Indian Institute of Technology Kharagpur
Department of Computer Science and Engineering

Machine Learning (CS60050)	**End-Semester Examination**	**Spring Semester, 2022-2023**
Date: 18-Apr-2023 (AN)	**Answer *all* questions.**	**Maximum Marks:** 100

Q1. [Bayesian Networks] **11 marks**

Let us learn some aspects about our life through inference in the following Bayesian Network shown.

$$\boxed{O} \longrightarrow \boxed{H} \longleftarrow \boxed{W} \longrightarrow \boxed{S} \longrightarrow \boxed{F}$$

The variables of our interest are as follows:

O: being Optimistic, **W**: Working hard, **H**: being Happy,
S: founding a Start-up company, **F**: being Famous.

The conditional probability tables for the model are given as:

$$\mathbb{P}(O = true) = 0.5 \qquad \mathbb{P}(W = true) = 0.4$$
$$\mathbb{P}(H = true \mid O = true,\ W = true) = 0.9 \qquad \mathbb{P}(H = true \mid O = true,\ W = false) = 0.7$$
$$\mathbb{P}(H = true \mid O = false,\ W = true) = 0.5 \qquad \mathbb{P}(H = true \mid O = false,\ W = false) = 0.2$$
$$\mathbb{P}(S = true \mid W = true) = 0.6 \qquad \mathbb{P}(S = true \mid W = false) = 0.2$$
$$\mathbb{P}(F = true \mid S = true) = 0.4 \qquad \mathbb{P}(F = true \mid S = false) = 0.1$$

Compute the following probabilities. Show your calculations in details.

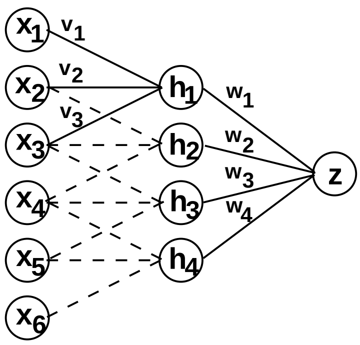
(a) $\mathbb{P}(H = false \mid O = false,\ W = true,\ S = true,\ F = true)$ **(3)**

(b) $\mathbb{P}(H = true \mid S = true,\ F = true)$ **(4)**

(c) $\mathbb{P}(F = true \mid H = true)$ **(4)**

Q2. [Artificial Neural Networks] **5 marks**

Consider the following convolutional neural network architecture.



In the first layer, we have a one-dimensional convolution with a single filter of size 3, such that $h_i = s\left(\sum_{j=1}^{3} v_j . x_{i+j-1}\right)$. The second layer is fully connected, such that $z = \sum_{i=1}^{4} w_i . h_i$. The hidden units' activation function s(x) is the logistic (sigmoid) function of the form $s(x) = \frac{1}{1+e^{-x}}$. The output unit is linear (no activation function). We perform gradient descent on the loss function, $\mathscr{L} = (y - z)^2$, where y is the training label for x.

Compute the following.

(a) What will be the expression for $\frac{\delta \mathscr{L}}{\delta w_i}$? **(2)**

(b) What will be the expression for $\frac{\delta \mathscr{L}}{\delta v_j}$? **(3)**

Instance	True Class	$\mathbb{P}(A,\ldots,Z,\ M_1)$	$\mathbb{P}(A,\ldots,Z,\ M_2)$
1	$+$	0.73	0.61
2	$+$	0.69	0.03
3	$-$	0.44	0.68
4	$-$	0.55	0.31
5	$+$	0.67	0.45
6	$+$	0.47	0.09
7	$-$	0.08	0.38
8	$-$	0.15	0.05
9	$+$	0.45	0.01
10	$-$	0.35	0.04

Q3. [Classifier Evaluation] **5 marks**

You are asked to evaluate the performance of two classification models, M_1 and M_2. The test set you have chosen contains 26 binary attributes, labeled as A through Z.

The above table shows the posterior probabilities obtained by applying the models to the test set. (Only the posterior probabilities for the positive class are shown). As this is a two-class problem, $\mathbb{P}(-) = 1 - \mathbb{P}(+)$ and $P(- \mid A,\ldots,Z) = 1 - \mathbb{P}(+ \mid A,\ldots,Z)$. Assume that, we are mostly interested in detecting instances from the positive class.

