
भारतीय प्रौद्योगिकी संस्थान जम्मू
INDIAN INSTITUTE OF TECHNOLOGY JAMMU

# Machine Learning (CSC033U3M)

Dr. Shaifu Gupta
shaifu.gupta@iitjammu.ac.in

# Content for the course

Data Preprocessing. Evaluation metrics. Supervised learning algorithms: Linear and Logistic Regression, Gradient Descent, Support Vector Machines, Kernels, Artificial Neural Networks, Decision Trees, ML and MAP Estimates, K-Nearest Neighbor, Naive Bayes, Introduction to Bayesian Networks. Unsupervised learning algorithms: K-Means clustering, Gaussian Mixture Models, Expectation Maximization. Dimensionality Reduction and Principal Component Analysis. Bias Variance Trade-off. Model Selection and Feature Selection. Regularization. Applications. Advanced Topics.

# Course İnformation

Course structure: 3:0:2

Prerequisite: COL 773
(Python, **probability and statistics**, linear algebra, optimization)

Revision will be helpful!

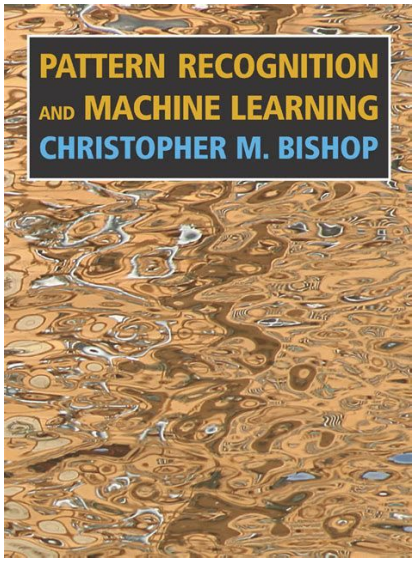
# Labs/ mini projects

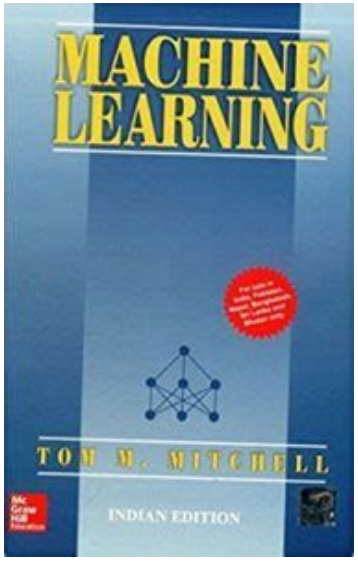
Linear and Logistic Regression, Support Vector Machines, Artificial Neural Networks, Decision Trees, K-Nearest Neighbor, Bayesian models, K-Means clustering, Gaussian Mixture Models, Principal Component Analysis etc.

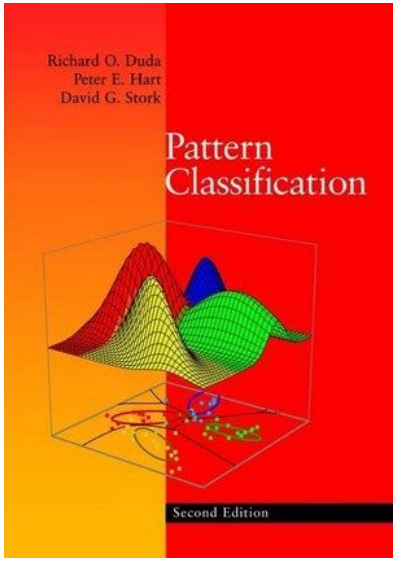
Write the code yourself (rather than using inbuilt libraries)!!

| Mode of Evaluation | % Weight |
|---|---|
| Class Test 1 | 15 |
| Mid Sem | 25 |
| Class Test 2 | 15 |
| End Sem | 25 |
| Lab Evaluation | 20 |
| Total % | 100 |

# Reference material




Other reference materials will also be shared from time to time!

