COURSE PLAN

For

Statistical Machine Learning (CSET211)

Faculty Name : Dr. Jagendra Singh
Ms. Sanchali Das

Course Type : Specialized Core-I

Semester and Year: III Semester / II Year

L-T-P : 3-0-2

Credits : 4

School : SCSET

Course Level : UG

School of Computer Science Engineering and Technology

Bennett University
Greater Noida, Uttar Pradesh

COURSE CONTEXT

SCHOOL	**SCSET**	**VERSION NO. OF CURRICULUM/SYLLABUS THAT THIS COURSE IS A PART OF**	**V1**
DEPARTMENT		**DATE THIS COURSE WILL BE EFFECTIVE FROM**	**July-Dec, 2022**
DEGREE	**B. Tech.**	**VERSION NUMBER OF THIS COURSE**	**2**

COURSE BRIEF

COURSE TITLE	**Statistical Machine Learning**	**PRE-REQUISITES**	**NA**
COURSE CODE	**CSET211**	**TOTAL CREDITS**	**4**
COURSE TYPE	**Specialized Core-I**	**L-T-P FORMAT**	**3-0-2**

COURSE SUMMARY

This course includes machine learning concepts, Statistical Theories, Supervised learning, high dimensional data and the role of sparsity, Learning theory, Risk minimization, Classification and regression, and EM algorithm. It also covers important topics such as parametric and non-parametric methods, theory of generalization, regularization, the role of sparsity in high dimensional data, and surrogate loss functions. In a broader sense, the course offers a thorough understanding of statistical ML concepts that help students design and implement daily life learning applications.

COURSE-SPECIFIC LEARNING OUTCOMES (CO)

By the end of this program, students should have the following knowledge, skills and values:

CO1: To articulate key features and methods of Statistical Machine Learning (SML).

CO2: To formulate and design the given application as a statistical machine learning problem.

CO3: To implement and evaluate common statistical machine learning techniques.

How are the above COs aligned with the Program-Specific Objectives (POs) of the degree?

The course outcomes are aligned to inculcating inquisitiveness in understanding cutting edge areas of computer science engineering and allied disciplines with their potential impacts.

CO - PO Mapping

COs ⟹ POs	PO1	PO2	PO3	PO4	PO5	PO6
CO1	✓					
CO2						
CO3						

Detailed Syllabus

Module 1 (Contact hours:7)
Statistical Theory, Supervised Learning, Unsupervised Learning, Data and Types, Feature variable, Machine Learning, Statistics terms, Supervised learning, Concentration inequalities, Generalization bounds, Plugin classifiers, Least-squares methods, Bias vs Variance, Theory of generalization, Understand Underfitting, Overfitting, Parametric methods, Maximum likelihood, Bayes algorithm, Minimax algorithm, Expectation-Maximization, Advantages and Disadvantages, Applications of EM Algorithm, Use case of EM Algorithm.

Module 2 (Contact hours:10)
Bayesian versus Non-Bayesian approaches, Density estimation, Gaussian Distributions, Gaussian Mixture Models, Gaussian Discriminant Analysis, Independent Component Analysis, Convexity and Optimization: Convexity, Conjugate functions, Nonparametric classifications methods, Unconstrained optimization, Constrained optimization, Nonparametric methods, KKT conditions, Lagrangian minimization, Primal feasibility, Dual feasibility, Complementary slackness.

Module 3 (Contact hours:13)
Basis pursuit, Polynomial Expansion, Feature maps, The "kernel trick", Vapnik-Chervonenkis (VC) dimension, VC generalization bounds, Sparsity: High dimensional data, The role of sparsity, Sparsistency, Consistency, Persistency, Sparsity in nonparametric regression, Sparsity in graphical models, Greedy algorithms, Sparse linear regression, Compressed sensing, Nonparametric Methods: Nonparametric regression, Density estimation, Factor Analysis, Matrix Factorization, The bootstrap, Subsampling, Nonparametric Bayes.

Module 4 (Contact hours: 12)
Probability Distributions for modelling, Markov Networks, Hidden Markov Model, Advanced Theory: Concentration of measure, Covering numbers, Learning theory, Exact learning (Dana Angluin), Probably approximately correct learning (PAC learning), VC theory (Vladimir Vapnik and Alexey Chervonenkis), Risk minimization and its approaches, Bundle Methods, Graph Analytics, Graph-based machine learning algorithms, Simulation methods, Variational methods, Tsybakov noise conditions, Surrogate loss functions, Minimax rates for classification, Minimax rates for regression, Manifold methods, Spectral methods.
STUDIO WORK / LABORATORY EXPERIMENTS:

Students will gain practical experience with the implementation of different statistical methods by using different statistical machine learning tools. Eventually, the lab works formulate the problem as a statistical machine learning problem followed by its implementation.

TEXTBOOKS/LEARNING RESOURCES:

a) Masashi Sugiyama, Introduction to Statistical Machine Learning (1 ed.), Morgan Kaufmann, 2017. ISBN 978-0128021217.
b) T. M. Mitchell, Machine Learning (1 ed.), McGraw Hill, 2017. ISBN 978-1259096952.

REFERENCE BOOKS/LEARNING RESOURCES:

a) Richard Golden, Statistical Machine Learning A Unified Framework (1 ed.), CRC Press, 2020. ISBN 9781351051490.

TEACHING-LEARNING STRATEGIES

The course will be taught using a combination of the best practices of teaching-learning. Multiple environments will be used to enhance the outcomes such as seminar, self-learning, MOOCs, group discussions and ICT based tools for class participation along with the classroom sessions. The teaching pedagogy being followed includes more exposure to hands-on experiment and practical implementations done in the lab sessions. To match with the latest trend in academics, case study, advanced topics and research oriented topics are covered to lay down the foundation and develop the interest in the students leading to further exploration of the related topics. To make the students aware of the industry trends, one session of expert lecture will be organized to provide a platform to the students for understanding the relevant industry needs.

