Birla Institute of Technology & Science, Pilani (Pilani Campus)

CS F320 Foundations of Data Science

Comprehensive Examination

Part-A (Closed book)

Date: 09/12/2023 Weightage: 17.5% Duration: 90 minutes Max. marks: 35

Instructions:

I. Please write all the answers in the Part-A answer sheet only. Do not submit this question paper.

II. For multiple choice questions, you have to write the answers only (not the question) in the answer sheet.

1. Please write only one of the four options for the following questions. If two options are written, then the answer will not be evaluated. There is no negative marking. [12]

i). A discrete time process is called a Markov chain if the current state at time t depends only on the state at time: [1]

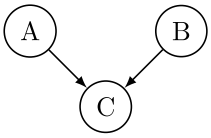
a). $t+1$ b). $2t$ c). $2t+1$ d). $t-1$

ii). In a hidden markov model (HMM), with p_i representing the hidden state, and x_i representing the visible state, the elements in the emission and transition matrix represent ______ and ______ respectively: [1]

a). $P(x_i|x_{i-1})$ and $P(p_i|p_{i-1})$ b). $P(x_i|p_i)$ and $P(p_i|p_{i-1})$

c). $P(p_i|p_{i-1})$ and $P(x_i|x_{i-1})$ d). $P(p_i|x_{i-1})$ and $P(x_i|p_{i-1})$

iii). In the following Bayesian network, the nodes A and B are ______ without C, and conditionally ______ given C: [1]



a). independent, dependent b). dependent, independent

c). independent, independent d). dependent, dependent.

iv). In the structural risk minimization principle, the structural risk is always ______ the expected (or true) risk: [1]

a). lesser than b). equal to

c). greater than or equal to d). lesser than or equal to.

v). In the ACID properties of a database, what does 'C' represent: [1]

a). Concurrency b). Consistency c). Continuity d). Computation

vi). For time-series alignment, the elements of two time-series need not satisfy: [1]

a). Monotonicity b). One to one relationship

c). Boundary conditions d). Continuity.

vii). Which of the following is not a stationary time-series? [1]

a). IID noise b). White noise

c). $X_t = Y$, where Y is a random variable d). Random walk

viii). Which of the following is true with regard to VC dimension for the following classes in $\mathbf{R^2}$: [2]
H_1:{Class of intervals}, H_2:{Class of non-linear classifiers}, H_3:{Class of linear classifiers}, H_4: {Class of axis-aligned rectangles}

a). $VCdim(H_2) > VCdim(H_4) > VCdim(H_3) > VCdim(H_1)$

b). $VCdim(H_2) > VCdim(H_3) > VCdim(H_4) > VCdim(H_1)$

c). $VCdim(H_1) > VCdim(H_2) > VCdim(H_3) > VCdim(H_4)$

d). $VCdim(H_3) > VCdim(H_2) > VCdim(H_1) > VCdim(H_4)$

ix). Which of the following is the correct alignment for the two time-series, $X = \{x_i, i = 1,2,3\} = \{0,2,0\}$, and $Y = \{y_i, i = 1,2,3\} = \{0,0,0.5\}$ using the dynamic time-warping algorithm: [2]

a). $(x_1, y_2), (x_1, y_2), (x_2, y_3)$ b). $(x_1, y_1), (x_1, y_2), (x_2, y_3), (x_3, y_3)$

c). $(x_1, y_1), (x_1, y_2), (x_2, y_2), (x_3, y_3)$ d). $(x_1, y_1), (x_2, y_2), (x_3, y_3)$

x). The minimum cost for the alignment in previous question is: [1]

a). 2 b). 1 c). 1.5 d). 3

2. Mark the following statements **True** or **False**. Please note that there is **negative marking** of **0.5** marks for each wrong answer: [8]

i). The optimal solution for the support vector machine optimization problem is a locally optimal solution.

ii). As the number of training examples increases, the model trained on that data will have higher variance.

iii). Images/videos are examples of structured data.

iv). Maximum aposteriory estimation (MAP) is based on the frequentist approach.

v). A Bayesian network is represented using a directed acyclic graph.

vi). Every IID sequence is white noise, but not vice-versa.

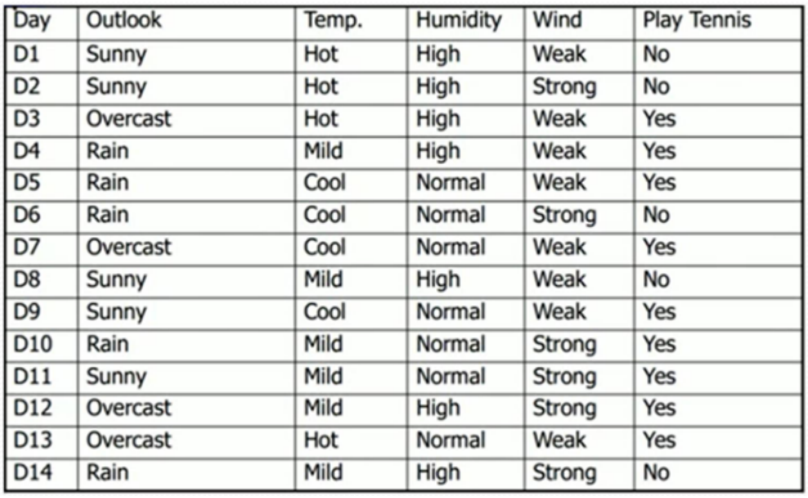
vii). The result of a coin toss is an example of IID noise.

viii). For a regression problem, the coefficient of determination, i.e. $R^2 = 0$ if all the points lie on the fitted curve.

