 marks]

b. Boosting maintains two set of weights. Explain the purpose of these weights. Explain how the weights are updated from round t to round $t+1$.

[5 marks]

4. a. Draw a diagram of linearly separable data set with five examples labeled '$+1$' and five labeled '-1' draw a maximum-margin separating hyperplane. Annotate your diagram so it is clear why the hyperplane you have chosen has the maximum margin. Indicate which vectors if any are support vectors.

[5 marks]

b. For a hard-margin SVM. If we remove one of

i. the examples which is a support vector from the training set,

ii. the examples which is not a support vector from the training set

and retrain without that example. How does the maximum margin change? Explain why.

[5 marks]

5. a. Explain what is a regret bound for an online algorithm and then give an example of a regret bound for an algorithm.

[5 marks]

b. Suppose we wish to use a perceptron with a kernel. Derive a "kernelized" perceptron. Explain the steps of your derivation.

[5 marks]

6. a. Explain the roles of ϵ and δ parameters in the PAC setting.

[5 marks]

b. **Outlier prediction.** Let $X = \mathbb{R}^2$, $Y = \{0, 1\}$, and let $\mathcal{H}$ be the class of origin-centered concentric circles in the plane, that is, $\mathcal{H} = \{h_r : r \in [0, \infty)\}$, where $h_r(x) = [\|x\| \leq r]$. Suppose there exists a distribution $\mathcal{D}$ over (X, Y) such that there is always a consistent learner for any set of samples. Argue that $\mathcal{H}$ is PAC-learnable, and prove a bound on the sample complexity.

[5 marks]

7. a. Explain the construction of the loss vector for exp3. Explain the rationale behind its construction.

[5 marks]

b. Suppose matrix $\boldsymbol{U}$ has a (k, ℓ)-biclustering argue that its rank is no larger $\min(k, l)$.

[5 marks]

8. This question concerns semi-supervised learning.

a. Describe the differences, in brief, of semi-supervised learning as compared to supervised learning and unsupervised learning.

[5 marks]

b. Explain the role of the graph in graph-based semi-supervised learning.

[5 marks]

9. This question concerns kernel methods.

   a. Suppose we wish to "kernelise" the K-nn algorithm, derive the prediction rule.

   [5 marks]

   b. Define the inner product space of bounded square summable sequences $\ell_2 := \{\mathbf{x} \in \mathbb{R}^\infty : \sum_{i=1}^\infty x_i^2 < \infty\}$ with inner product $\langle \mathbf{x}, \mathbf{x}' \rangle = \sum_{i=1}^\infty x_i x_i'$.

   Definition: Given an $X \subset \ell_2$ and a $\Lambda \in (0, \infty)$ define,

   $$\mathcal{H}_{X,\Lambda} = \{\mathbf{x} \mapsto \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) : \mathbf{x} \in X, \mathbf{w} \in \ell_2, \|\mathbf{w}\|^2 \leq \Lambda, \langle \mathbf{w}, \mathbf{x} \rangle \geq 1\}$$

   Prove

   $$\text{VCdim}(\mathcal{H}_{X,\Lambda}) \leq \Lambda^2 \max \|\mathbf{x}\|^2_{\mathbf{x} \in X}$$

   [5 marks]

10. a. What is meant by overfitting? Illustrate your answer with a graph showing the typical training and test performance observed as the complexity of a function class of learners increases.

   b. The VC-dimension appears in PAC learning bounds. Explain "intuitively" or analytically why it can act as a measure of complexity of a hypothesis class.