Welcome to:

EE353 Intro to Data Science and Machine Learning

EE769 Intro to Machine Learning

Course Introduction

Amit Sethi, EE (and KCDH, CMINDS, DSSE), IITB

MeDAL Lab (1st flr, EE), 3528, 7483, asethi@iitb.ac.in

Instructor Introduction

Employment: Current: IIT Bombay

Previous: IIT Guwahati,

ZS Associates, Chicago

NEC Labs, Cupertino

Research: Computational pathology, medical image analysis

Deep learning, machine learning,

Image processing, signal processing

Education: IIT Delhi, B Tech in EE

UIUC, PhD in ECE

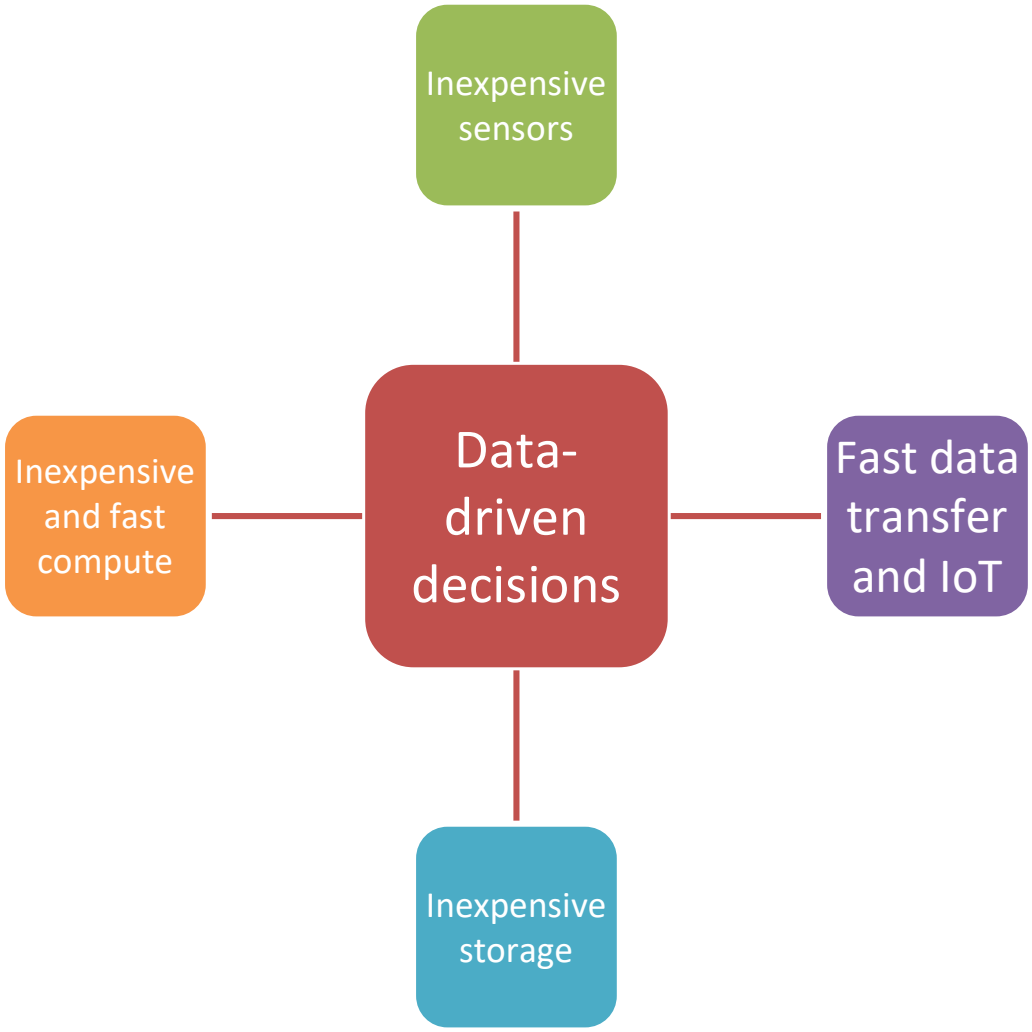
Sample research

Reverse Knowledge Distillation: Training a Large Model using a Small One for Retinal Image Matching on Limited Data SA Nasser, N Gupte, A Sethi Proceedings of the IEEE/CVF Winter Conference on Applications of Computer ...	1	2024
WaveMixSR: Resource-Efficient Neural Network for Image Super-Resolution P Jeevan, A Srinidhi, P Prathiba, A Sethi Proceedings of the IEEE/CVF Winter Conference on Applications of Computer ...		2024
Utilizing Radiomic Feature Analysis For Automated MRI Keypoint Detection: Enhancing Graph Applications SA Nasser, S Pathak, K Singhal, M Meena, N Gupte, A Chinmaya, P Garg, ... arXiv preprint arXiv:2311.18281		2023
Classification of Various Types of Damages in Honeycomb Composite Sandwich Structures using Guided Wave Structural Health Monitoring S Sawant, J Thalapil, S Tallur, S Banerjee, A Sethi arXiv preprint arXiv:2311.03765		2023
Utilizing Radiomic Feature Analysis For Automated MRI Keypoint Detection: Enhancing Graph Applications S Almahfouz Nasser, S Pathak, K Singhal, M Meena, N Gupte, ... arXiv e-prints, arXiv: 2311.18281		2023
Leveraging Segmentation to Improve Medical Image Registration SA Nasser, M Meena, G Sresth, A Sethi Authorea Preprints		2023
Combining Datasets with Different Label Sets for Improved Nucleus Segmentation and Classification A Parulekar, U Kanwat, RK Gupta, M Chippa, T Jacob, T Bameta, S Rane, ... arXiv preprint arXiv:2310.03346		2023