# INTRODUCTION

# What is machine learning?

Herbert Simon (1970)

Any process by which a system improves its performance

Tom Mitchell (1990)

A computer program that improves its performance at some task through experience

Wikipedia

Machine learning (ML) is the study of computer algorithms that improve automatically through experience - by the use of data.

# Big Data

- Widespread use of personal computers, social networks, web search etc.. leads to generation of “big data”
- We are both producers and consumers of data
- Data is not random, it has structure, e.g., customer behavior
- We need mechanism to extract that structure from data for

  (a) Understanding the process

  (b) Making predictions for the future

Example in retail: Customer transactions to consumer behavior:

People who bought “Blink” also bought “Outliers” [*books by Malcolm Gladwell] (www.amazon.com)

# Data Mining

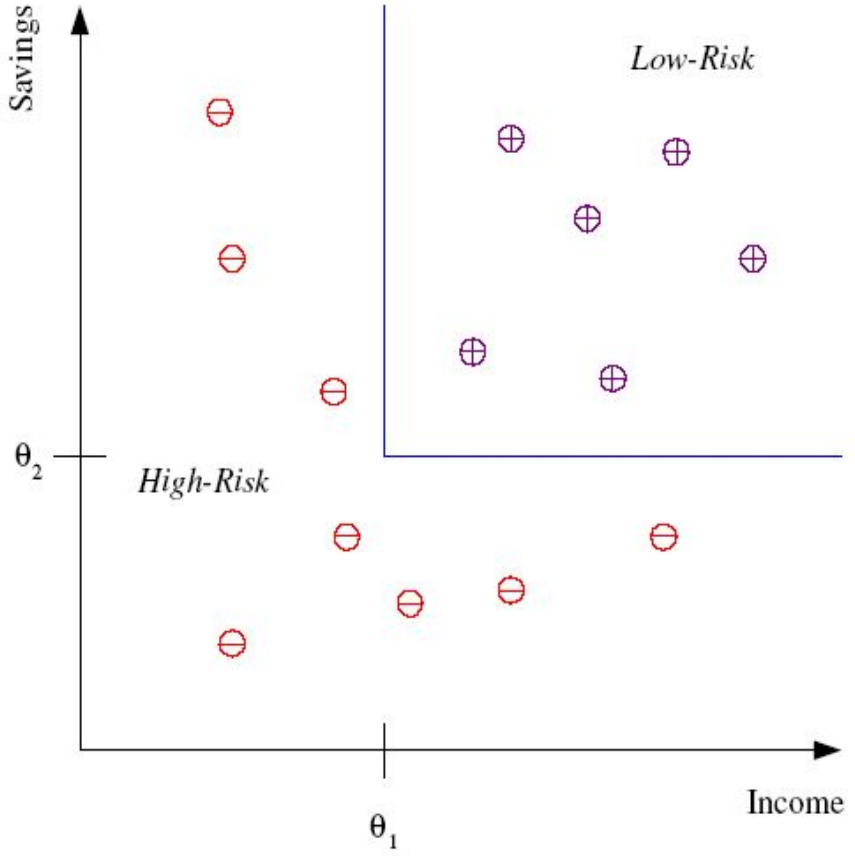
- Retail: Market basket analysis, Customer relationship management (CRM)
- Finance: Credit scoring, fraud detection
- Manufacturing: Control, robotics, troubleshooting
- Medicine: Medical diagnosis
- Telecommunications: Spam filters, intrusion detection
- Bioinformatics: Motifs, alignment
- Web mining: Search engines
- …

# Learning paradigms

- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning

# Classification

- Example: Credit scoring (financial metric used by money lenders)
- Differentiating between low-risk and high-risk customers from their *income* and *savings*




Discriminant: IF *income* > $\theta_1$ AND *savings* > $\theta_2$
THEN low-risk ELSE high-risk

# Classification: Applications

- Face recognition: Pose, lighting, occlusion (glasses, beard), makeup, hair style
- Character recognition: Different handwriting styles.
- Speech recognition: Temporal dependency.
- Medical diagnosis: From symptoms to illnesses
- Biometrics: Recognition/authentication using physical and/or behavioral characteristics: Face, iris, signature, etc
- Outlier/novelty detection
- Fault diagnosis

