CODE CST 284	Mathematics for Machine Learning	CATEGORY	L	T	P	CREDIT
		MINOR	3	1	0	4

Preamble: This is a foundational course for awarding B. Tech. Minor in Computer Science and Engineering with specialization in ***Machine Learning***. The purpose of this course is to introduce mathematical foundations of basic Machine Learning concepts among learners, on which Machine Learning systems are built. This course covers Linear Algebra, Vector Calculus, Probability and Distributions, Optimization and Machine Learning problems. Concepts in this course help the learners to understand the mathematical principles in Machine Learning and aid in the creation of new Machine Learning solutions, understand & debug existing ones, and learn about the inherent assumptions & limitations of the current methodologies.

Prerequisite:

1. A sound background in higher secondary school Mathematics.
2. Python for Machine Learning (CST 253)

Course Outcomes: After the completion of the course the student will be able to

CO 1	Make use of the concepts, rules and results about linear equations, matrix algebra, vector spaces, eigenvalues & eigenvectors and orthogonality & diagonalization to solve computational problems (Cognitive Knowledge Level: **Apply**)
CO 2	Perform calculus operations on functions of several variables and matrices, including partial derivatives and gradients (Cognitive Knowledge Level: **Apply**)
CO 3	Utilize the concepts, rules and results about probability, random variables, additive & multiplicative rules, conditional probability, probability distributions and Bayes' theorem to find solutions of computational problems (Cognitive Knowledge Level: **Apply**)
CO 4	Train Machine Learning Models using unconstrained and constrained optimization methods (Cognitive Knowledge Level: **Apply**)
CO 5	Illustrate how the mathematical objects - linear algebra, probability, and calculus can be used to design machine learning algorithms (Cognitive Knowledge Level: **Understand**)

Mapping of course outcomes with program outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12
CO 1	√	√	√	√								√
CO 2	√	√	√									√
CO 3	√	√	√	√								√
CO 4	√	√	√	√		√						√
CO 5	√	√	√	√	√	√				√		√

Abstract POs defined by National Board of Accreditation			
PO#	**Broad PO**	**PO#**	**Broad PO**
PO1	Engineering Knowledge	**PO7**	Environment and Sustainability
PO2	Problem Analysis	**PO8**	Ethics
PO3	Design/Development of solutions	**PO9**	Individual and team work
PO4	Conduct investigations of complex problems	**PO10**	Communication
PO5	Modern tool usage	**PO11**	Project Management and Finance
PO6	The Engineer and Society	**PO12**	Life long learning

Assessment Pattern

Bloom's Category	**Continuous Assessment Tests**		**End Semester Examination**
	1	**2**	
Remember	20%	20%	20%
Understand	40%	40%	40%
Apply	40%	40%	40%
Analyse			
Evaluate			
Create			

Mark Distribution

Total Marks	**CIE Marks**	**ESE Marks**	**ESE Duration**
150	**50**	**100**	**3 hours**

Continuous Internal Evaluation Pattern:

Attendance	: 10 marks
Continuous Assessment Tests	: 25 marks
Continuous Assessment Assignment	: 15 marks

Internal Examination Pattern:

Each of the two internal examinations has to be conducted out of 50 marks

First Internal Examination shall be preferably conducted after completing the first half of the syllabus and the Second Internal Examination shall be preferably conducted after completing remaining part of the syllabus.

There will be two parts: Part A and Part B. Part A contains 5 questions (preferably, 2 questions each from the completed modules and 1 question from the partly covered module), having 3 marks for each question adding up to 15 marks for part A. Students should answer all questions from Part A. Part B contains 7 questions (preferably, 3 questions each from the completed modules and 1 question from the partly covered module), each with 7 marks. Out of the 7 questions in Part B, a student should answer any 5.

End Semester Examination Pattern: There will be two parts; Part A and Part B. Part A contains 10 questions with 2 questions from each module, having 3 marks for each question. Students should answer all questions. Part B contains 2 questions from each module of which student should answer anyone. Each question can have maximum 2 sub-divisions and carries 14 marks.

