

25UAMT412	STATISTICAL LEARNING & INFERENCE	Category	L	T	P	R	Credit	Year of Introduction
		PEC	3	0	0	0	3	2025

Preamble

The objective of this course is to provide students with a strong foundation in statistical inference and modern statistical learning techniques used for data analysis and predictive modeling. The course covers probability theory, parameter estimation, hypothesis testing, regression models, classification techniques, and model evaluation methods.

Prerequisite: 25UMAT301 Probability, Random Process and Numerical Methods; 25UAMT Object Oriented Programming in Python

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

CO	Course Outcomes	Knowledge Level
CO1	Apply probability theory and statistical inference techniques to model uncertainty and draw meaningful conclusions from data.	K3 Apply
CO2	Apply estimation and hypothesis testing methods to solve practical statistical problems.	K3 Apply
CO3	Develop and implement regression and classification models for predictive analysis.	K3 Apply
CO4	Apply model evaluation and validation techniques to assess the performance of statistical learning models.	K3 Apply

Mapping of course outcomes with program outcomes and program specific outcomes

CO/PO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PSO1	PSO2
CO1	3	3	-	-	-	-	-	-	-	-	2	2	-
CO2	3	3	-	-	-	-	-	-	-	-	2	2	-
CO3	3	3	2	-	-	-	-	-	-	-	2	3	2
CO4	3	3	2	-	-	-	-	-	-	-	2	3	2
AVG	3	3	2	-	-	-	-	-	-	-	2	3	2

Mapping of course outcomes with Knowledge and Attitude Profile (WK)

	WK1	WK2	WK3	WK4	WK5	WK6	WK7	WK8	WK9
CO1	*	*	*	*				*	
CO2	*	*	*	*				*	
CO3	*	*	*	*	*			*	

CO4	*	*	*	*	*			*	
-----	---	---	---	---	---	--	--	---	--

Wks Mapped with the course

WK1	WK2	WK3	WK4	WK5	WK6	WK7	WK8	WK9
*	*	*	*	*			*	

Mark Distribution:

Total Marks	CIE Marks	ESE Marks	ESE Duration
100	40	60	2.5 hrs

Continuous Internal Evaluation Pattern:

Attendance : 5 marks

Continuous Assessment Test : 20 marks

Continuous Assessment Assignment : 15 marks

SYLLABUS

Module 1 (10 hrs)

Probability concepts for machine learning, random variables in data modeling (discrete and continuous), expectation and variance in data analysis, commonly used distributions in machine learning (Binomial, Poisson, Normal), and the role of the Central Limit Theorem in data-driven model development and performance evaluation.

Module 2 (10 hrs)

Point estimation (Maximum Likelihood Estimation), confidence intervals, hypothesis testing (Z-test, t-test, chi-square test), p-values, Type I and Type II errors.

Module 3 (12 hrs)

Simple and multiple linear regression, logistic regression, bias–variance tradeoff, classification concepts, decision boundaries, introduction to k-Nearest Neighbors and decision trees.

Module 4 (12 hrs)

Model evaluation and performance metrics: confusion matrix, accuracy, precision, recall, F1-score, ROC curve and AUC. Cross-validation techniques and model selection. Overfitting and underfitting, regularization methods (L1 and L2). Introduction to ensemble learning: bagging, boosting, and random forests. Basic concepts of dimensionality reduction and feature selection.

Text Books:

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R (2nd ed.). Springer.
2. Murphy, K. P. (2022). Probabilistic machine learning: An introduction. MIT Press.
3. Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (3rd ed.). O'Reilly Media.

Reference Books:

1. Bishop, C. M., & Bishop, H. (2024). Deep learning: Foundations and concepts. Springer.
2. Hastie, T., Tibshirani, R., & Friedman, J. (2021). The elements of statistical learning (2nd ed.). Springer.
3. Wasserman, L. (2004). All of statistics: A concise course in statistical inference. Springer.