For both models, M_1 and M_2, suppose you choose the cutoff threshold to be $t = 0.5$. In other words, any test instances whose posterior probability is greater than t will be classified as a positive example. Compute the precision, recall, and F-measure for both models at this threshold value. **(5)**

Q4. [Computational Learning Theory] **12 marks**

Answer the following questions.

(a) Let the growth function $m_H(N)$ for some hypothesis set, H (N = number of training examples), be $m_H(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$. Determine the Generalization Bound (Ω) for E_{out} with at least 95% probability (confidence) when the number of training examples are 1000. **(2)**

(b) Consider the feature transform $\mathbf{z} = [L_0(x)\ \ L_1(x)\ \ L_2(x)]^T$ with Legendre polynomials and the linear model $h(x) = \mathbf{w}^T.\mathbf{z}$. For the regularized hypothesis with $\mathbf{w} = [-1\ \ +2\ \ -1]^T$, what is $h(x)$ explicitly as a function of x? **(2)**

(d) What the VC-dimension of axis-aligned rectangles in a 2-dimensional plane? Derive / Prove. **(4)**

(e) What is the VC-dimension of axis-aligned squares in a 2-dimensional plane? Derive / Prove. **(4)**

Q5. [Bias and Variance] **6 marks**

For $z \in \mathbb{R}$, you are trying to estimate a true function $g(z) = 2z^2$ with *linear (least-squares) regression*, where the regression function is a line $h(z) = wz$ that goes through the origin and $w \in \mathbb{R}$. Each sample point $x \in \mathbb{R}$ is drawn from the *uniform distribution* on $[-1, 1]$ and has a corresponding label $y = g(x) \in \mathbb{R}$. There is no noise in the labels. We train the model with *just one sample point*! Call it x, and assume $x \neq 0$. We want to apply the bias-variance decomposition to this model.

What is the bias and variance of your model $h(z)$ as a function of a test point $z \in \mathbb{R}$? Your final bias and variance both should not include an x; work out the expectations. **(3+3)**

(*Hint: start by working out the value of the least-squares weight w.*)

Q6. [Unsupervised Learning] **16 marks**

Suppose, six points (P_1, P_2, P_3, P_4, P_5 and P_6) are provided in a 2-dimensional plane. The Euclidean distance between a pair of these points are provided in the table (above).

	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0.00					
P_2	0.12	0.00				
P_3	0.51	0.25	0.00			
P_4	0.84	0.16	0.14	0.00		
P_5	0.28	0.77	0.70	0.45	0.00	
P_6	0.34	0.61	0.93	0.20	0.67	0.00

If you use *Hierarchical Agglomerative Clustering* technique to form the single-link dendrogram, initially each point will form separate clusters, denoted as, $\{P_1\}$, $\{P_2\}$, $\{P_3\}$, $\{P_4\}$, $\{P_5\}$ and $\{P_6\}$. Then, at the first (bottom-most grouping) phase, the algorithm selects $\{P_1\}$ and $\{P_2\}$ clusters to merge and form new cluster $\{P_1, P_2\}$, as the distance considered for grouping here was, $dist(P_1, P_2) = 0.12$ (the minimum among all pairs), for both *single-linkage* and *complete-linkage* variants.

Now, you need to complete the rest of the phases mentioning the next new cluster formed and the distance considered that time for both *single-linkage* and *complete-linkage* variations in the following.

(a) Using ***Single*** *Linkage Hierarchical Agglomerative Clustering* technique to form the single-link dendrogram, complete the remaining phases (missing entries) in the following table. **(4)**

Phase $\rightarrow$	1st	2nd	3rd	4th	5th
New Cluster Formed	$\{P_1, P_2\}$				
Distance Considered	0.12				

Show final result of hierarchical clustering with *single linkage* by drawing a dendrogram. **(1)**

(b) Using ***Complete*** *Linkage Hierarchical Agglomerative Clustering* technique to form the complete-link dendrogram, complete the remaining phases (missing entries) in the following table. **(4)**

Phase $\rightarrow$	1st	2nd	3rd	4th	5th
New Cluster Formed	$\{P_1, P_2\}$				
Distance Considered	0.12				

