IE Technical Elective: Data Mining

FALL, 2026 Course Syllabus

INSTRUCTOR:
Associate Professor Changxi Wang, Ph.D.
Office: SCUPI Building N403
Email: changxi.wang@scai-ji.cn
Office Hours: Thursday 12:20 to 15:20

LECTURES:
Thursday 8:15-11:00, SCUPI Building N209; three 45-minute sections.

RECITATION:
N.A.

TEXTBOOK:
Deep Learning Foundations and Concepts, Christopher M. Bishop and Hugh Bishop, 2024 edition (Springer)

REFERENCES:
Data Mining: Concepts and Techniques, Third Edition, by Jiawei Han, Micheline Kamber, and Jian Pei, Morgan Kaufmann, 2012.

TEACHING ASSISTANT:
Hang Chen 陈航(email: chen_hang@stu.scu.edu.cn)
Office: SCUPI Building N407;
Office Hours: Thursday 12:30 to 15:20
QQ Group: 1062491638

PREREQUISITE:
N.A.

DESCRIPTION:
This elective provides a practical introduction to data mining and machine learning. The course progresses from data tools and model evaluation to clustering, density estimation, regression, classification, gradient-based optimization, neural networks, convolutional neural networks, and an introduction to Transformers and large language models. Emphasis is placed on conceptual understanding, implementation, model evaluation, and a group project; advanced mathematical proofs are not required.

COURSE OBJECTIVES:
By the end of this course, students will be able to:
1. Understand the data analytics and machine-learning lifecycle.
2. Identify industrial problems and formulate them as data-mining or machine-learning tasks.
3. Prepare, transform, visualize, and explore data using Python-based tools.
4. Apply core clustering, regression, and classification models to practical datasets.
5. Select model parameters, regularization methods, and evaluation metrics appropriately.

6. Explain the roles of optimization and modern neural-network models in data mining.

LEARNING OUTCOMES FOR THIS COURSE:

1) Formulate a real-world problem as a data-mining or machine-learning problem.
2) Load, clean, combine, and visualize structured data using Python and pandas.
3) Construct training, validation, and test workflows and avoid common evaluation mistakes.
4) Train and interpret introductory clustering, regression, and classification models.
5) Use regularization and gradient-based optimization to improve model performance.
6) Explain the basic ideas behind neural networks, CNNs, Transformers, and large language models.
7) Evaluate model trade-offs and communicate project methods, results, and limitations.

GRADE DETERMINATION:

Attendance:	10%
Quiz:	10%
Homework:	25%
Project:	30%
Final Examination:	25%
	100%

Attendance

Attendance may be checked periodically throughout the semester.

Homework & Exercises

Homework will be assigned weekly and needed to be finished before the next class. Homework solutions must be submitted to the Blackboard system.

Exams

There will be one open book open notes exam.

Group Project

Group project will be described in separate handouts as they are assigned.

MATERIAL COVERED: The sequence of topics covered in this class is:

Week	Descriptions
1 (Sep 3)	Lecture 01: Introduction: Section 1 (45 min): course scope, objectives, assessment, and the machine-learning lifecycle; Section 2 (45 min): artificial intelligence, machine learning, and supervised versus unsupervised learning; Section 3 (45 min): recognizing machine-learning problems and framing an industrial application.
2 (Sep 10)	Lecture 02: Data Tools: Section 1 (45 min): pandas Series, DataFrame, data types, loading, and inspection; Section 2 (45 min): selection, filtering, sorting, missing values, and column operations; Section 3 (45 min): aggregation, groupby, joins, visualization, and a guided data exercise.

3 (Sep 17)	Lecture 03: ML Mechanics: Section 1 (45 min): generalization and train-validation-test splits; Section 2 (45 min): feature engineering, encoding, and standardization; Section 3 (45 min): scikit-learn workflow, model families, hyperparameters, and baseline evaluation.
4 (Sep 24)	Lecture 04: K-means and Probability: Section 1 (45 min): clustering objectives and Lloyd's K-means algorithm; Section 2 (45 min): choosing K, interpreting clusters, limitations, and a code demonstration; Section 3 (45 min): joint, conditional, and empirical probability plus an intuitive Bayes update.
5 (Oct 8)	Lecture 05: Density Estimation and GMM: Section 1 (45 min): expectation, variance, covariance, and probability-density concepts; Section 2 (45 min): likelihood and maximum-likelihood estimation through simple examples, without full derivations; Section 3 (45 min): Gaussian mixture models, latent assignments, and the EM algorithm at a conceptual level.
6 (Oct 15)	Lecture 06: Linear Regression (1): Section 1 (45 min): regression tasks, linear models, basis functions, and feature representations; Section 2 (45 min): least-squares loss and the normal-equation idea without geometric proof; Section 3 (45 min): predictions, residual plots, MSE, R-squared, and a regression demonstration.
7 (Oct 22)	Lecture 07: Linear Regression (2): Section 1 (45 min): model complexity, overfitting, and the purpose of regularization; Section 2 (45 min): Ridge and Lasso regularization, excluding Lagrangian and SVD derivations; Section 3 (45 min): practical comparison of unregularized, Ridge, and Lasso models.
8 (Oct 29)	Lecture 09: Bias-Variance Trade-off: Section 1 (45 min): bias, noise, and model variance; Section 2 (45 min): the bias-variance trade-off through visual examples rather than formula derivation; Section 3 (45 min): validation curves, model complexity, hyperparameter selection, and the effect of more data.
9 (Nov 5)	Lecture 10: Logistic Regression (1): Section 1 (45 min): binary classification, sigmoid probabilities, and linear decision boundaries; Section 2 (45 min): cross-entropy and maximum-likelihood intuition without the full gradient derivation; Section 3 (45 min): regularization, classification thresholds, and a logistic-regression demonstration.
10 (Nov 12)	Lecture 11: Logistic Regression (2): Section 1 (45 min): multiclass classification and softmax at an introductory level; Section 2 (45 min): confusion matrix, accuracy, precision, recall, and F1 score; Section 3 (45 min): threshold selection, precision-recall curves, ROC curves, and an evaluation exercise.
11 (Nov 19)	Lectures 12-13: Gradient Descent and SGD: Section 1 (45 min): objectives, gradients, batch gradient descent, and learning-rate effects; Section 2 (45 min): momentum, learning-rate schedules, RMSProp, and Adam at a practical level; Section 3 (45 min): stochastic and mini-batch gradient descent, batch-size choices, and a code demonstration.
12 (Nov 26)	Lecture 14: Neural Networks and PyTorch: Section 1 (45 min): artificial neurons, activation functions, multilayer perceptrons, and hidden layers; Section 2 (45 min): deep networks and universal approximation through examples, without theorem proof; Section 3 (45 min): PyTorch tensors, nn.Module, autograd, optimizers, and a small training exercise.
13 (Dec 3)	Lecture 18: Convolutional Neural Networks: Section 1 (45 min): image inductive biases, receptive fields, convolution, and weight sharing; Section 2 (45 min): kernels, channels, padding, stride, and pooling with limited dimension calculations; Section 3 (45 min): basic CNN design, learned filters, historical examples, and an application demonstration.
14 (Dec 10)	Lecture 21: Transformers and LLMs: Section 1 (45 min): tokens, embeddings, self-attention, and the Transformer architecture at a conceptual level; Section 2 (45 min): tokenization, causal

	language modeling, and large language models; Section 3 (45 min): prompting, RAG, applications, limitations, responsible use, and course synthesis.
15 (Dec 21)	Project presentation
TBD(Dec 25)	Final Exam