CS5103: Applied Machine Learning
(29/11/2025 10 AM)

Duration: 2.5 hours **Total marks: 50**

Instructions: Please keep all answers to the questions from a part together.

Part - I

1. (a) **(Marks: 3)** Please draw a (5 x 5) grid and plot the points $(2,3),(2,4),(4,2),(4,1)$. We randomly assigned the points into two clusters. Suppose, points $\{(2,3),(4,1)\}$ are assigned to cluster 1 and the remaining two points are assigned to cluster 2. Using Euclidean distance as the measure of dissimilarity, please run the k-means clustering algorithm for three iterations or until convergence whichever is earlier.

   (b) **(Marks: 2)** When should we apply hierarchical clustering? In Figure 1, how many clusters will be generated for the given cut line?

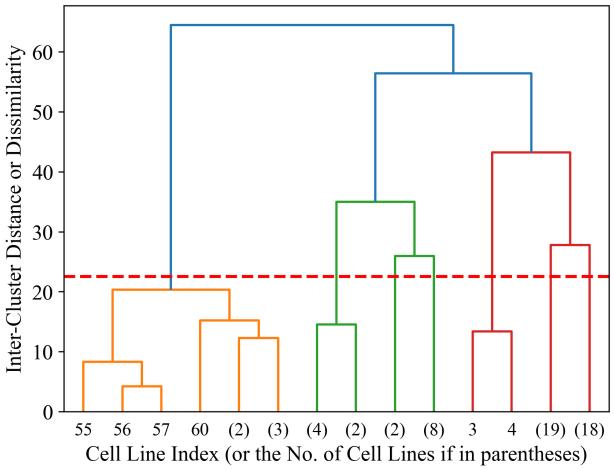


Figure 1: An example of hierarchical clustering

2. (a) **(Marks: 1)** The number (or count) of particles emitted during the decay of radioactive isotopes depends on factors such as time duration, the distance of the particle detector from the source and background radiation. Given the values of these factors, if we wish to predict the number of particles emitted, which regression method would be the most suitable?

   (b) **(Marks: 3)** Why are ridge regression and lasso regression preferred over linear (ordinary least squares a.k.a. OLS) regression? When should we choose lasso regression over ridge regression? When should we choose principal component regression (PCR) or partial least squares (PLS) regression over ridge or lasso regression?

   (c) **(Marks: 5)** Rainfall (in mm) is a non-negative real number that can be predicted using hundreds of meteorological factors along with multiple human-activity related factors. Suppose, we have the daily values of rainfall and 'n' (> 100) of these factors over the last 365 days. Please explain how we can calculate the top two principal components by applying the singular value decomposition (SVD) followed by the principal component analysis (PCA). You may use ellipsis (...) to represent a long series, e.g., 1 2 3 ... n.

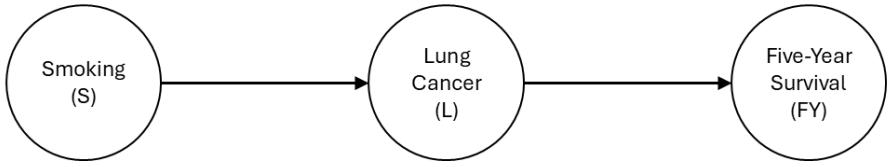
3. (a) **(Marks: 3)** Suppose, we are performing a binary classification task with a decision tree. A particular node contains 200 samples with 100 from each class. There exists a potential split condition

that can separate the samples such a way that the left child will have all 100 samples of the first class and the right child will have all 100 samples of the second class. Please calculate the GINI index of the potential split.

(b) **(Marks: 1)** How does a decision tree classifier find the threshold values (a.k.a. cut points) for a continuous feature at a particular node?

(c) **(Marks: 1)** How is a random forest different from an ensemble of bagged trees?

4. (a) **(Marks: 3)** Lung cancer is the leading cause of cancer-related deaths. The Bayesian network in Figure 2 represents the dependencies between smoking, lung cancer, and five-year survival after being diagnosed with lung cancer. Given the Bayesian network, what is the probability (rounded to 2 decimal places) that a smoker would not survive next five years? (We do not know whether the smoker has lung cancer or not.)



S = True	S = False
0.48	0.52

	L = True	L = False
S = True	0.2	0.8
S = False	0.25	0.75

	FY = True	FY = False
L = True	0.7	0.3
L = False	0.95	0.05

Figure 2: Dependencies between smoking, lung cancer, and five-year survival after being diagnosed with lung cancer. The conditional probability table (CPT) of smoking indicates that 48% of people are smokers. The CPT of lung cancer represents that 20% of smokers develop lung cancer while 25% of non-smokers also develop lung cancer (due to other factors such as genetics). The CPT of five-year survival shows that 30% of lung cancer patients could not survive five years after they are diagnosed with lung cancer. On the other hand, 95% of people who are tested to not have lung cancer survive five years after testing.