Syllabus

Module 1

LINEAR ALGEBRA : Systems of Linear Equations – Matrices, Solving Systems of Linear Equations. Vector Spaces - Linear Independence, Basis and Rank, Linear Mappings, Norms, - Inner Products - Lengths and Distances - Angles and Orthogonality - Orthonormal Basis - Orthogonal Complement - Orthogonal Projections. Matrix Decompositions - Determinant and Trace, Eigenvalues and Eigenvectors, Cholesky Decomposition, Eigen decomposition and Diagonalization, Singular Value Decomposition, Matrix Approximation.

Module 2

VECTOR CALCULUS : Differentiation of Univariate Functions - Partial Differentiation and Gradients, Gradients of Vector Valued Functions, Gradients of Matrices, Useful Identities for Computing Gradients. Back propagation and Automatic Differentiation - Higher Order Derivatives- Linearization and Multivariate Taylor Series.

Module 3

Probability and Distributions : Construction of a Probability Space - Discrete and Continuous Probabilities, Sum Rule, Product Rule, and Bayes' Theorem. Summary Statistics and Independence – Important Probability distributions - Conjugacy and the Exponential Family - Change of Variables/Inverse Transform.

Module 4

Optimization : Optimization Using Gradient Descent - Gradient Descent With Momentum, Stochastic Gradient Descent. Constrained Optimization and Lagrange Multipliers - Convex Optimization - Linear Programming - Quadratic Programming.

Module 5

CENTRAL MACHINE LEARNING PROBLEMS : Data and Learning Model- Empirical Risk Minimization - Parameter Estimation - Directed Graphical Models.

Linear Regression - Bayesian Linear Regression - Maximum Likelihood as Orthogonal Projection.

Dimensionality Reduction with Principal Component Analysis - Maximum Variance Perspective, Projection Perspective. Eigenvector Computation and Low Rank Approximations.

Density Estimation with Gaussian Mixture Models - Gaussian Mixture Model, Parameter Learning via Maximum Likelihood, EM Algorithm.

Classification with Support Vector Machines - Separating Hyperplanes, Primal Support Vector Machine, Dual Support Vector Machine, Kernels.

Text book:

1. Mathematics for Machine Learning by Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong published by Cambridge University Press (freely available at https:// mml - book.github.io)

Reference books:

1. Linear Algebra and Its Applications, 4th Edition by Gilbert Strang
2. Linear Algebra Done Right by Axler, Sheldon, 2015 published by Springer
3. Introduction to Applied Linear Algebra by Stephen Boyd and Lieven Vandenberghe, 2018 published by Cambridge University Press
4. Convex Optimization by Stephen Boyd and Lieven Vandenberghe, 2004 published by Cambridge University Press
5. Pattern Recognition and Machine Learning by Christopher M Bishop, 2006, published by Springer
6. Learning with Kernels – Support Vector Machines, Regularization, Optimization, and Beyond by Bernhard Scholkopf and Smola, Alexander J Smola, 2002, bublished by MIT Press
7. Information Theory, Inference, and Learning Algorithms by David J. C MacKay, 2003 published by Cambridge University Press
8. Machine Learning: A Probabilistic Perspective by Kevin P Murphy, 2012 published by MIT Press.
9. The Nature of Statistical Learning Theory by Vladimir N Vapnik, 2000, published by Springer

Sample Course Level Assessment Questions.

Course Outcome 1 (CO1):

1. Find the set ***S*** of all solutions in ***x*** of the following inhomogeneous linear systems ***Ax* = *b***, where ***A*** and ***b*** are defined as follows:

$$\boldsymbol{A} = \begin{bmatrix} 1 & -1 & 0 & 0 & 1 \\ 1 & 1 & 0 & -3 & 0 \\ 2 & -1 & 0 & 1 & -1 \\ -1 & 2 & 0 & -2 & -1 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 3 \\ 6 \\ 5 \\ -1 \end{bmatrix}$$