EVALUATION POLICY

Components of Course Evaluation	Percentage Distribution
Mid Term Examination	10
End Term Examination	30
Lab Continuous Evaluation	30
Class Participation	10
Project Demonstration (Hackathon)	20
Total	100

Lecture Wise Plan

No.	Content Planned
1.	Course handout and Assessment mechanism (15) Statistical Theory (15) Supervised Learning (10) Unsupervised Learning (10)
2.	Data and Types (15) Feature variable (15) Machine Learning, Statistics terms (20)
3.	Concentration inequalities (15) Generalization bounds (15) Plugin classifiers (15)
4.	Least-squares methods (30) Bias vs Variance (20)
5.	Theory of generalization (20) Understand Underfitting (15) Overfitting (15)
6.	Parametric methods (10) Maximum likelihood (20) Bayes algorithm (10) Minimax algorithm (10)
7.	Expectation-Maximization (EM) (30) Advantages and Disadvantages (5) Applications of EM Algorithm (5) Use case of EM Algorithm (10)
8.	**Buffer Lecture/Assessment**
9.	Bayesian versus Non-Bayesian approaches (30) Density estimation (20)
10.	Gaussian Distributions (20) Gaussian Mixture Models (GMMs) (30)
11.	Gaussian Discriminant Analysis (25) Independent Component Analysis (25)
12.	Convexity and Optimization: Convexity (30) Conjugate functions (20)
13.	Nonparametric classifications methods (50)
14.	Unconstrained optimization (25) Constrained optimization (25)
15.	Nonparametric methods (25) KKT conditions (5) Lagrangian minimization (5) Primal feasibility (5) Dual feasibility (5) Complementary slackness (5)
16.	**Industry Talk (50)**
17.	**Buffer /Startups related to Machine Learning and Analytics (50)**
18.	Basis pursuit (25) Polynomial Expansion (25)
19.	Feature maps (30)

	The “kernel trick” (20)
20.	Vapnik-Chervonenkis (VC) dimension (25) VC generalization bounds (25)
21.	Sparsity: High dimensional data (30) The role of sparsity (25)
22.	Sparsistency (15) Consistency (15) Persistency (20)
23.	Sparsity in nonparametric regression (25) Sparsity in graphical models (25)
24.	Greedy algorithms (20) Sparse linear regression (30)
25.	Compressed sensing (50)
26.	Nonparametric Methods: Nonparametric regression (25) Density estimation (25)
27.	Factor Analysis (20) Matrix Factorization (30)
28.	Analytics and Visualization Tools in SML (50)
29.	The bootstrap (15) Subsampling (20) Nonparametric Bayes (15)
30.	**Buffer Lecture/Expert Talks (50)**
31.	Probability Distributions for modelling (50)
32.	Markov Networks (25) Hidden Markov Model (25)
33.	Advanced Theory: Concentration of measure (30) Covering numbers (20)
34.	Learning theory (15) Exact learning (Dana Angluin) (15) Probably approximately correct learning (PAC learning) (10) VC theory (Vladimir Vapnik and Alexey Chervonenkis) (10)
35.	Risk minimization and its approaches (30) Bundle Methods (20)
36.	Graph Analytics (25) Graph-based machine learning algorithms (25)
37.	Advanced Topics in SML (50)
38.	Simulation methods (30) Variational methods (20)
39.	Tsybakov noise conditions (25) Surrogate loss functions (25)
40.	Minimax rates for classification (25) Minimax rates for regression (25)
41.	Manifold methods (25) Spectral methods (25)
42.	**Buffer Lecture/Assessment**

Lab Wise Plan

No.	Content Planned
1.	Implementation of Data Pre-processing techniques using Python Libraries
2	Implementation of simple least square regression
3	Implementation of polynomial expansion of data
4	Implementation of Maximum Likelihood Estimation
5	Implementation of Linear Classification model
6	Implementation of Sparsity operations in linear algebraic
7	**Buffer/Mid-Term Assessment**
8	Implement and apply Global Minima, Maxima, Local Minima and Maxima
9	Apply Constrained Minimization using Inequality constraints
10	Bi-Section Method for solving optimization problems
11	Implementation of probability and distribution functions
12	Project Demo
13	**Buffer/End-Term Assessment**
14	**Buffer/End-Term Assessment**

Hardware required:
Systems available in labs are sufficient for Experimentation.

Software/Platform Required:
All software and tools are available as open-source.

Project Evaluation Component:

1. Working project (Demonstration)
2. Presentation
3. Project report
4. Project video (Upload in YouTube)(max 5 min)
5. Project Poster
6. Github, Social Media Post

Tools :

a. Acadly
b. Mentimeter
c. Anaconda/Jupiter
d. Google Colab
e. PyCharm

Innovation:

- Proposed Industry Talks: Dr Mukesh Prasad (University of Technology Sydney). tentatively in 3rd week of September 2022.
- Startups talks related to the Course

- Mentimeter/ Acadly(https://www.acadly.com/) is used wherever possible for making interaction to the students during online theory/practical classes.
- Case Studies
- Advanced Research Topics
- Certification Mapping: Statistics and Machine Learning (Coursera).
  https://www.coursera.org/specializations/data-science-statistics-machine-learning.
  Statistics for Machine Learning (Udemy).
  https://www.udemy.com/course/statistics-for-machine-learning
  Statistical Learning(EDX)
  https://www.edx.org/course/statistical-learning