Domain-Adaptive Learning: Unsupervised Adaptation for Histology Images with Improved Loss Function Combination RK Gupta, S Das, A Sethi arXiv preprint arXiv:2309.17172		2023

Sample research

Transforming Breast Cancer Diagnosis: Towards Real-Time Ultrasound to Mammogram Conversion for Cost-Effective Diagnosis SA Nasser, A Sharma, A Saraf, AM Parulekar, P Haria, A Sethi arXiv preprint arXiv:2308.05449		2023
Heterogeneous graphs model spatial relationships between biological entities for breast cancer diagnosis A Krishna K, R Kant Gupta, N Cherian Kurian, P Jeevan, A Sethi arXiv e-prints, arXiv: 2307.08132		2023
WavePaint: Resource-efficient Token-mixer for Self-supervised Inpainting P Jeevan, DS Kumar, A Sethi arXiv preprint arXiv:2307.00407	1	2023
Leveraging Segmentation to Improve Medical Image Registration S Almahfouz Nasser, M Meena, G Sresth, A Sethi TechRxiv		2023
The ACROBAT 2022 Challenge: Automatic Registration Of Breast Cancer Tissue P Weitz, M Valkonen, L Solorzano, C Carr, K Kartasalo, C Boissin, ... arXiv preprint arXiv:2305.18033	2	2023
Multiscale deep learning framework captures systemic immune features in lymph nodes predictive of triple negative breast cancer outcome in large-scale studies G Verghese, M Li, F Liu, A Lohan, NC Kurian, S Meena, P Gazinska, ... The Journal of Pathology	3	2023
Quantification of subtype purity in Luminal A breast cancer predicts clinical characteristics and survival N Kumar, PH Gann, SM McGregor, A Sethi Breast Cancer Research and Treatment, 1-11	1	2023
EP178 ARTIFICIAL INTELLIGENCE BASED EOSINOPHIL COUNT IN GASTROINTESTINAL TRACT BIOPSY HC Shah, AD Amarpurkar, T Jacob, AM Parulekar, A Sethi Gastroenterology 164 (6), S-1229		2023
CHATTY: Coupled Holistic Adversarial Transport Terms with Yield for Unsupervised Domain Adaptation M Wagle, RK Gupta, A Sethi	1	2023

