Indian Association for the Cultivation of Science

(Deemed to be University under *de novo* Category)

Master's/Integrated Master's-PhD Program/Integrated Bachelor's-Master's Program/PhD Course

End-Semester (Sem-II/IV) Examination-Spring 2021

Subject: Introduction to Machine Learning ***Subject Code(s): COM-4203/PHD-226***

Full Marks: 50 ***Time Allotted: 3 h***

Answer all questions. Marks for each question are indicated in [].

1. **(a)** Consider an unthreshold perceptron. Suppose you use gradient descent to train it. Will the algorithm always converge to a weight vector with minimum error on the training examples? Explain. [2]

   (b) State the differences between perceptron rule and delta rule for training a perceptron. [2]

   (c) Consider the *k*-means algorithm. At the end of the current iteration, suppose we have 3 centroids (0, 1), (2, 1), (-1, 2) corresponding to 3 clusters C_1, C_2, C_3, respectively. In the next iteration, to which cluster(s) will the points (1.5, 3) and (2, 0.6) be assigned? Explain. [2]

   (d) Suppose you have 2 decision trees: *D*1 and *D*2. You trained *D*1 and *D*2 on the same training sets. *D*1 is grown to full depth while *D*2 is post-pruned to reduce the size of the tree. Which of the two trees is likely to have larger bias and which one larger variance? [2]

   (e) Consider the instance space corresponding to all points in the *x*, *y* plane. Let *H* be the set of all circles in the *x*, *y* plane. Points inside the circle are classified as positive examples. Then, is the following statement TRUE or FALSE? Justify your answer. [2]

   "VC-dimension of *H* is at least 3."

2. **(a)** Which of the following 3 statements is/are TRUE for *k*-NN algorithm? Justify your answer. [6]

   (I) For very large *k*, points from classes other than the true class may also get included in the neighborhood.

   (II) For very small *k*, the algorithm is not sensitive to noise.

   (III) It takes a small training time but large inference time.

(b) Consider 3 data points in the 2D space: (-2, -3), (0,0), (2,3). What is the first principal component (write down the actual vector)? [4]

3. **(a)** The following four training instances are provided: [4]
(i) $x_1 = -1, y_1 = 0.03$
(ii) $x_2 = 0, y_2 = 0.80$
(iii) $x_3 = 1, y_3 = 1.90$
(iv) $x_4 = 2, y_4 = 3.05$
Suppose you model this data using the ordinary least squares regression form of

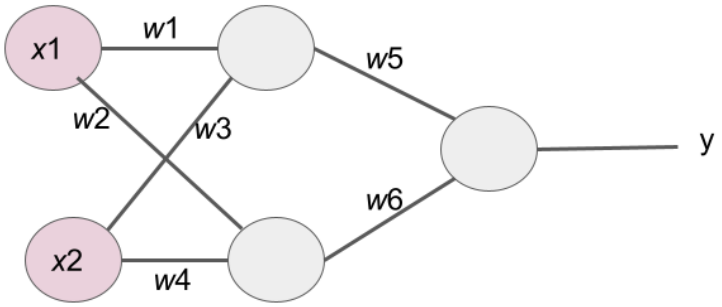
$$y = f(x) = a_0 + a_1 x$$

Which of the following parameters (a_0, a_1) would you use to best model this data? Justify your answer.
(I) (1,2)
(II) (1,1)
(III) (2,1)
(IV) (2,2)

(b) Why do we allow misclassifications in support vector classifier? [3]

(c) Briefly explain the principle behind SVM. [3]

4. Consider a neural network, consisting of two inputs (shown as $x1$, $x2$), a single hidden layer containing two units, and one output unit. The weights are $w1$ through $w6$. The activation function used by the hidden and output units is specified in each sub-question.



(a) Suppose each input is binary-valued. The activation function used by each unit is the threshold function $f(x)$ given by: [4]
$f(x) = 1$ if $x > t$, and
$f(x) = 0$ otherwise.
Assume t is the threshold which can be different for different units and must be set properly. How will you implement the equivalence gate (defined as $Q=A.B+A'.B'$ where A, B are the input boolean variables, A' denotes the complement of A, '+' denotes the OR operator and '.' denotes the AND operator) with this neural network? Explain with a diagram.

(b) Assume each input is a real number. (There is no bias term for the units.) [3]
Consider 2 cases:
Case-I: each unit uses sigmoid activation function, i.e., $f(x) = 1/(1+\exp(-x))$,
Case-II: each unit uses linear activation function, i.e., $f(x) = Cx$ where C is a constant.
Write down the expression for the output y in each case.

(c) Consider **Case-II** from part (b) above. Can you draw an equivalent neural network (with linear activation function) that produces the same output but without any hidden layer? [3]
(You have to express the weights of the new network in terms of the old weights $w1$ through $w6$.)

5. (**a**) In the PAC learning model, the sample complexity for a consistent learner with finite hypothesis space is given by [4]
$m >= (1/\varepsilon) (\ln |H| + \ln (1/\delta))$
where the symbols have their usual significance.
Suppose the concept class C = set of boolean conjunctions based on n boolean variables. Show that the class C is PAC-learnable by a consistent learner with hypothesis space $H = C$.
(Assume we are only concerned with the number of training examples required and not with the computational resources when defining PAC-learnability.)

(b) Why is it unrealistic to expect a classifier to give zero test error even when trained on examples that are very representative of the underlying data distribution? [3]

(**c**) Briefly explain the significance of VC-dimension in computational learning theory. [3]