

221TC S100	ADVANCED MACHINE LEARNING	CATEGORY	L	T	P	CREDIT
		CORE	3	0	0	3

Preamble: This course introduces machine learning concepts and popular machine learning algorithms. It will cover the standard and most popular supervised learning algorithms including linear regression, logistic regression, decision trees, k-nearest neighbour, an introduction to Bayesian learning and the naive Bayes algorithm, support vector machines and kernels and basic clustering algorithms. Dimensionality reduction methods and some applications to real world problems will also be discussed. It helps the learners to develop application machine learning based solutions for real world applications.

Course Outcomes:

After the completion of the course the student will be able to:*

CO 1	Analyse the Machine Learning concepts, classifications of Machine Learning algorithms and basic parameter estimation methods. (Cognitive Knowledge Level: Analyse)
CO 2	Illustrate the concepts of regression and classification techniques (Cognitive Knowledge Level: Apply)
CO 3	Describe unsupervised learning concepts and dimensionality reduction techniques. (Cognitive Knowledge Level: Apply)
CO 4	Explain Support Vector Machine concepts and graphical models. (Cognitive Knowledge Level: Apply)
CO 5	Choose suitable model parameters for different machine learning techniques and to evaluate a model performance. (Cognitive Knowledge Level: Apply)
CO6	Design, implement and analyse machine learning solution for a real world problem. (Cognitive Knowledge Level: Create)

Program Outcomes (PO)

Outcomes are the attributes that are to be demonstrated by a graduate after completing the course.

PO1: An ability to independently carry out research/investigation and developmentwork in engineering and allied streams

PO2: An ability to communicate effectively, write and present technical reports on complex engineering activities by interacting with the engineering fraternity and with society at large.

PO3: An ability to demonstrate a degree of mastery over the area as per the specialization of the program. The mastery should be at a level higher than the requirements in the appropriate bachelor program

PO4: An ability to apply stream knowledge to design or develop solutions for real world problems by following the standards

PO5: An ability to identify, select and apply appropriate techniques, resources and state-of-the-art tool to model, analyse and solve practical engineering problems.

PO6: An ability to engage in life-long learning for the design and development related to the stream related problems taking into consideration sustainability, societal, ethical and environmental aspects

PO7: An ability to develop cognitive load management skills related to project management and finance which focus on Entrepreneurship and Industry relevance.

Mapping of course outcomes with program outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7
CO 1	☑		☑		☑	☑	
CO 2	☑		☑	☑	☑	☑	
CO 3	☑		☑	☑	☑	☑	
CO 4	☑		☑	☑	☑	☑	
CO 5	☑		☑	☑	☑	☑	
CO 6	☑	☑	☑	☑	☑	☑	☑

Assessment Pattern

Bloom's Category	End Semester Examination
Apply	60-80%
Analyse	20-40%
Evaluate	
Create	

Mark distribution

Total Marks	CIE	ESE	ESE Duration
100	40	60	2.5 hours

Continuous Internal Evaluation Pattern:

Evaluation shall only be based on application, analysis or design based questions (for both internal and end semester examinations).

Continuous Internal Evaluation : 40 marks

Micro project/Course based project : 20 marks

Course based task/Seminar/Quiz : 10 marks

Test paper, 1 no. : 10 marks

The project shall be done individually. Group projects not permitted.

Test paper shall include minimum 80% of the syllabus.

Course based task/test paper questions shall be useful in the testing of knowledge, skills, comprehension, application, analysis, synthesis, evaluation and understanding of the students.

End Semester Examination Pattern:

The end semester examination will be conducted by the University. There will be two parts; Part A and Part B. Part A contain 5 numerical questions with 1 question from each module, having 5 marks for each question. (such questions shall be useful in the testing of knowledge, skills, comprehension, application, analysis, synthesis, evaluation and understanding of the students). Students shall answer all questions.

Part B will contain 7 questions (such questions shall be useful in the testing of overall achievement and maturity of the students in a course, through long answer questions relating to theoretical/practical knowledge, derivations, problem solving and quantitative evaluation), with minimum one question from each module of which student should answer any five. Each question can carry 7 marks.

Total duration of the examination will be 150 minutes.

Course Level Assessment Questions

Course Outcome 1 (CO1):

1. Suppose that X is a discrete random variable with the following probability mass function: where $0 \leq \theta \leq 1$ is a parameter. The following 10 independent observations were taken from such a distribution: $(3, 0, 2, 1, 3, 2, 1, 0, 2, 1)$. What is the maximum likelihood estimate of θ .