Show final result of hierarchical clustering with *complete linkage* by drawing a dendrogram. **(1)**

(c) Suppose, for both the above variants (single and complete linkage) of hierarchical clustering, we stop after 4th phase. Compute the average silhouette coefficient (SC) of the overall clustering for both these cases. **(3+3)**

Q7. [Ensemble Learning] **10 marks**

In this problem, we study how boosting algorithm performs on a very simple classification problem. We are given with four training points, P_1, P_2, P_3, P_4, in a 1-dimensional line (x-valued) having their respective values as $x = 1$, $x = 2$, $x = 3$, $x = 4$ and their corresponding 2-class ($+/-$) labels as $-$, $+$, $-$, $+$, respectively.

We shall use decision stumps as our weak learner / hypothesis. Decision stump classifier chooses a constant value c and classifies all points where $x \geq c$ as one class and other points where $x < c$ as the other class. In our given example, let us chose one such decision stump as follows: $x \geq 3$ region is classified as '$+$' zone and $x < 3$ region is classified as '$-$' zone.

Answer the following questions.

(a) What is the initial weight assigned to each data point? **(1)**

(b) How many different decision stumps are possible for the data points given? **(1)**

(c) Which data point(s) will have weights increased after the boosting process as per the decision stump considered in the problem? **(1)**

(d) What will be weights of all the data points after boosting is performed? Show your approach. **(4)**

(e) Indicate whether the following statements are <u>true</u> / <u>false</u>. Give a brief justification.

(i) We cannot perfectly classify all the training examples given in this problem by only applying boosting algorithm (AdaBoost). **(1.5)**

(ii) The training error of boosting classifier (combination of all the weak classifier) monotonically decreases as the number of iterations in the boosting algorithm increases. **(1.5)**

Q8. [Principal Component Analysis] $\boxed{\textbf{5 marks}}$

Given the (x,y)-coordinates of four data points in two-dimensional space: $(4,1)$, $(2,3)$, $(5,4)$ and $(1,0)$, calculate the first principal component. Show your calculations in details. **(5)**

Q9. [Kernel Functions] $\boxed{\textbf{5 marks}}$

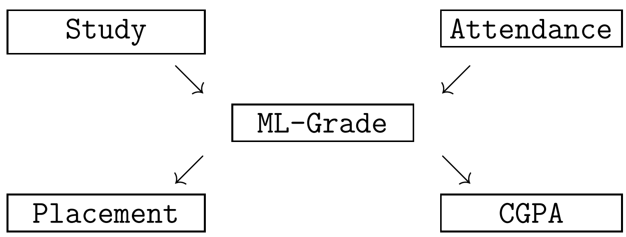
Answer the following.

(a) Let k_1 and k_2 be (valid) kernels; that is, $k_1(\mathbf{x},\mathbf{y}) = \Phi_1(\mathbf{x})^T.\Phi_1(\mathbf{y})$ and $k_2(\mathbf{x},\mathbf{y}) = \Phi_2(\mathbf{x})^T.\Phi_2(\mathbf{y})$. Show that $k = k_1 + k_2$ is a valid kernel by explicitly constructing a corresponding feature mapping $\Phi(\mathbf{z})$. **(2)**

(b) The polynomial kernel is defined to be, $k(\mathbf{x},\mathbf{y}) = (\mathbf{x}^T.\mathbf{y} + c)^d$,
where $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$, and $c \geq 0$. When we take $d = 2$, this kernel is called the *quadratic kernel*. Find the feature mapping $\Phi(\mathbf{z})$ that corresponds to the quadratic kernel. **(3)**

Q10. [Expectation-Maximization Algorithm] $\boxed{\textbf{15 marks}}$

Consider the Bayes Network structure shown below. From the figure below, we abbreviate as follows: **S** = <u>`Study`</u> well, **A** = high <u>`Attendance`</u>, **G** = good <u>`ML-Grade`</u>, **P** = better <u>`Placement`</u>, **C** = high <u>`CGPA`</u>.