Why this course



- Lots of data generated and stored
- Data-driven decisions lead to better outcomes
- Same story across industries:
  - finance
  - healthcare
  - retail,
  - industrial automation,
  - environment and climate monitoring,
  - power,
  - automobiles,
  - …
- Handling data and programming are now a basic engineering skills

What is data science

- Extract useful insights from data that can be
  - Large in volume
  - Structured or unstructured
  - Captured and stored in different formats
- Using any of the following disciplines
  - Scientific method and statistics
  - Data mining and data visualization
  - Machine learning and deep learning
  - Algorithms, programming, and parallel processing

Types of data analysis

- Exploratory: what can be said about this data?
- Descriptive: does the data answer a question?
- Predictive: does the help predict something?
- Prescriptive: how can the data help us meet an objective?

Example 1: Increase company revenue

- Exploratory:
  - What data do we have on our customers?
  - Are there gaps in that data, e.g. certain seasons?
- Descriptive:
  - Do women really spend more than men?
  - Which age group spends the most?
- Predictive:
  - Can we predict when customers is ready to skip coming to our stores?
- Prescriptive:
  - Will a well-designed coupon campaign increase customer loyalty and sales?

Example 2: Increase car fuel efficiency

- Exploratory:
  - What data do we have about our car?
  - What data do we have about our customers and their driving habits?
- Descriptive:
  - What phase of the performance is crucial for sales?
- Predictive:
  - Can we predict among new fuel system design choices which ones will meet the performance and efficiency objectives?
- Prescriptive:
  - Which fuel system design choice will keep performance customers happy while fending criticism about our efficiency?

Types of data

- Structured:
  - Records with fields
  - Ordered on a grid
    - Time series
    - Images, videos, audio
    - Text
  - Relational
    - Graphs
- Unstructured

- Variables:
  - Nominal
  - Ordinal
  - Continuous
    - Floating point
    - Quantized continuous

Why make machines learn?

$Input_i$ → **Model** → $Output_i$

- We need models (functions, algorithms) to make predictions about inputs
- Many models are unknown and difficult to define
- Machine learning is the art and science of iteratively adjusting models based on inputs and some properties of the output

Some recent success stories

Autonomous driving

- Road recognition
- Automatic navigation

Speech recognition

- Speech to text
- Automated services over the phone

Face detection

- Facebook face tagging suggestions
- Camera autofocus for portraits

ML gives a model

- Elements of a model:
  - Input x_i
  - Function $f_\theta(x_i)$
- The model has to be useful:
  - Some notion of ideal output y_i
  - Loss $L(y_i, f_\theta(x_i), \theta)$

Good examples of ML problems

- Is a given face image of a male or female?
- Is there a Coke bottle in a given image?
- Is this image artistic?
- How can this text be improved or question answered?
- Is this customer likely to spend more later if we give her a discount now?
- Can I represent a 50-d data using 2-d?
- Can I divide my customers into logical groups?
- Can I generate music that sounds like Mozart?
- Inverse problem: source separation, super-resolution

Bad examples of ML problems

- Predict the next lunar eclipse
- Understand this text
- Should I pursue a PhD or not

Performance criteria and requirements

Performance criteria

- High accuracy
- Low risk
- More explanability
- Less memory
- Less compute

Requirements

- Good amount of data
- Clean and organized data
- Data labels and annotations
- Computational power

Learning outcomes for the course

- Translate real-world problem statements into different types of data analysis problems
- Define various data science tasks
- Demonstrate beginner-level proficiency in setting up all parts of a data analysis pipelines
- Define machine learning and cast ML problems
- Define and code popular ML algorithms
- Critique and compare ML algorithms and models