X	0	1	2	3
$P(X)$	$2\theta/3$	$\theta/3$	$2(1 - \theta)/3$	$(1 - \theta)/3$

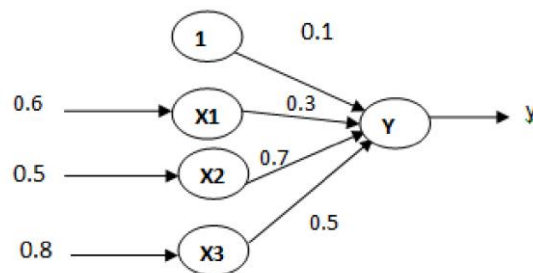
2. What is the difference between Maximum Likelihood estimation (MLE) and Maximum a Posteriori (MAP) estimation?
3. A gamma distribution with parameters α, β has the following density function, where $\Gamma(t)$ is the gamma function.

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

If the posterior distribution is in the same family as the prior distribution, then we say that the prior distribution is the conjugate prior for the likelihood function. Using the Gamma distribution as a prior, show that the Exponential distribution is a conjugate prior of the Gamma distribution. Also, find the maximum a posteriori estimator for the parameter of the Exponential distribution as a function of α and β .

Course Outcome 2 (CO2)

1. How can we interpret the output of a two-class logistic regression classifier as a probability?
2. Calculate the output of the following neuron Y if the activation function is a binary sigmoid.



3. Suppose you have a 3-dimensional input $x = (x_1, x_2, x_3) = (2, 2, 1)$ fully connected with weights $(0.5, 0.3, 0.2)$ to one neuron which is in the hidden layer with sigmoid activation function. Calculate the output of the hidden layer neuron.
4. Consider the case of the XOR function in which the two points $\{(0, 0), (1, 1)\}$ belong to one class, and the other two points $\{(1, 0), (0, 1)\}$ belong to the other class. Design a multilayer perceptron for this binary classification problem.
5. Why does a single perceptron cannot simulate simple XOR function? Explain how this limitation is overcome?
6. Consider a naive Bayes classifier with 3 boolean input variables, X_1, X_2 and X_3 , and one boolean output, Y . How many parameters must be estimated to train such a naive Bayes classifier? How many parameters would have to be estimated to learn the above classifier if we do not make the naive Bayes conditional independence assumption?

Course Outcome 3(CO3):

1. Describe the basic operation of k-means clustering.
2. A Poisson distribution is used to model data that consists of non-negative integers. Suppose you observe m integers in your training set. Your model assumption is that each integer is sampled from one of two different Gaussian distributions. You would like to learn this model using the EM algorithm. List all the parameters of the model. Derive the E-step and M-step for this model.
3. A uni-variate Gaussian distribution is used to model data that consists of non-negative integers. Suppose you observe m integers in your training set. Your model assumption is that each integer is sampled from one of two different Gaussian distributions. You would like to learn this model using the EM algorithm. List all the parameters of the model. Derive the E-step and M-step for the model.
4. Suppose you want to cluster the eight points shown below using **k**-means

	A_1	A_2
x_1	2	10
x_2	2	5
x_3	8	4
x_4	5	8
x_5	7	5
x_6	6	4
x_7	1	2
x_8	4	9

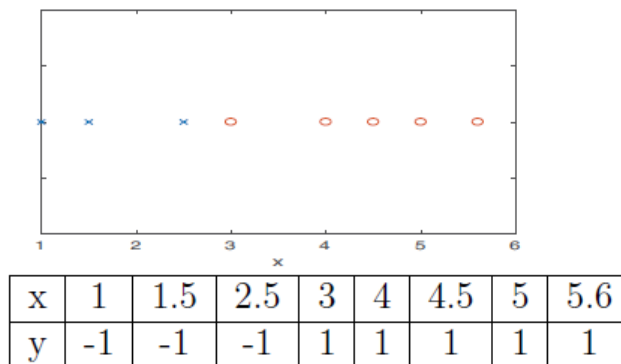
Assume that $k = 3$ and that initially the points are assigned to clusters as follows:

$C1 = \{x_1, x_2, x_3\}$, $C2 = \{x_4, x_5, x_6\}$, $C3 = \{x_7, x_8\}$. Apply the **k**-means algorithm until convergence, using the Manhattan distance.