2. Determine the inverses of the following matrix if possible

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

3. Are the following sets of vectors linearly independent?

$$\boldsymbol{x}_1 = \begin{bmatrix} 2 \\ -1 \\ 3 \end{bmatrix}, \quad \boldsymbol{x}_2 = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}, \quad \boldsymbol{x}_3 = \begin{bmatrix} 3 \\ -3 \\ 8 \end{bmatrix}$$

4. A set of ***n*** linearly independent vectors in $\boldsymbol{R}^n$ forms a basis. Does the set of vectors **(2, 4,−3)** , **(0, 1, 1)** , **(0, 1,−1**) form a basis for $\boldsymbol{R}^3$? Explain your reasons.
5. Consider the transformation ***T (x, y) = (x + y, x + 2y, 2x + 3y)***. Obtain ***ker T*** and use this to calculate the nullity. Also find the transformation matrix for ***T***.

6. Find the characteristic equation, eigenvalues, and eigenspaces corresponding to each eigenvalue of the following matrix

$$\begin{bmatrix} 2 & 0 & 4 \\ 0 & 3 & 0 \\ 0 & 1 & 2 \end{bmatrix}$$

7. Diagonalize the following matrix, if possible

$$\begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & 3 \end{bmatrix}$$

8. Find the singular value decomposition (SVD) of the following matrix

$$\begin{bmatrix} 0 & 1 & 1 \\ \sqrt{2} & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

Course Outcome 2 (CO2):

1. For a scalar function $f(x, y, z) = x^2 + 3y^2 + 2z^2$, find the gradient and its magnitude at the point **(1, 2, -1)**.
2. Find the maximum and minimum values of the function $f(x, y) = 4x + 4y - x^2 - y^2$ subject to the condition $x^2 + y^2 <= 2$.
3. Suppose you were trying to minimize $f(x, y) = x^2 + 2y + 2y^2$. Along what vector should you travel from (5, 12)?
4. Find the second order Taylor series expansion for $f(x, y) = (x + y)^2$ about **(0 , 0)**.
5. Find the critical points of $f(x, y) = x^2 - 3xy + 5x - 2y + 6y^2 + 8$.
6. Compute the gradient of the Rectified Linear Unit (ReLU) function ***ReLU(z) = max(0 , z)***.
7. Let $L = ||Ax - b||^2_2$, where A is a matrix and x and b are vectors. Derive ***dL*** in terms of ***dx***.

Course Outcome 3 (CO3):

1. Let ***J*** and ***T*** be independent events, where ***P(J)=0.4*** and ***P(T)=0.7***.

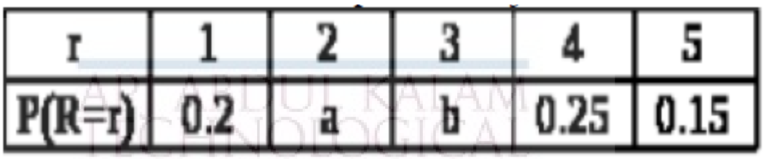
   i. Find ***P(J∩T)***

   ii. Find ***P(J∪T)***

   iii. Find ***P(J∩T')***

2. Let ***A*** and ***B*** be events such that ***P(A)=0.45*** , ***P(B)=0.35*** and ***P(A∪B)=0.5***. Find ***P(A∣B)***.

3. A random variable **R** has the probability distribution as shown in the following table:



r	1	2	3	4	5
P(R=r)	0.2	a	b	0.25	0.15

   i. Given that ***E(R)=2.85***, find ***a*** and ***b***.

   ii. Find ***P(R>2)***.

4. A biased coin (with probability of obtaining a head equal to ***p*** > ***0***) is tossed repeatedly and independently until the first head is observed. Compute the probability that the first head appears at an even numbered toss.

5. Two players A and B are competing at a trivia quiz game involving a series of questions. On any individual question, the probabilities that A and B give the correct answer are ***p*** and ***q*** respectively, for all questions, with outcomes for different questions being independent. The game finishes when a player wins by answering a question correctly. Compute the probability that A wins if

   i. A answers the first question,

   ii. B answers the first question.

