AIPE3010 MACHINE LEARNING WITH UNSTRUCTURED DATA (3-0-0)

Course Objectives

- Understand characteristics and challenges of unstructured data (text, image, audio, video).
- Learn preprocessing, feature extraction, and representation techniques.
- Apply ML/DL models (CNNs, RNNs, Transformers) for unstructured data tasks.
- Develop end-to-end solutions and evaluate/deploy models for real-world applications.

Module - I

Foundations and Unstructured Data Types

- Introduction to structured vs semi structured vs unstructured data
- Sources of unstructured data: text, images, audio, video, sensor streams, documents (PDF, scans)
- Basics of machine learning: supervised, unsupervised, evaluation metrics (brief recap)
- Data management and pipelines for unstructured data (storage, retrieval, annotation)

Module - II

Text Data and NLP Basics

- Text representation: tokenization, stopwords, stemming/lemmatization
- Bag of Words, TF IDF, word and sentence embeddings (Word2Vec, GloVe, transformer embeddings)
- Core NLP tasks: text classification, sentiment analysis, topic modeling (LDA basics)
- Named Entity Recognition, text preprocessing for downstream ML

Module - III

Images, Vision, and Multimodal Data

- Image as data: pixels, channels, basic preprocessing and augmentation
- Feature extraction: SIFT/HOG (overview), transition to deep features
- Convolutional Neural Networks (CNNs): architecture, transfer learning for classification
- Intro to multimodal learning: aligning text and image representations in simple applications

Module - IV

Advanced Models for Unstructured Data

- Sequence models: RNNs, LSTMs, GRUs for text and time series
- Attention and transformers for text (high level), pre trained language models (BERT style overview)
- Autoencoders and representation learning for images and documents
- Unsupervised and semi supervised learning on unstructured data (clustering, anomaly detection)

Module - V

End to End Applications and MLOps

- Case studies: document understanding (OCR + layout analysis), tabular extraction from PDFs, NER for information extraction, basic recommender with text/image features
- Data quality, bias, and ethics for unstructured ML systems
- Deployment: exporting models, APIs, and monitoring pipelines for unstructured data services

Course Outcomes

Students will be able to:

1. Categorize types of unstructured data and identify associated challenges.
2. Students explain and implement advanced deep architectures for unstructured data and apply unsupervised/semi supervised techniques where labels are scarce.
3. Preprocess and represent text, image, and audio data.
4. Implement ML/DL models such as CNNs, RNNs, LSTMs, and Transformers.
5. Analyze NLP, computer vision, and audio tasks using modern ML frameworks.
6. Build, evaluate, and deploy ML solutions with ethical considerations.

Textbook

1. H. S. Koley and A. Chowdhury, “Machine Learning for Text and Image Data Analysis”, Wiley, 2024.

Reference books

1. C. Bishop, “Pattern Recognition and Machine Learning”, Springer, 2006 (general ML, foundations).
2. Goodfellow, Y. Bengio, and A. Courville, “Deep Learning”, MIT Press, 2016 (deep learning for text and images).
3. V. Verdhan, “Data Without Labels: Unsupervised Learning in Practice”, Manning, 2025 (emphasis on unstructured and unlabeled data)