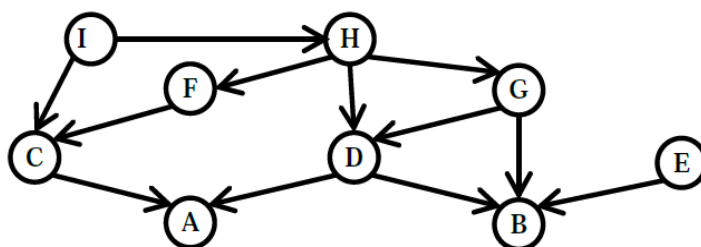
Course Outcome 4 (CO4):

1. Describe how Support Vector Machines can be extended to make use of kernels. Illustrate with reference to the Gaussian kernel $K(x, y) = e^{-\gamma}$, where $\gamma = (x-y)^2$.
2. 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? Justify your answer.
3. What is the primary motivation for using the kernel trick in machine learning algorithms?

4. Show that the Boolean function $(x_1 \wedge x_2) \vee (\neg x_1 \wedge \neg x_2)$ is not linearly separable (i.e. there is no linear classifier $\text{sign}(w_1 x_1 + w_2 x_2 + b)$ that classifies all 4 possible input points correctly). Assume that “true” is represented by 1 and “false” is represented by -1. Show that there is a linear separator for this Boolean function when we use the kernel $K(x, y) = (x \cdot y)^2$ ($x \cdot y$ denotes the ordinary inner product). Give the weights and the value of b for one such separator.
5. 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. Suppose a SVM is used to classify this data. Indicate which are the support vectors and mark the decision boundary. Give the value of the cost function and of the model parameters after training.



6. Write down the factored conditional probability expression that corresponds to the graphical Bayesian Network shown below.



7. How do we learn the conditional probability tables(CPT) in Bayesian networks if information about some variables is missing? How are these variables called?

Course Outcome 5 (CO5):

- Suppose 10000 patients get tested for flu; out of them, 9000 are actually healthy and 1000 are actually sick. For the sick people, a test was positive for 620 and negative for 380. For healthy people, the same test was positive for 180 and negative for 8820. Construct a confusion matrix for the data and compute the accuracy, precision and recall for the data.
- Given the following data, construct the ROC curve of the data. Compute the AUC.

Thres hold	TP	TN	FP	FN
1	0	25	0	29
2	7	25	0	22
3	18	24	1	11
4	26	20	5	3
5	29	11	14	0
6	29	0	25	0
7	29	0	25	0

3. With an example classification problem, explain the following terms: a) Hyper parameters
b) Training set c) Validation sets d) Bias e) Variance.
4. What is ensemble learning? Can ensemble learning using linear classifiers learn classification of linearly non-separable sets?
5. Describe boosting. What is the relation between boosting and ensemble learning?
6. Classifier A attains 100% accuracy on the training set and 70% accuracy on the test set. Classifier B attains 70% accuracy on the training set and 75% accuracy on the test set. Which one is a better classifier. Justify your answer.
7. What are ROC space and ROC curve in machine learning? In ROC space, which points correspond to perfect prediction, always positive prediction and always negative prediction? Why?
8. Suppose there are three classifiers A,B and C. The (FPR, TPR) measures of the three classifiers are as follows – A (0, 1), B (1, 1) , C (1,0.5). Which can be considered as a perfect classifier? Justify your answer.
9. What does it mean for a classifier to have a high precision but low recall?

Model Question Paper

QP CODE:

Reg No: _____

Name: _____

PAGES : 4

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

FIRST SEMESTER M.TECH DEGREE EXAMINATION, MONTH & YEAR

Course Code: 221TCS100

Course Name: ADVANCED MACHINE LEARNING

Max. Marks : 60

Duration: 2.5

Hours

PART A

Answer All Questions. Each Question Carries 5 Marks

1. Explain the principle of the gradient descent algorithm.
2. In a two-class logistic regression model, the weight vector $\mathbf{w} = [4, 3, 2, 1, 0]$. We apply it to some object that we would like to classify; the vectorized feature representation of this object is $\mathbf{x} = [-2, 0, -3, 0.5, 3]$. What is the probability, according to the model, that this instance belongs to the positive class?
3. Expectation maximization (EM) is designed to find a maximum likelihood setting of the parameters of model when some of the data is missing. Does the algorithm converge? If so, do you obtain a locally or globally optimal set of parameters?
4. What is the basic idea of a Support Vector Machine?
5. What is the trade-off between bias and variance? (5x5=25)