We are given the following $K = 8$ training examples as shown below, where only two examples

K	**S**	**A**	**G**	**P**	**C**
$k=1$	1	0	1	1	1
$k=2$	0	1	1	1	0
$k=3$	1	1	1	1	1
$k=4$	0	0	0	0	1
$k=5$	0	0	0	1	0
$k=6$	0	0	0	0	0
$k=7$	1	1	1	?	1
$k=8$	1	1	1	1	?

contain unobserved values (marked with ?), namely, p_7 and c_8. We like to simulate a few steps of the simplified EM algorithm by hand.

Notation: Here, s_k, a_k, g_k, p_k, and c_k indicate the values of **S**, **A**, **G**, **P**, and **C**, respectively, as seen in the k-th example/row. For example, $s_1 = 1$, $a_1 = 0$, $g_1 = 1$, $p_1 = 1$, and $c_1 = 1$.

Answer the following questions:

(a) Given that *all variables are Boolean*, how many basic parameters we need to estimate for the given Bayes Network?

For example, one parameter will be $\theta(g \mid 11)$, which stands for $\mathbb{P}(G=1 \mid S=1, A=1)$. **(2)**

(b) Now, we like to simulate the first E-step of the EM algorithm. Before we start, we initialize all the parameters as 0.5, and then proceed to execute the E-step. What are the following expectation values that will get calculated in this E-step? In particular, calculate the following: **(2+2)**

- $\mathbb{E}(p_7 = 1 \mid s_7, a_7, g_7, c_7, \theta) = ?$
- $\mathbb{E}(c_8 = 1 \mid s_8, a_8, g_8, p_8, \theta) = ?$

(*Note that, only two examples (k=7 and k=8) contains unobserved variables, where $p_7 = ?$, but $s_7 = a_7 = g_7 = c_7 = 1$; and $c_8 = ?$, but $s_8 = a_8 = g_8 = p_8 = 1$, respectively.*)

(c) Now, we like to simulate the first M-step of the EM algorithm. What will be the estimated values of all the model parameters (which you identified in part (a)) that we obtain in this M-step? **(5)**

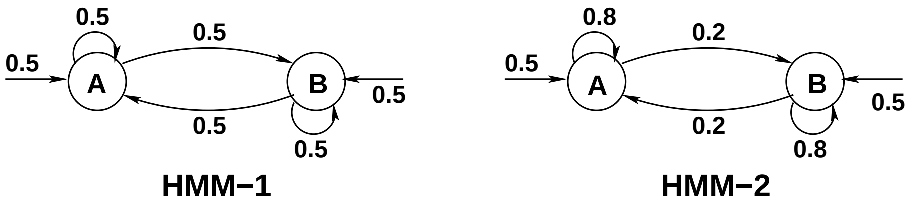
(*Note that, we use the expected count only when the variable is unobserved in an example*)

(d) Last, let us (again) simulate the second E-step of the EM algorithm. What a re the following expectation values that will get calculated in this E-step? In particular, calculate the following: **(2+2)**

- $\mathbb{E}(p_7 = 1 \mid s_7, a_7, g_7, c_7, \theta) = ?$
- $\mathbb{E}(c_8 = 1 \mid s_8, a_8, g_8, p_8, \theta) = ?$

Q11. [Hidden Markov Models] **10 marks**

The following figure above presents two HMMs. States are represented by circles and transitions by directed edges. In both, emissions are deterministic and listed inside the states (either A or B).



Transition probabilities and starting probabilities are listed next to the relevant edges. For example, in HMM-1 we have a probability of 0.5 to start with the state that emits A and a probability of 0.5 to transition to the state that emits B if we are now in the state that emits A.

Notation: In the questions below, $O_{100} = A$ means that the 100-th symbol emitted by an HMM is A.

Answer the following.

(a) Calculate $\mathbb{P}(O_{100} = A,\ O_{101} = A,\ O_{102} = A)$ for HMM-1 and HMM-2, respectively. **(2+2)**

(b) Calculate $\mathbb{P}(O_{100} = A,\ O_{101} = B,\ O_{102} = A,\ O_{103} = B)$ for HMM-1 and HMM-2. respectively. **(2+2)**

(c) Assume you are told that a casino has been using one of the two HMMs to generate streams of letters. You are also told that among the first 1000 letters emitted, 500 are As and 500 are Bs. Can you tell which of the HMMs is being used by this casino? Explain. **(2)**

— END —