6. A coin for which ***P(heads) = p*** is tossed until two successive tails are obtained. Find the probability that the experiment is completed on the ***n***th toss.

7. You roll a fair dice twice. Let the random variable ***X*** be the product of the outcomes of the two rolls. What is the probability mass function of ***X***? What are the expected value and the standard deviation of ***X***?

8. While watching a game of Cricket, you observe someone who is clearly supporting Mumbai Indians. What is the probability that they were actually born within 25KM of Mumbai? Assume that:

   - the probability that a randomly selected person is born within 25KM of Mumbai is 1/20;
   - the chance that a person born within 25KMs of Mumbai actually supports MI is 7/10 ;
   - the probability that a person not born within 25KM of Mumbai supports MI with probability 1/10.

9. What is an exponential family? Why are exponential families useful?
10. Let Z_1 and Z_2 be independent random variables each having the standard normal distribution. Define the random variables X and Y by $X = Z_1 + 3Z_2$ and Y $= Z_1 + Z_2$. Argue that the joint distribution of (X, Y) is a bivariate normal distribution. What are the parameters of this distribution?
11. Given a continuous random variable x, with cumulative distribution function $F_x(x)$, show that the random variable $y = F_x(x)$ is uniformly distributed.
12. Explain Normal distribution, Binomial distribution and Poisson distribution in the exponential family form.

Course Outcome 4(CO4):

1. Find the extrema of $f(x, y) = x$ subject to $g(x, y) = x^2 + 2y^2 = 3$.
2. Maximize the function $f(x, y, z) = xy + yz + xz$ on the unit sphere $g(x, y, z) = x^2 + y^2 + z^2 = 1$.
3. Provide necessary and suffcient conditions under which a quadratic optimization problem be written as a linear least squares problem.
4. Consider the univariate function $f(x) = x^3 + 6x^2 - 3x - 5$. Find its stationary points and indicate whether they are maximum, minimum, or saddle points.
5. Consider the update equation for stochastic gradient descent. Write down the update when we use a mini-batch size of one.

6. Consider the function

$$f(x) = (x_1 - x_2)^2 + \frac{1}{1 + x_1^2 + x_2^2}.$$

   i. Is ***f(x)*** a convex function? Justify your answer.

   ii. Is (1 , -1) a local/global minimum? Justify your answer.

7. Is the function $\boldsymbol{f(x, y) = 2x^2 + y^2 + 6xy - x + 3y - 7}$ convex, concave, or neither? Justify your answer.

8. Consider the following convex optimization problem

$$\text{minimize } \frac{x^2}{2} + x + 4y^2 - 2y$$

   Subject to the constraint $\boldsymbol{x + y >= 4, x, y >= 1}$.

   Derive an explicit form of the Lagrangian dual problem.

9. Solve the following LP problem with the simplex method.

$$max\ \ 5x_1 + 6x_2 + 9x_3 + 8x_4$$

   subject to the constraints

$$\begin{array}{ccccccccc} x_1 & + & 2x_2 & + & 3x_3 & + & x_4 & \leq & 5 \\ x_1 & + & x_2 & + & 2x_3 & + & 3x_4 & \leq & 3 \end{array}$$

$$x_1, x_2, x_3, x_4 \geq 0$$

Course Outcome 5 (CO5):

1. What is a loss function? Give examples.
2. What are training/validation/test sets? What is cross-validation? Name one or two examples of cross-validation methods.
3. Explain generalization, overfitting, model selection, kernel trick, Bayesian learning

4. Distinguish between Maximum Likelihood Estimation (MLE) and Maximum A Posteriori Estimation (MAP)?

5. What is the link between structural risk minimization and regularization?

6. What is a kernel? What is a dot product? Give examples of kernels that are valid dot products.

7. What is ridge regression? How can one train a ridge regression linear model?

8. What is Principal Component Analysis (PCA)? Which eigen value indicates the direction of largest variance? In what sense is the representation obtained from a projection onto the eigen directions corresponding the the largest eigen values optimal for data reconstruction?