(b) **(Marks: 3)** Suppose, we have a noisy 2D binary image I with pixel values either 1 or -1. Let us choose a particular pixel which has the value 1. The pixels on the left and right sides of our chosen pixel have values -1 and 1, respectively. The pixels on top and bottom of our chosen pixel both have values -1. If we denoise our chosen pixel using a Markov random field (MRF) with parameter values $(\eta = 2.0, \zeta = 1.5)$, what will be the value of our chosen pixel after denoising?

Part - II

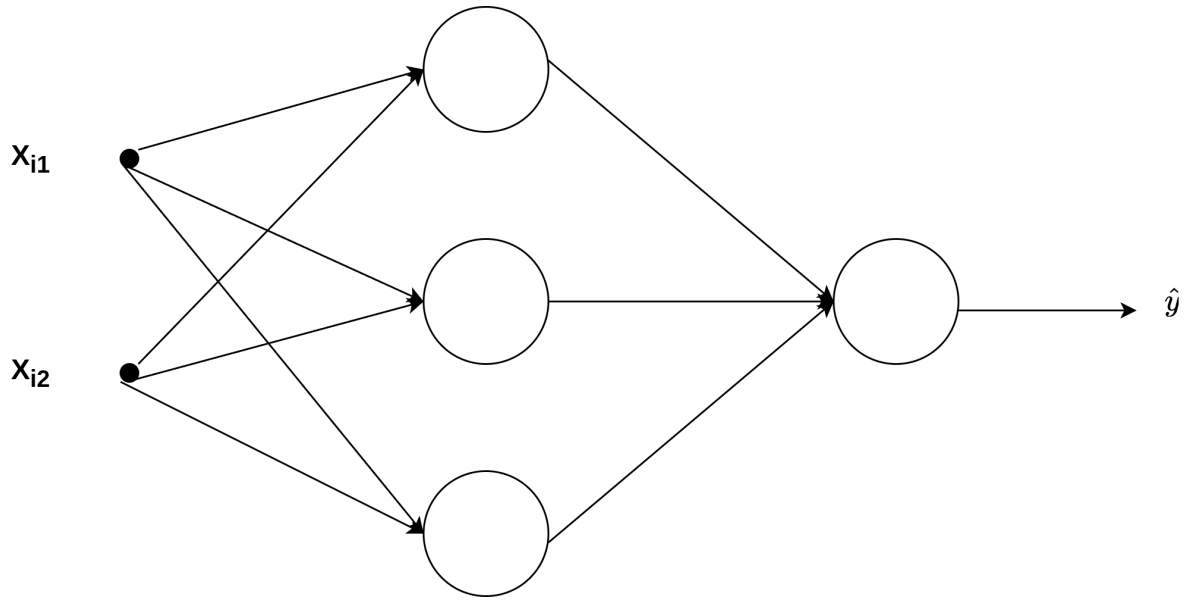


Figure 3: Unlabelled MLP (Question 1), Part II

1. **(Marks: 15)** (a) **(Marks: 4)** Label Figure 3 with all the weights, biases and outputs.
   (b) **(Marks: 7)** For the loss function ($\mathcal{L}$) as $(y - \hat{y})$ and in the absence of an activation function, deduce the partial derivative of the loss function w.r.t w_{11}^2 and w_{11}^1.
   (c) **(Marks: 4)** Recalculate the partial derivative of the loss function w.r.t w_{11}^2 if the *sigmoid* function is used as the activation function.

2. **(Marks: 10)** Table 1 shows a legal situation (case **C**) for which the appropriate Sections of law applicable are sought by a law student. The student has two prior legal situations (P_1 and P_2) with the Sections of law applied. She decides to write high-quality prompts to a Large Language Model (LLM). Help her by writing a suitable `zero-shot` prompt and a suitable `two-shot` prompt (using P_1 and P_2) for the same. The prompts should be detailed and engaging *Role Prompting*.

Case	Situation	Section(s)
C	A digital media company in India published a satirical political video on an OTT platform. An FIR has been filed against them for defamation, promoting enmity, and against the sovereignty of India. The video was labeled as satire and had no calls for violence. What are the specific Indian Penal Code (IPC) and other legal sections likely invoked in the FIR, and what is a key legal defense available to the company?	?
P_1	A comedian, on a TV show, cracked jokes about a religious community.	Indian Penal Code Sections 499 and 153A
P_2	A journalist posted tweets criticizing the government's handling of a situation.	Indian Penal Code Sections 124A and 505

Table 1: **C** is a current case situation, P_i indicates a previous case (Question 2, Part II); **Section(s)** indicates the applicable law.