NANYANG TECHNOLOGICAL UNIVERSITY

SEMESTER 1 EXAMINATION 2022-2023

CE4041/CZ4041 – MACHINE LEARNING

Nov/Dec 2022 Time Allowed: 2 hours

INSTRUCTIONS

1. This paper contains 4 questions and comprises 6 pages.
2. Answer **ALL** questions.
3. This is an **open-book** examination.
4. Keep your answers to **TWO (2)** decimal places whenever needed.
5. All questions carry equal marks.

---

1. (a) State whether each of the following statements is "TRUE" or "FALSE". Each question carries one mark.

(10 marks)

(i) An unlabeled data set is referred to as a set of pairs, each of which consists of an input instance and its corresponding output.

(ii) Suppose A, B and C are binary variables (1 or 0). If $\mathrm{P}(A = 1) = 0.1$, $\mathrm{P}(A = 0, B = 1) = 0.7$ and $\mathrm{P}(A = 0, B = 0, C = 1) = 0.1$, then $\mathrm{P}(C = 1 | A = 0, B = 0) = 0.5$.

(iii) Naïve Bayes Classifiers are a special case of Bayesian Belief Networks.

(iv) Information gain is used to measure node impurity.

(v) According to Occam's Razor, if two decision trees have similar pessimistic errors, the more complex one should be preferred.

(vi) The backpropagation algorithm is used to learn hyper-parameters of a neural network (e.g., number of nodes, number of layers, etc.).

Note: Question No. 1 continues on Page 2

(vii) Support Vector Machines aim to learn a linear hyperplane with the largest margin.

(viii) In the Random Forest algorithm, any classifier can be used as the base classifier.

(ix) In non-parametric density estimation, data instances can only be assumed to follow a Gaussian distribution.

(x) The K-means algorithm is to classify a data instance into one of a set of K classes.

(b) Consider the dataset shown in Table Q1b for a binary classification task, where *Member* and *Level* are input features.

Table Q1b. Training Dataset for Question 1(b).

ID	Member	Level	Class Label
1	Yes	Gold	+1
2	No	Standard	+1
3	No	Gold	−1
4	Yes	Silver	+1
5	Yes	Bronze	+1
6	No	Bronze	+1
7	Yes	Silver	+1
8	No	Silver	−1
9	No	Standard	+1
10	Yes	Gold	−1

(i) Use gain ratio to induce a depth-1 decision tree, i.e., the decision tree only contains a root node and leaf nodes without any other intermediate nodes. Show calculations and the final decision tree. Note: assume multi-way split is adopted when splitting on *Level*.

(9 marks)

(ii) Suppose a penalty term of $k = 0.25$ is defined on each leaf node. Compute the pessimistic errors of the induced decision tree in Question 1(b)(i).

(3 marks)

Note: Question No. 1 continues on Page 3

(c) In binary classification, when majority voting is adopted to make predictions, why is the odd value of K preferred over even values in the K-NN algorithm?

(3 marks)

2. (a) Given the training data set in Table Q2a for a binary classification task, where "Gender" and "Age" are the binary and continuous features, respectively. Use a Naïve Bayes classifier to predict the label of the test instance (M, 8). Note: use the original estimation, not the Laplace estimate or the M-estimate for "Gender". Suppose $\pi = 3.14$, $e = 2.72$.

(8 marks)

Table Q2a. Training Dataset for Question 2(a).

ID	Gender	Age	Class Label
1	M	11	+1
2	F	10	+1
3	M	7	−1
4	F	9	+1
5	M	3	−1

(b) Consider the Bayesian Belief Network shown in Figure Q2b. Compute the following probabilities.

(i) $P(B = \text{Y} \mid C = \text{Y},\ D = \text{N})$

(8 marks)

(ii) $P(A = \text{Y} \mid E = \text{Y}, C = \text{N}, D = \text{N})$

(9 marks)

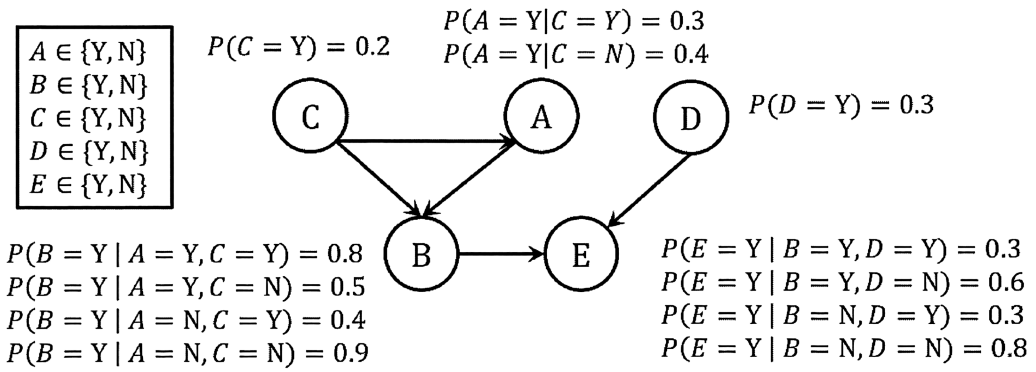

Figure Q2b. The Bayesian Belief Network for Question 2(b).

