NATIONAL UNIVERSITY OF SINGAPORE

CS5339: Theory and Algorithms for Machine Learning

(Semester 2: AY2018/19)

Time Allowed: 120 Minutes

INSTRUCTIONS TO CANDIDATES

a) This paper consists of **THREE (3)** questions and **EIGHT (8)** printed pages including this page. Answer all questions.

b) You have 120 minutes to earn 40 marks. Do not spend too much time on any problem. Read them all through first and attack them in the order that allows you to make the most progress.

c) You may quote results that are stated in the lecture notes or homework solutions without re-deriving them if you need the results as part of your answer.

d) Show your work, as partial credit will be given. You will be graded on the correctness and optimality of your answers, and also on your clarity. Be clear and always explain your reasoning.

e) This is an **OPEN BOOK** assessment.

f) Please write your Student Number only. Do not write your name.

Student number:

For Examiner's Use Only		
	Max Marks	Earned Marks
Problem 1	18	
Problem 2	12	
Problem 3	10	
TOTAL:	40	

Problem 1. Short Questions (18 Marks)

a) Assume that the density model used for multi-output regression is $p(\boldsymbol{y}|f(x)) = \frac{1}{Z}\exp\left(-(\boldsymbol{y} - f(x))^T \boldsymbol{A}(\boldsymbol{y} - f(x))\right)$, where $\boldsymbol{y}$ is a d-dimension vector, **A** is a *d* by *d* positive definite matrix and Z is the normalizing constant. What is the corresponding loss function for empirical risk minimization such that minimizing the empirical risk corresponds to maximizing the likelihood?

b) Consider representing the Boolean function $f: \{0,1\}^d \to \{0,1\}$, where *f*(*x*) outputs the value 1 if $\sum_{i=1}^{d} x_i > d - 1$ and *f*(*x*) the value 0 otherwise. Argue that the height of a decision tree representing *f* is at least *d*, where the height is the longest path from the root to a leaf.

c) Consider the quadratic kernel $K(x, x') = (1 + \langle x, x' \rangle)^2$ in $\mathbb{R}^2$. Let $f(x) = 2K\big((2,3), x\big) + 3K((1,2), x)$. Let w be the weight vector representing $f(x)$ in feature space. What is the value of $\|w\|$?

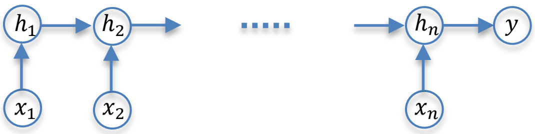
d) **True or False.** Let $K(x, x') = (1 + \langle x, x' \rangle)^2$ be the quadratic kernel in $\mathbb{R}^2$. Consider the class of function defined by thresholding $f(x) = \sum_{i=1}^{100} \alpha_i K(x_i, x)$ where x_i are fixed but α_i are allowed to vary for $i = 1, \dots, 100$. Then the VC-dimension of this class of functions is no more than 6. Justify your answer.

e) **True or False.** Consider using the following class for doing weak learning for classifying strings: $H = \{1_s(x)\colon s\ is\ a\ substring\ of\ length\ d\} \cup \{f(x) = 1\}$, where $f(x) = 1$ is a constant function taking the value 1, and $1_s(x)$ is an indicator function that takes the value 1 when string *s* is a substring in *x* and takes the value 0 otherwise. Assume that the target function *t* is a disjunction of *k* indicator functions $1_{s1}(x) \vee \ldots \vee 1_{sk}(x)$. Then there is a *O*(1/*k*)-weak learning algorithm that uses *H* for learning *t*. Justify your answer.

f) **True or False.** The function class *H* contains only a *single* Boolean function of *d* variables, namely the Parity function. The *Rademacher complexity* of *H* is at least 2*d* as representing Parity using a decision tree requires a tree of height *d* and such decision trees have VC-dimension at least 2*d*. Justify your answer.

Problem 2. Recurrent Neural Networks (12 Marks)

Consider a recurrent neural network that takes in sequences $x_1, x_2, \ldots, x_n$, where $x_i \in \mathbb{R}$ as shown below.



Each vector of hidden units has dimension *k*. There are *k* by *k* weights forming a weight matrix ***W*** between ***h**ᵢ* and ***h**ᵢ₊₁* and a vector ***U*** of length *k* parameterizing the connections between *xᵢ* and ***h**ᵢ*. The *j*-th hidden variable at time *i* is defined by $h_{i,j} = \sigma\left(\sum_{l=1}^{k} W_{j,l} h_{i-1,l} + U_j x_i\right)$, where σ is an activation function. There is only one binary output *y* which has *hₙ* as input and is parameterized by weight vector *w* of length *k*: $y = \sum_{j=1}^{k} w_j h_{n,j}$. For simplicity, none of the units or output has bias. In the following, give efficient solutions in *O*-notation in terms of the parameters of the problem, *n* and *k*.

a) Assume that all parameters are represented in a computer using 64 bits for each parameter. What is the sample complexity of PAC learning this function class? Justify your solution.

b) In the usual recurrent neural networks, the same parameters *W* and *U* are used for every time instance *i*. Assume instead that a different set of parameters *W* and *U* are used for each time instance *i*. Assume that all parameters are still represented in a computer using 64 bits for each parameter. What is the sample complexity for PAC learning this function class?

c) Assume that the gradient is computed using the back-propagation algorithm as described in the lecture notes. What is the amount of memory used by the algorithm, excluding the memory used for storing the parameters of the model? Justify your solution.

d) Assume that we are given the input sequence $x_1, x_2, \ldots, x_n$, and we would like to compute the value y. Describe a memory efficient algorithm for computing y and give the amount of memory used, excluding the memory used for storing $x_1, x_2, \ldots, x_n$ and the parameters of the model.

Problem 3. Winnow (10 Marks)

We analyse the Winnow algorithm, which can be used for online learning of disjunctions of Boolean variables. A disjunction is a concept of the form $x_{c1} \vee x_{c2} \vee \ldots \vee x_{ck}$ where $x_{c1}, x_{c2}, \ldots, x_{ck}$ are *k* Boolean variables selected from a larger set of variables $x_1, x_2, \ldots, x_n$. The Winnow algorithm works as follows:

- Each variable x_i is initialized with the weight $w_i = 1$.
- If $\sum_{i=1}^{n} w_i x_i \geq n$ the algorithm predicts 1, otherwise it predicts 0.
- If the prediction is 1 but correct label is 0, set $w_i = w_i/2$ for all *i* with x_i=1.
- If the prediction is 0 but correct label is 1, set $w_i = 2w_i$ for all *i* with x_i=1.

Let *P* be the number of mistakes where the correct label is 1 and *N* be the number of mistakes where the correct label is 0.

a) Argue that the *increase* in total weights on a mistake where the correct label is 1 is at most *n*.

b) Argue that the *decrease* in total weights on a mistake where the correct label is 0 is at least *n*/2.

c) Using (i) and (ii), argue that $N < 2P + 2$.

d) Argue that $P = O(k \log n)$ hence $P + N = O(k \log n)$.

END OF PAPER