9. Suppose that you have a linear support vector machine (SVM) binary classifier. Consider a point that is currently classified correctly, and is far away from the decision boundary. If you remove the point from the training set, and re-train the classifier, will the decision boundary change or stay the same? Explain your answer in one sentence.

10. Suppose you have ***n*** independent and identically distributed (i.i.d) sample data points $x_1, \ldots , x_n$. These data points come from a distribution where the probability of a given datapoint ***x*** is

$$P(x) = \frac{1}{\theta} e^{-\frac{1}{\theta} x}.$$

    Prove that the MLE estimate of parameter is the sample mean.

11. Suppose the data set $y_1,\ldots,y_n$ is a drawn from a random sample consisting of i.i.d. discrete uniform distributions with range 1 to N. Find the maximum likelihood estimate of N.

12. Ram has two coins: one fair coin and one biased coin which lands heads with probability 3/4. He picks one coin at random (50-50) and flips it repeatedly until he gets a tails. Given that he observes 3 heads before the first tails, find the posterior probability that he picked each coin.

    i. What are the prior and posterior odds for the fair coin?

    ii. What are the prior and posterior predictive probabilities of heads on the next flip? Here prior predictive means prior to considering the data of the first four flips.

Model Question paper

QP Code : **Total Pages: 4**

Reg No.:_______________ Name:__________________________

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

IV SEMESTER B.TECH (MINOR) DEGREE EXAMINATION, MONTH and YEAR

Course Code: CST 284

Course Name: MATHEMATICS FOR MACHINE LEARNING

Max. Marks: 100 Duration: 3 Hours

PART A

Answer all questions, each carries 3 marks. Marks

1 Show that with the usual operation of scalar multiplication but with addition on reals given by $x \# y = 2(x + y)$ is not a vector space.

2 Find the eigenvalues of the following matrix in terms of k. Can you find an eigenvector corresponding to each of the eigenvalues?

$$\begin{bmatrix} 1 & k \\ 2 & 1 \end{bmatrix}$$

3 Let $f(x, y, z) = xye^r$, where $r = x^2+z^2-5$. Calculate the gradient of f at the point **(1, 3, -2)**.

4 Compute the Taylor polynomials T_n, $n = 0 , \dots , 5$ of $f(x) = \sin(x) + \cos(x)$ at $x_0 = 0$.

5 Let X be a continuous random variable with probability density function on $0 <= x <= 1$ defined by $f(x) = 3x^2$. Find the pdf of $Y = X^2$.

6 Show that if two events A and B are independent, then A and B' are independent.

7 Explain the principle of the gradient descent algorithm.

8 Briey explain the difference between (batch) gradient descent and stochastic gradient descent. Give an example of when you might prefer one over the other.

9 What is the empirical risk? What is “empirical risk minimization”?

10 Explain the concept of a Kernel function in Support Vector Machines. Why are kernels so useful? What properties a kernel should posses to be used in an SVM?

PART B

Answer any one Question from each module. Each question carries 14 Marks

11 a) i. Find all solutions to the system of linear equations (6)

$$\begin{aligned} -4x + 5z &= -2 \\ -3x - 3y + 5z &= 3 \\ -x + 2y + 2z &= -1 \end{aligned}$$

ii. Prove that all vectors orthogonal to **[2, −3, 1]**^{T} forms a subspace ***W*** of ***R***^{3}. What is ***dim (W)*** and why?

b) Use the Gramm-Schmidt process to find an orthogonal basis for the column space of the following matrix (8)

$$\begin{bmatrix} 2 & 1 & 0 \\ 1 & -1 & 1 \\ 0 & 3 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

OR

12 a) i. Let $\boldsymbol{L}$ be the line through the origin in $\boldsymbol{R^2}$ that is parallel to the vector (6)

[3, 4]^T. Find the standard matrix of the orthogonal projection onto L. Also find the point on $\boldsymbol{L}$ which is closest to the point **(7 , 1)** and find the point on $\boldsymbol{L}$ which is closest to the point ***(-3 , 5)***.