List of Topics

- Intro to random variables
- Intro to databases
- Exploratory data analysis
- Graphs and plots
- Statistical testing
- ML as a black box
- Linear regression
- Linear classification
- Regularization
- Kernelized classification
- Feature engineering
- Neural networks
- Deep neural networks
- Clustering
- Dimension reduction
- Density estimation

Prerequisites

- Basic linear algebra
  - Matrix-vector products, dot products, eigen vector definition, norms, …
- Intermediate probability
  - Continuous random variable, PDF, conditional distribution, marginalization
- Basic calculus
  - Derivatives, partial derivatives, critical points of a function
- Intermediate programming in python
  - Loops, functions, arrays, i/o, file i/o, graph plotting

Course eligibility

EE 353

- Core for EE juniors (BTech and DD)
- No other students allowed

EE 769

- Priority given to EE PGs
- Other students may be allowed, if their advisors recommend

Ineligible:

- First year and second year undergrads are not allowed
- CSE, IEOR, ME, CSRE students (except guide's reco)
- Any student who has done or is doing an equivalent (e.g. CS 419, 337, 725, 747, GNR 652, IE 506, 611, ME 781...) or advanced IITB course on ML (e.g. CS 726, 729, IE 643, 663, 712, GNR 638, EE 782...)

Tentative evaluation plan

Item	Wt.
Participation	10
Assignments (4)	40
Mid-sem exam	20
End-sem exam	30
Total	**100**

Min. marks	Grade
90	AA
80	AB
70	BB
60	BC
50	CC
40	CD
30	DD/AU

Evaluation items

- Precise understanding of concepts
- Express concepts mathematically
- Make basic mathematical derivations
- Program diligently
- Design experiments diligently
- Interpret results
- **Zero tolerance for academic malpractice**

Allowed vs. not-allowed

Allowed

- Exams
  - Notes
- Assignments
  - Discuss the assignment before starting a portion with friends
  - Consult code on the internet or LLMs
  - Disclosing sources of “inspiration,” indicating the exact lines copied and modified

Not allowed

- Exam
  - Open the internet
  - Communicate with others
- Assignments
  - Copy code from friend or internet and make trivial changes
  - Not disclosing sources of “inspiration” and the exact lines copied

How to get the most out of this course

- Attend lectures and take notes
- Read books and internet resources
- Attempt problems
- Discuss with classmates offline and online
- Become comfortable with programming
- Ask TAs and instructors any remaining doubts

Resources

- People:
  - Instructor: Amit Sethi (Chat on MS Teams, tag in channel)
  - TAs: TBA
- Office hours (Wednesdays 6pm to 7pm):
  - Amit Sethi's office, easternmost on second floor, EE building (7483)
  - MeDAL lab, first floor, EE building (3528)
- LMS
  - MS Teams code **yggbq3k** (Official channel for announcements and material, General channel for discussions and queries)
  - Moodle for assignment submissions and grades
- Books
  - Pattern Recognition and Machine Learning, by Christopher Bishop

Emergency absence policy

- Doctor’s note or documentation essential
- Assignments: extension of deadline
- Exams: Extrapolation of your performance based on your percentile (not percent)

Resume verification

- Sorry, I cannot verify resumes for such a large class
- Make a GitHub Repo, and upload your assignment and project reports on ArXiv for the recruiters to see

Immediate tasks

- **Python:**
  - https://www.learnpython.org/
- **Numpy and Google CoLab: (Basic python, numpy, pandas, matplotlib, scikit-learn)**
  - https://cs231n.github.io/python-numpy-tutorial/
- **Linear algebra and calculus:**
  - https://stanford.edu/~shervine/teaching/cs-229/refresher-algebra-calculus
- **Probability:**
  - https://stanford.edu/~shervine/teaching/cs-229/refresher-probabilities-statistics