3. (a) Consider the multi-layer neural network in Figure Q3a. Suppose that the quadratic function $z = \Sigma_{i=1}^{d} w_i X_i^2 - \theta$ and the sign function $a(z) = sign(z)$ are used for both hidden and output nodes. The outputs of the quadratic function and the sign function at each node n_i are denoted as z_i and h_i, respectively. This network uses an error function for each training instance $E = \frac{1}{2}(y_i - \hat{y}_i)^2$.

An initialization of the parameters is shown in Table Q3a, where θ_3, θ_4, and θ_5 are the bias terms for the nodes n_3, n_4, and n_5, respectively. The learning rate λ is set at 0.1. Given a training instance P1 = $[X_1, X_2]$ = [3, 0.6] with the ground truth label of "-1", the backpropagation algorithm is applied after one forward pass of P1. Draw the computational graph for w_{23} and θ_3. Show their update rules in terms of derivative-free expressions and compute their updated values.

(11 marks)

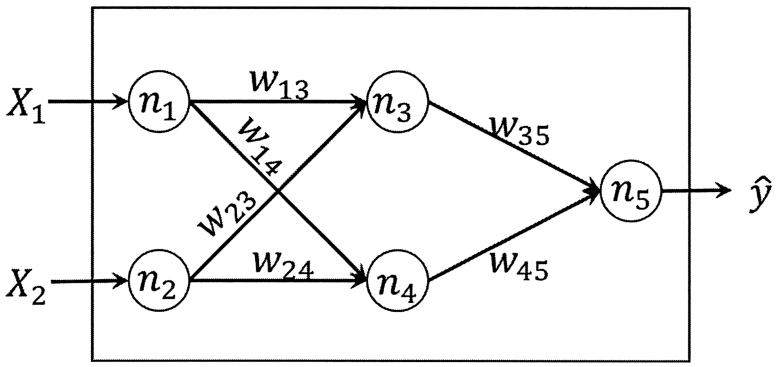

Figure Q3a. The Multi-Layer Neural Network for Question 3(a).

Table Q3a. Parameter Initialization for Question 3(a).

w_{13}	w_{14}	w_{23}	w_{24}	θ_3	θ_4	w_{35}	w_{45}	θ_5
1	0.5	-0.25	1	3.8	5	0.5	0.25	-0.8

(b) In an SVM, the decision boundary is represented by the equation $5x_1 - 2x_2 + b = 0$, where b is an unknown variable. We have two support vectors: (-1, 4) from class 1 and (3, -3) from class 2. Write down the equations for the two parallel hyperplanes (lines in this case). The equations should not have any unknown variable other than x_1 and x_2.

(8 marks)

Note: Question No. 3 continues on Page 5

(c) Which of the two ensemble methods, Bagging or Boosting, would you expect to be more robust to noise in the data (e.g., outliers)? Provide a short justification of your answer in 2-3 sentences.

(6 marks)

4. (a) Consider four 2-dimentional instances A = (0, 1), B = (1, 0), C = (-1, 0), and D = (0, -1). The hierarchical agglomerative clustering algorithm is run on them with the Euclidean distance as the proximity metric and the ties on the proximity broken arbitrarily. The algorithm is stopped when two clusters remain.

(i) Show the initial proximity matrix.

(4 marks)

(ii) Draw the resultant dendrogram with the Complete Link as the inter-class similarity. Show the updated proximity matrix at each step.

(6 marks)

(b) Suppose a dataset D contains n 1-dimensional instances, $x_1, x_2, \ldots, x_n$, each drawn independently from a distribution with the following probability density function:

$$p(x; \sigma) = \frac{1}{2\sigma} e^{-\frac{|x|}{\sigma}}.$$

Find the maximum likelihood estimate for σ.

(8 marks)

(c) You are given a data matrix $\boldsymbol{X}$ shown below, which contains four 2-dimentional instances.

$$\boldsymbol{X} = \begin{bmatrix} 6 & -4 \\ -3 & 5 \\ -2 & 6 \\ 7 & -3 \end{bmatrix}$$

After performing SVD on the mean-centered matrix $\boldsymbol{X_c}$ of $\boldsymbol{X}$ via $\boldsymbol{X_c} = \boldsymbol{VDU}^{\mathrm{T}}$, we obtain the matrices $\boldsymbol{V}$, $\boldsymbol{D}$, and $\boldsymbol{U}$ below.

Note: Question No. 4 continues on Page 6

$$\boldsymbol{V} = \begin{bmatrix} -0.5 & -0.5 \\ 0.5 & -0.5 \\ 0.5 & 0.5 \\ -0.5 & 0.5 \end{bmatrix}; \quad \boldsymbol{D} = \begin{bmatrix} 12.73 & 0 \\ 0 & 1.41 \end{bmatrix}; \quad \boldsymbol{U} = \begin{bmatrix} -0.71 & 0.71 \\ 0.71 & 0.71 \end{bmatrix}$$

PCA is used to project the four instances in $\boldsymbol{X}$ to a one-dimensional space. Compute the percentage of variance preserved by this projection.

(7 marks)

END OF PAPER

CE4041 MACHINE LEARNING
CZ4041 MACHINE LEARNING

Please read the following instructions carefully:

1. **Please do not turn over the question paper until you are told to do so. Disciplinary action may be taken against you if you do so.**

2. You are not allowed to leave the examination hall unless accompanied by an invigilator. You may raise your hand if you need to communicate with the invigilator.

3. Please write your Matriculation Number on the front of the answer book.

4. Please indicate clearly in the answer book (at the appropriate place) if you are continuing the answer to a question elsewhere in the book.