ii. Find the rank-1 approximation of

$$\begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix}$$

b) i. Find an orthonormal basis of $\boldsymbol{R^3}$ consisting of eigenvectors for the following matrix (8)

$$\begin{bmatrix} 1 & 0 & -2 \\ 0 & 5 & 0 \\ -2 & 0 & 4 \end{bmatrix}$$

ii. Find a 3 × 3 orthogonal matrix $\boldsymbol{S}$ and a 3 × 3 diagonal matrix $\boldsymbol{D}$ such that $\boldsymbol{A = SDS^T}$.

13 a) A skier is on a mountain with equation $z = 100 - 0.4x^2 - 0.3y^2$, where z denotes height. (8)

i. The skier is located at the point with xy-coordinates **(1 , 1)**, and wants to ski downhill along the steepest possible path. In which direction (indicated by a vector **(a , b**) in the xy-plane) should the skier begin skiing.

ii. The skier begins skiing in the direction given by the xy-vector (**a , b)** you found in part (i), so the skier heads in a direction in space given by the vector **(a , b , c**). Find the value of **c**.

b) Find the linear approximation to the function ***f(x,y) = 2 - sin(-x - 3y)*** at the point **(0 , π)**, and then use your answer to estimate ***f(0.001 , π)***. (6)

OR

14 a) Let ***g*** be the function given by (8)

$$g(x,y) = \begin{cases} \dfrac{x^2y}{x^2+y^2} & \text{if } (x,y) \neq (0,0); \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

i. Calculate the partial derivatives of ***g*** at **(0 , 0)**.

ii. Show that ***g*** is not differentiable at **(0 , 0)**.

b) Find the second order Taylor series expansion for $f(x,y) = e^{-(x2+y2)} \cos(xy)$ about **(0 , 0)**. (6)

15 a) There are two bags. The first bag contains four mangos and two apples; the second bag contains four mangos and four apples. We also have a biased coin, which shows "heads" with probability 0.6 and "tails" with probability 0.4. If the coin shows "heads". we pick a fruit at random from bag 1; otherwise we pick a fruit at random from bag 2. Your friend flips the coin (you cannot see the result), picks a fruit at random from the corresponding bag, and presents you a mango. (6)

What is the probability that the mango was picked from bag 2?

b) Suppose that one has written a computer program that sometimes compiles and sometimes not (code does not change). You decide to model the apparent stochasticity (success vs. no success) ***x*** of the compiler using a Bernoulli distribution with parameter μ: (8)

$$p(x \mid \mu) = \mu^x (1-\mu)^{1-x}, \quad x \in \{0,1\}$$

Choose a conjugate prior for the Bernoulli likelihood and compute the posterior distribution $p(\mu \mid x_1, \ldots, x_N)$.

OR

16 a) Consider a mixture of two Gaussian distributions (8)

$$0.4\mathcal{N}\left(\begin{bmatrix}10\\2\end{bmatrix},\begin{bmatrix}1&0\\0&1\end{bmatrix}\right)+0.6\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}8.4&2.0\\2.0&1.7\end{bmatrix}\right)$$

i. Compute the marginal distributions for each dimension.

ii. Compute the mean, mode and median for each marginal distribution.

iii. Compute the mean and mode for the two-dimensional distribution.

b) Express the Binomial distribution as an exponential family distribution. Also express the Beta distribution is an exponential family distribution. Show that the product of the Beta and the Binomial distribution is also a member of the exponential family. (6)

17 a) Find the extrema of ***f(x,y,z) = x - y + z*** subject to ***g(x,y,z) = x² + y² + z² = 2***. (8)

b) Let

$$P=\begin{bmatrix}13&12&-2\\12&17&6\\-2&6&12\end{bmatrix},\ q=\begin{bmatrix}-22.0\\-14.5\\13.0\end{bmatrix},\ \text{and}\ r=1.$$