Part B

(Answer any five questions. Each question carries 7 marks)

6. Suppose x_1, \dots, x_n are independent and identically distributed(iid) samples from a distribution with density (7)

$$f_X(x | \theta) = \begin{cases} \frac{\theta x^{\theta-1}}{3^\theta}, & 0 \leq x \leq 3 \\ 0, & \text{otherwise} \end{cases}$$

Find the maximum likelihood estimate(MLE) for θ .

7. Derive the gradient descent training rule assuming for the target function $\mathbf{o}_d = \mathbf{w}_0 + \mathbf{w}_1 \mathbf{x}_1 + \dots + \mathbf{w}_n \mathbf{x}_n$. Define explicitly the squared cost/error function E , assuming that a set of training examples D is provided, where each training example $\mathbf{d} \in D$ is associated with the target output t_d . (7)

8. Cluster the following eight points representing locations into three clusters: $A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)$. (7)

Initial cluster centers are: $A1(2, 10), A4(5, 8)$ and $A7(1, 2)$.

The distance function between two points $\mathbf{a} = (x_1, y_1)$ and $\mathbf{b} = (x_2, y_2)$ is defined as $D(\mathbf{a}, \mathbf{b}) = |x_2 - x_1| + |y_2 - y_1|$

Use k-Means Algorithm to find the three cluster centers after the second iteration.

9. Describe Principal Component Analysis. What criterion does the method minimize? What is the objective of the method? Give a way to compute the solution from a matrix X encoding the features. (7)

10. Consider a support vector machine whose input space is 2-D, and the inner products are computed by means of the kernel $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^2 - 1$ ($\mathbf{x} \cdot \mathbf{y}$ denotes the ordinary inner product). Show that the mapping to feature space that is implicitly defined by this kernel is the mapping to 5-D given by (7)

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \rightarrow \phi(\mathbf{x}) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2} x_1 x_2 \\ \sqrt{2} x_1 \\ \sqrt{2} x_2 \end{bmatrix}.$$

11. How does random forest classifier work? Why is a random forest better than a decision tree? (7)

12. Consider a two-class classification problem of predicting whether a photograph contains a man or a woman. Suppose we have a test dataset of 10 records with expected outcomes and a set of predictions from our classification algorithm. Compute the confusion matrix, accuracy, precision, recall, sensitivity and specificity on the following data. (7)

Sl.No.	Actual	Predicted
1	man	woman
2	man	man
3	woman	woman
4	man	man
5	man	woman
6	woman	woman
7	woman	man
8	man	man
9	man	woman
10	woman	woman

Syllabus

Module-1 (Parameter Estimation and Regression) 8 hours

Overview of machine learning: supervised, semi-supervised, unsupervised learning, reinforcement learning. Basics of parameter estimation: Maximum Likelihood Estimation(MLE), Maximum a Posteriori Estimation (MAP). Gradient Descent Algorithm, Batch Gradient Descent, Stochastic Gradient Descent. Regression algorithms: least squares linear regression, normal equations and closed form solution, Polynomial regression.

Module-2 (Regularization techniques and Classification algorithms) 9 hours

Overfitting, Regularization techniques - LASSO and RIDGE. Classification algorithms: linear and non-linear algorithms, Perceptrons, Logistic regression, Naive Bayes, Decision trees. Neural networks : Concept of Artificial neuron, Feed-Forward Neural Network, Back propagation algorithm.

Module-3 (Unsupervised learning) 8 hours

Unsupervised learning: clustering, k-means, Hierarchical clustering, Principal component analysis, Density-based spatial clustering of applications with noise (DBSCAN). Gaussian mixture models:

Expectation Maximization (EM) algorithm for Gaussian mixture model.

Module-4 (Support Vector Machine and Graphical Models) 7 hours

Support vector machines and kernels : Max margin classification, Nonlinear SVM and the kernel trick, nonlinear decision boundaries, Kernel functions. Basics of graphical models - Bayesian networks, Hidden Markov model - Inference and estimation.