END SEMESTER EXAMINATION – MODEL QUESTION PAPER

Course Code	25UAMT412
Course Name	STATISTICAL LEARNING & INFERENCE
Max. Marks: 60	Duration: 2.5 Hours

Part A (Answer all questions. Each question carries 3 marks)				
Qn No.	Question	Marks	CO	BL
1.	A random variable X follows a Normal distribution with $\mu = 50$ and $\sigma = 5$. Find $P(45 < X < 60)$. State and explain the Central Limit Theorem and its significance in statistical inference.	3	CO1	L3
2.	Explain the Binomial and Poisson distributions. When is the Poisson distribution used as an approximation to the Binomial distribution	3	CO1	L2
3.	Explain Maximum Likelihood Estimation (MLE). Derive the MLE estimator for the mean of a Normal distribution.	3	CO2	L3
4.	A sample of 25 students had a mean score of 72 with standard deviation 10. Test at 5% significance level whether the population mean is 70. (Use t-test, critical value $t(0.025, 24) = 2.064$)	3	CO2	L3
5.	Explain the bias-variance tradeoff in machine learning models. How does model complexity affect bias and variance?	3	CO3	L2
6.	Derive the cost function for logistic regression and explain the sigmoid function and its properties.	3	CO3	L3
7.	Explain the ROC curve and AUC. For two classifiers with AUC values of 0.85 and 0.72, which is better and why?	3	CO4	L2
8.	Explain L1 and L2 regularization. How do they prevent overfitting? What is the effect on model weights?	3	CO4	L2

Part B (Answer any one full question from each Module. Each carry 9 marks)				
Module I				
Qn No.	Question	Marks	CO	BL

9.1	Explain discrete and continuous random variables with examples. Derive the mean and variance of the Binomial distribution.	5	CO1	L3
9.2	Explain the concept of expectation and variance. If X and Y are independent random variables, prove that $E(X+Y) = E(X) + E(Y)$ and $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$.	4	CO1	L3
OR				
10.1	Explain the Normal distribution and its properties. Derive the standard normal form and explain how z-scores are used in probability calculations.	5	CO1	L2
10.2	State and prove the Central Limit Theorem. Explain its importance in hypothesis testing and confidence interval construction.	4	CO1	L3
Module II				
Qn No.	Question	Marks	CO	BL
11	Explain hypothesis testing with the following components: null and alternative hypothesis, Type I and Type II errors, p-values, and significance levels. Apply Z-test and t-test to the following problem: A manufacturer claims the mean weight of a product is 500g. A sample of 36 items has mean 498g with $\sigma=6$. Test at $\alpha=0.05$.	9	CO2	L3
OR				
12.1	Explain confidence intervals for population mean. Construct a 95% confidence interval for a sample with $n=100$, $\text{mean}=75$, and $s=12$.	5	CO2	L3
12.2	Explain the chi-square test for goodness of fit with a suitable example. When is it used instead of a t-test?	4	CO2	L2
Module III				
Qn No.	Question	Marks	CO	BL
13	Explain multiple linear regression. Describe the assumptions of the linear regression model and methods to validate them (residual plots, QQ plot). Fit a regression model to the following data and compute the coefficient of determination R^2 : [Provide dataset in exam]. Interpret the regression coefficients.	9	CO3	L3
OR				
14.1	Explain logistic regression for binary classification. Derive the decision boundary and explain how probabilities are estimated.	5	CO3	L3
14.2	Explain the k-Nearest Neighbors (k-NN) classification algorithm. How is the optimal k chosen? What are its computational	4	CO3	L3

	limitations.			
Module IV				
Qn No.	Question	Marks	CO	BL
15.1	Explain ensemble learning methods: Bagging, Boosting, and Random Forests. How does Boosting differ from Bagging in reducing variance and bias.	5	CO4	L3
15.2	Explain k-fold cross-validation with an example. What is the relationship between cross-validation score and model generalization?	4	CO4	L3
OR				
16	Describe model evaluation metrics for classification: confusion matrix, accuracy, precision, recall, F1-score, and ROC-AUC. Compute all metrics for the following confusion matrix: TP=80, FP=10, FN=20, TN=90. Explain when precision is preferred over recall and vice versa.	9	CO4	L3