Show that ***x**** = **(1 , 1/2 , -1)** is optimal for the optimization problem

$$\begin{array}{ll}\min & \frac{1}{2}x^{\mathsf{T}}Px+q^{\mathsf{T}}x+r\\ \text{s.t.} & -1\le x_i\le 1,\ i=1,2,3.\end{array}$$

(6)

OR

18 a) Derive the gradient descent training rule assuming that the target function is represented as $o_d = w_0 + w_1x_1 + \ldots + w_nx_n$. Define explicitly the cost/ error function ***E***, assuming that a set of training examples ***D*** is provided, where each training example $d \in D$ is associated with the target output t_d. (8)

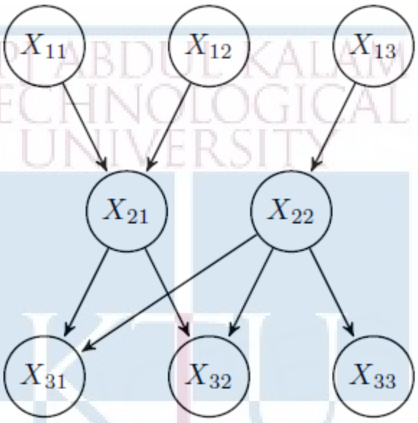
b) Find the maximum value of ***f(x,y,z)* = *xyz*** given that ***g(x,y,z)* = *x* + *y* + *z* = *3*** and ***x,y,z* >= *0***. (6)

19 a) Consider the following probability distribution (7)

$$P_\theta(x) = 2\theta x e^{-\theta x^2}$$

where $\boldsymbol{\theta}$ is a parameter and $\boldsymbol{x}$ is a positive real number. Suppose you get $\boldsymbol{m}$ i.i.d. samples $\boldsymbol{x_i}$ drawn from this distribution. Compute the maximum likelihood estimator for $\boldsymbol{\theta}$ based on these samples.

b) Consider the following Bayesian network with boolean variables. (7)



i. List variable(s) conditionally independent of X_{33} given $\boldsymbol{X_{11}}$ and $\boldsymbol{X_{12}}$

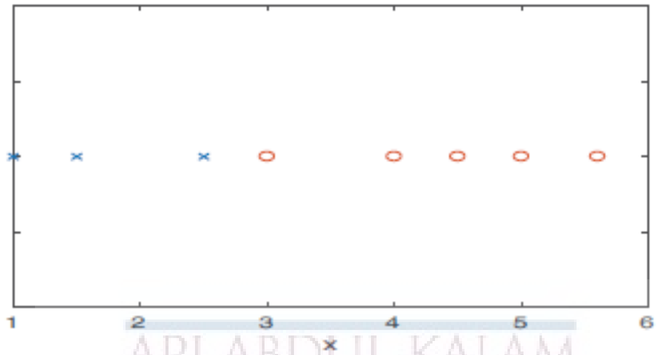
ii. List variable(s) conditionally independent of X_{33} and $\boldsymbol{X_{22}}$

iii. Write the joint probability $\boldsymbol{P(X_{11}, X_{12}, X_{13}, X_{21}, X_{22}, X_{31}, X_{32}, X_{33})}$ factored according to the Bayes net. How many parameters are necessary to define the conditional probability distributions for this Bayesian network?

iv. Write an expression for $\boldsymbol{P(X_{13} = 0, X_{22} = 1, X_{33} = 0)}$ in terms of the conditional probability distributions given in your answer to part (iii). Justify your answer.

OR

20 a) Consider the following one dimensional training data set, ’x’ denotes negative examples and ’o’ positive examples. The exact data points and their labels are given in the table below. Suppose a SVM is used to classify this data. (6)



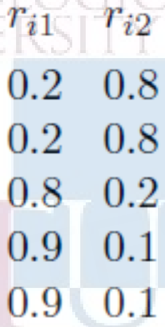
x	1	1.5	2.5	3	4	4.5	5	5.6
y	-1	-1	-1	1	1	1	1	1

i. Indicate which are the support vectors and mark the decision boundary.

ii. Give the value of the cost function and the model parameter after training.