# Face Recognition

## Training examples of a person

## Test images

ORL dataset,
AT&T Laboratories, Cambridge UK

# Regression

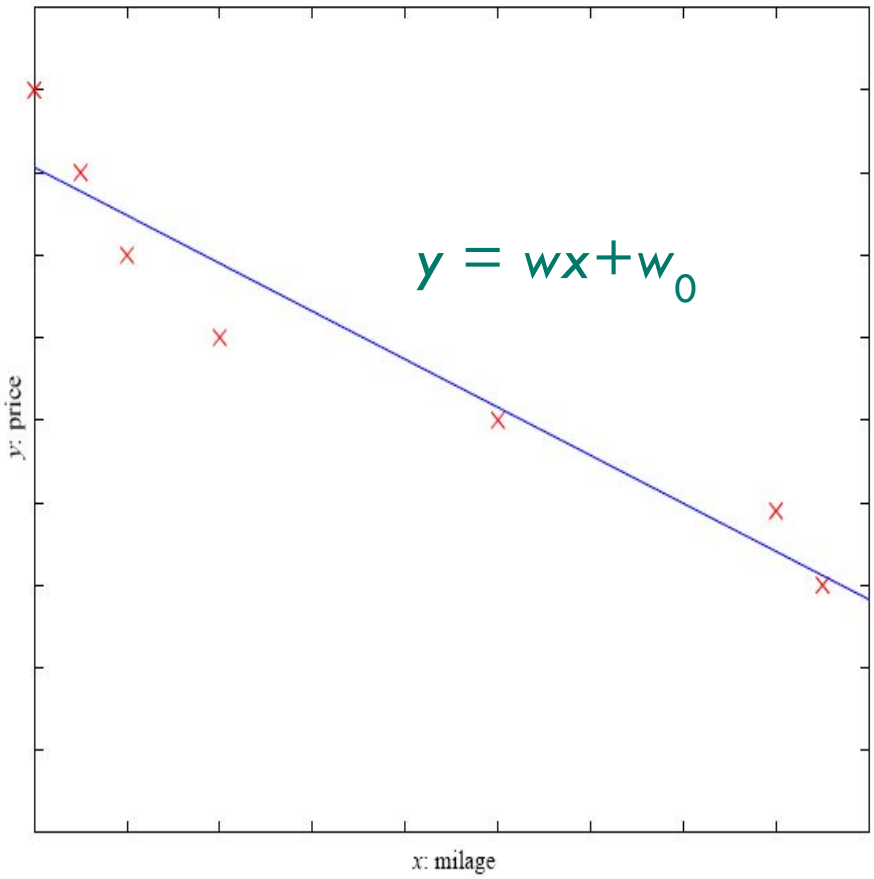
- Example: Price of a used car
- $x$ : car attributes

  $y$ : price

  $y = g(x \mid \theta)$

  $g(\ )$ model,

  $\theta$ parameters



# Regression applications

- Predict water given to soil based on weather conditions
- Predict future resource usage of server/data center
- Predict weather conditions (temperature, humidity etc. )
- Predict mobility of public based on active covid cases
- Predict sales based on season, customer interest, quality of product etc.

# Unsupervised Learning

- Clustering: Grouping similar instances
- Example applications
  - Customer segmentation in CRM
  - Grouping sensors in an organization
  - Grouping of movies based on reviews, genre etc.

# Other learning paradigms (Meta learning, EXplainable AI)

- Transfer learning
  - Transfer of knowledge between multiple domains
- Active learning
  - Learning algorithms interactively queries an oracle to obtain desired outputs for new data
- Online learning
  - Learning on the fly
  - zero shot learning
- Representation learning
  - Learning representation from data
  - Embeddings
- Reinforcement learning
  - Learn to act in an environment
  - Actions: rewards and penalties

# Machine learning challenges

- Curse of dimensionality
- Sample size
- Choosing algorithm
- Too many hyper-parameters to tweak
- Choosing right values of parameters to learn

# Resources: Datasets

- UCI Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html
- Statlib: http://lib.stat.cmu.edu/