3. a). For the dataset $D = \{(x_i, y_i), i = 1,2,...\} = \{(1,3),(2,1),(4,4)\}$, perform simple linear regression, and obtain the equation for the best fitting line, i.e. $\hat{y}_i = ax_i + b$ by minimizing the total squared error. [3]

b). Under the assumption that the error $E_i \sim \mathcal{N}(0, \sigma^2)$, show that the least squares method for performing simple linear regression is equivalent to finding the maximum likelihood estimate of the parameters (a, b), for the regression function $\hat{y}(x; a, b)$ on a dataset with $(x_i, y_i, i = 1,2,\ldots,n)$ as the values of the predictor and response variables. [3]

4. Calculate the Gini impurity for the node split using 'Outlook' and 'Temperature' as the splitting criteria. Which of these two splitting criteria is more suitable for constructing a decision tree for classification of the label 'Play/No-play'? [3+3+1=7]



Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

5. Give examples of two kernel functions with their mathematical equations. [2]

Birla Institute of Technology & Science, Pilani (Pilani Campus)

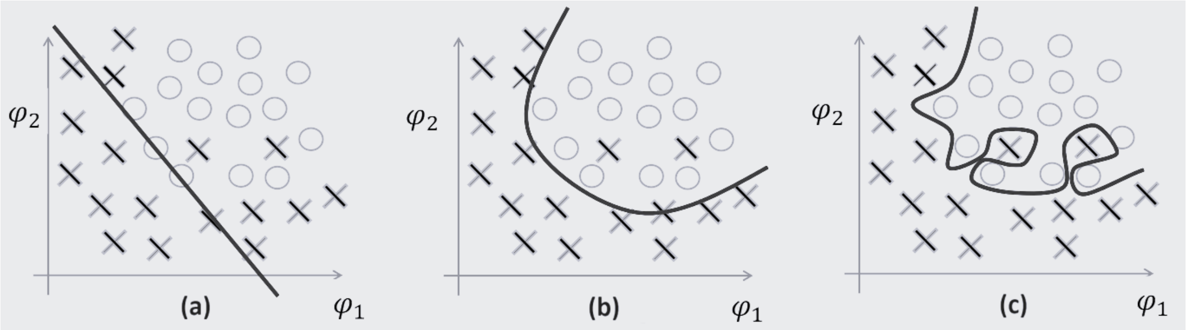
CS F320 Foundations of Data Science

Comprehensive Examination

Part-B (Open book)

Date: 09/12/2023 Weightage: 17.5% Duration: 90 minutes Max. marks: 35

1. Which of the following depicts underfitting, overfitting and optimal fit, and why? [3]



2. Consider the dataset below: [3+1+1=5]

$\mathbf{X_1}$	$\mathbf{X_2}$
6	-4
-3	5
-2	6
7	-3

a). Calculate the principal components of this dataset by first calculating the covariance matrix, and applying the procedure for principal component analysis.

b). What will be the dataset after dimension reduction using PCA to one dimension?

c). Show the dataset and the principal component axes pictorially on a graph.

3. a). Find the maximum and minimum values of the following function [3+3=6]

$$f(x, y) = x + y, \text{ subject to the constraint } x^2 + y^2 = 1.$$

b). A company manufactures two items i_1 and i_2. The total cost of manufacturing depends on these two items as,

$$C = f(i_1, i_2) = 2i_1^2 + i_1 i_2 + i_2^2 + 500$$

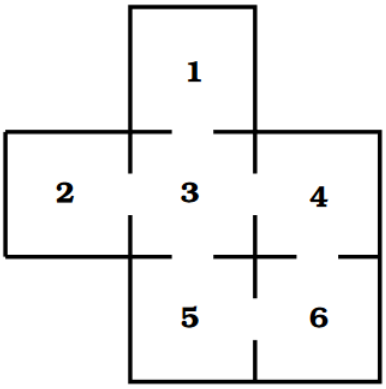
So, how many of these two items should be manufactured by the company to minimize the cost for manufacturing a total of 200 items (including both i_1 and i_2)?

4. Find the singular values of the matrix

$$A = \begin{bmatrix} 7 & 1 \\ 0 & 0 \\ 5 & 5 \end{bmatrix}$$

and find the singular value decomposition of A. [2+4=6]

5. Suppose that you are in the following maze with 6 rooms. You leave the room you are in by choosing a door randomly. The doors are shown with gaps in the image. [3+3=6]



a). What is the transition matrix T for this Markov chain, where the states are the rooms.
b). Find the stationary distribution of this Markov chain.

6. Consider the dataset below [1+3+1+2+2=9]

class	x_1	x_2
+	1	1
+	2	2
+	2	0
−	0	0
−	1	0
−	0	1

a). Are the classes in this data linearly separable?
b). Plot the dataset, and using a graphical approach obtain the equations for the supporting hyperplanes, and the classifying hyperplane for linear SVM? Plot the hyperplanes on a graph.
c). What is the weight vector **w** here? Also, show it on the graph.
d). What are the support vectors here? Also, mark them on the graph.
e). How will the size of the margin change if we remove any one of the support vectors in this dataset? Show the scenario after removal of each of the support vectors in this dataset on a graph. Give explaination for the change in margin after removal of each support vector.

* * * * * * * * *