b) Suppose that we are fitting a Gaussian mixture model for data items consisting of a single real value, x, using $K = 2$ components. We have $N = 5$ training cases, in which the values of x are as **5, 15, 25, 30, 40**. Using the EM algorithm to find the maximum likeihood estimates for the model parameters, what are the mixing proportions for the two components, π_1 and π_2, and the means for the two components, μ_1 and μ_2. The standard deviations for the two components are fixed at 10. (8)

Suppose that at some point in the EM algorithm, the **E** step found that the responsibilities of the two components for the five data items were as follows:



r_{i1}	r_{i2}
0.2	0.8
0.2	0.8
0.8	0.2
0.9	0.1
0.9	0.1

What values for the parameters π_1, π_2 , μ_1, and μ_2 will be found in the next **M** step of the algorithm?

****

Teaching Plan

No	Topic	No. of Lectures (45)
1	**Module-I (LINEAR ALGEBRA)**	**8**
1.	Systems of Linear Equations – Matrices, Solving Systems of Linear Equations. Vector Spaces - Linear Independence.	1
2.	Vector Spaces - Basis and Rank	1
3.	Linear Mappings	1
4.	Norms, Inner Products, Lengths and Distances, Angles and Orthogonality, Orthonormal Basis, Orthogonal Complement	1
5.	Orthogonal Projections, Matrix Decompositions, Determinant and Trace.	1
6.	Eigenvalues and Eigenvectors	1
7.	Cholesky Decomposition, Eigen decomposition and Diagonalization	1
8.	Singular Value Decomposition - Matrix Approximation	1
	Module-II (VECTOR CALCULUS)	**6**
1	Differentiation of Univariate Functions, Partial Differentiation and Gradients	1
2	Gradients of Vector Valued Functions, Gradients of Matrices	1
3	Useful Identities for Computing Gradients	1
4	Backpropagation and Automatic Differentiation	1
5	Higher Order Derivatives	1
6	Linearization and Multivariate Taylor Series	1
3	**Module-III (Probability and Distributions)**	**10**
1	Construction of a Probability Space - Discrete and Continuous Probabilities (Lecture 1)	1

2	Construction of a Probability Space - Discrete and Continuous Probabilities (Lecture 2)	1
3	Sum Rule, Product Rule	1
4	Bayes’ Theorem	1
5	Summary Statistics and Independence	1
6	Important probability Distributions (Lecture 1)	1
7	Important probability Distributions (Lecture 2)	1
8	Conjugacy and the Exponential Family (Lecture 1)	1
9	Conjugacy and the Exponential Family (Lecture 2)	1
10	Change of Variables/Inverse Transform	1
4	**Module-IV (Optimization)**	**7**
1	Optimization Using Gradient Descent.	1
2	Gradient Descent With Momentum, Stochastic Gradient Descent	1
3	Constrained Optimization and Lagrange Multipliers (Lecture 1)	1
4	Constrained Optimization and Lagrange Multipliers (Lecture 2)	1
5	Convex Optimization	1
6.	Linear Programming	1
7.	Quadratic Programming	1
5	**Module-V (CENTRAL MACHINE LEARNING PROBLEMS)**	**14**
1.	Data and Learning models - Empirical Risk Minimization,	1
2.	Parameter Estimation	1
3.	Directed Graphical Models	1
4.	Linear Regression	1
5.	Bayesian Linear Regression	1
6.	Maximum Likelihood as Orthogonal Projection	1
7.	Dimensionality Reduction with Principal Component Analysis - Maximum Variance Perspective, Projection Perspective.	1
8.	Eigenvector Computation and Low Rank Approximations	1
9.	Density Estimation with Gaussian Mixture Models	1

10.	Parameter Learning via Maximum Likelihood	1
11.	EM Algorithm	1
12.	Classification with Support Vector Machines - Separating Hyperplanes	1
13.	Primal Support Vector Machines, Dual Support Vector Machines	1
14.	Kernels	1
*Assignments may include applications of the above theory. With respect to module V, programming assignments may be given.		