Module-5 (Evaluation Metrics and Sampling Methods) 8 hours

Classification Performance Evaluation Metrics: Accuracy, Precision, Precision, Recall, Specificity, False Positive Rate (FPR), F1 Score, Receiver Operator Characteristic (ROC) Curve, AUC. Regression Performance Evaluation Metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R Squared/Coefficient of Determination. Clustering Performance Evaluation Metrics: Purity, Jaccard index, Normalized Mutual Information, Clustering Accuracy, Silhouette Coefficient, Dunn's Index. Boosting: AdaBoost, gradient boosting machines. Resampling methods: cross-validation, bootstrap. Ensemble methods: bagging, boosting, random forests Practical aspects in machine learning: data preprocessing, overfitting, accuracy estimation, parameter and model selection Bias-Variance tradeoff

Course Plan

No	Topics	No. of Lectures (40)
1	Module-1 (Parameter Estimation and Regression) 8 hours	
1.1	Overview of machine learning: supervised, semi-supervised, unsupervised learning, reinforcement learning.	1
1.2	Basics of parameter estimation: Maximum Likelihood Estimation(MLE)	1
1.3	Basics of parameter estimation: Maximum Likelihood Estimation(MLE) - Examples	1
1.4	Basics of parameter estimation: Maximum a Posteriori Estimation (MAP)	1
1.5	Basics of parameter estimation: Maximum a Posteriori Estimation (MAP) - Example	1
1.6	Gradient Descent Algorithm, Batch Gradient Descent, Stochastic Gradient Descent	1
1.7	Regression algorithms: least squares linear regression, normal equations and closed form solution	1
1.8	Polynomial regression	1
2	Module-2 (Regularization techniques and Classification algorithms) 9 hours	
2.1	Overfitting, Regularization techniques - LASSO and RIDGE	

2.2	Classification algorithms: linear and non-linear algorithms	
2.3	Perceptrons	
2.4	Logistic regression	
2.5	Naive Bayes	
2.6	Decision trees	
2.7	Neural networks : Concept of Artificial neuron	
2.8	Feed-Forward Neural Network	
2.9	Back propagation algorithm	
3	Module-3 (Unsupervised learning) 8 hours	
3.1	Unsupervised learning: clustering, k-means	
3.2	Hierarchical clustering	
3.3	Principal component analysis	
3.4	Density-based spatial clustering of applications with noise (DBSCAN)	
3.5	Gaussian mixture models: Expectation Maximization (EM) algorithm for Gaussian mixture model	
3.6	Gaussian mixture models: Expectation Maximization (EM) algorithm for Gaussian mixture model	
4	Module-4 (Support Vector Machine and Graphical Models) 7 hours	
4.1	Support vector machines and kernels : Max margin classification	
4.2	Support vector machines: Max margin classification	
4.3	Nonlinear SVM and the kernel trick, nonlinear decision boundaries	
4.3	Kernel functions	
4.5	Basics of graphical models - Bayesian networks	
4.6	Hidden Markov model - Inference and estimation	
4.7	Hidden Markov model - Inference and estimation	
4.8	Hidden Markov model - Inference and estimation	
5	Module-5 (Evaluation Metrics and Sampling Methods) 8 hours	
5.1	Classification Performance Evaluation Metrics: Accuracy, Precision, Precision, Recall, Specificity, False Positive Rate (FPR), F1 Score, Receiver Operator Characteristic (ROC) Curve, AUC	
5.2	Regression Performance Evaluation Metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R Squared/Coefficient of Determination	

5.3	Clustering Performance Evaluation Metrics: Purity, Jaccard index, Normalized Mutual Information, Clustering Accuracy, Silhouette Coefficient, Dunn's Index	
5.4	Boosting: AdaBoost, gradient boosting machines.	
5.5	Resampling methods: cross-validation, bootstrap.	
5.6	Ensemble methods: bagging, boosting, random forests	
5.7	Practical aspects in machine learning: data preprocessing, overfitting, accuracy estimation, parameter and model selection	
5.8	Bias-Variance tradeoff	

Reference Books

1. Christopher Bishop. Neural Networks for Pattern Recognition, Oxford University Press, 1995.
2. Kevin P. Murphy. Machine Learning: A Probabilistic Perspective, MIT Press 2012.
3. Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements Of Statistical Learning, Second edition Springer 2007.
4. Ethem Alpaydin, Introduction to Machine Learning, 2nd edition, MIT Press 2010.
5. Tom Mitchell, Machine Learning, McGraw-Hill, 1997.