Lecture 1: Introduction

Madhavan Mukund

https://www.cmi.ac.in/~madhavan

Data Mining and Machine Learning
August–December 2020

What is this course about?

Data Mining

- Identify "hidden" patterns in data
- Also data collection, cleaning, uniformization, storage
  - Won't emphasize these aspects

Machine Learning

- "Learn" mathematical models of processes from data
- Supervised learning — learn from experience
- Unsupervised learning — search for structure

Supervised Learning

Extrapolate from historical data

- Predict board exam scores from model exams
- Should this loan application be granted?
- Do these symptoms indicate CoViD-19?

“Manually” labelled historical data is available

- Past exam scores: model exams and board exam
- Customer profiles: age, income, . . . , repayment/default status
- Patient health records, diagnosis

Historical data → model to predict outcome

Supervised learning . . .

What are we trying to predict?

Numerical values

- Board exam scores
- House price (valuation for insurance)
- Net worth of a person (for loan eligibility)

Categories

- Email: is this message junk?
- Insurance claim: pay out, or check for fraud?
- Credit card approval: reject, normal, premium

Supervised learning . . .

How do we predict?

- Build a mathematical model
  - Different types of models
  - Parameters to be tuned
- Fit parameters based on input data
  - Different historical data produces different models
  - e.g., each user's junk mail filter fits their individual preferences
- Study different models, how they are built from historical data

Unsupervised learning

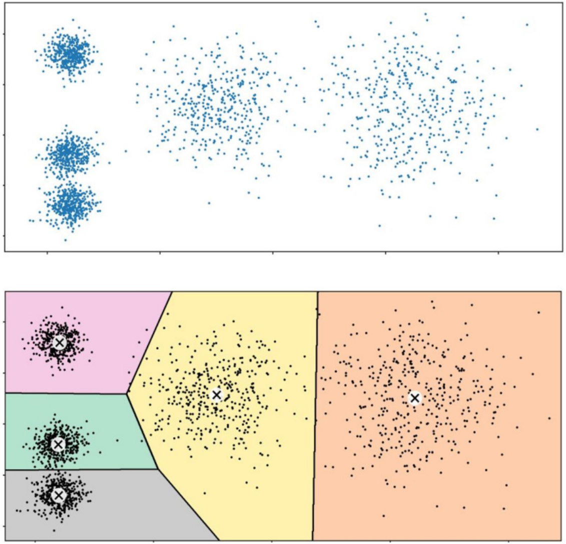
- Supervised learning builds models to reconstruct “known” patterns given by historical data
- Unsupervised learning tries to identify patterns without guidance

Customer segmentation

- Different types of newspaper readers
- Age vs product profile of retail shop customers
- Viewer recommendations on video platform

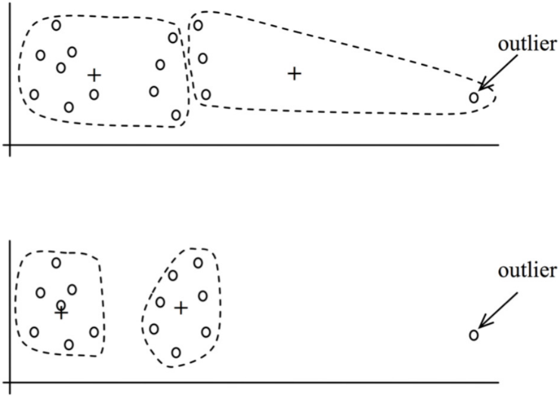
Clustering

- Organize data into “similar” groups — clusters
- Define a similarity measure, or distance function
- Clusters are groups of data items that are “close together”



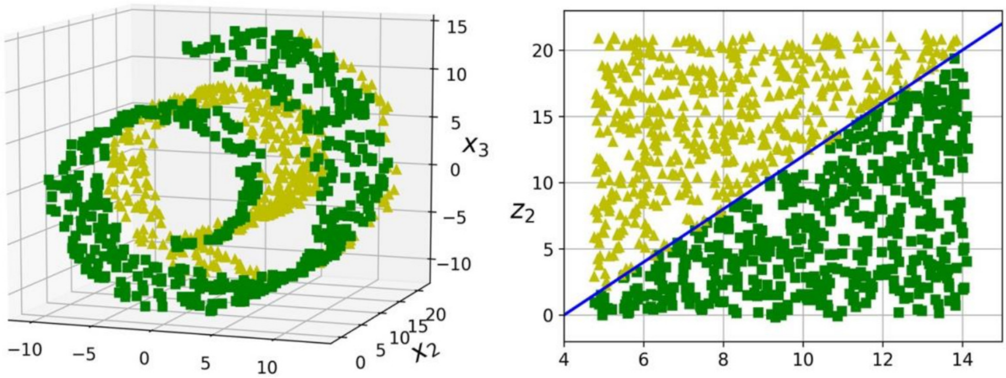
Outliers

- Outliers are anomalous values
  - Net worth of Bill Gates, Mukesh Ambani
- Outliers distort clustering and other analysis
- How can we identify outliers?



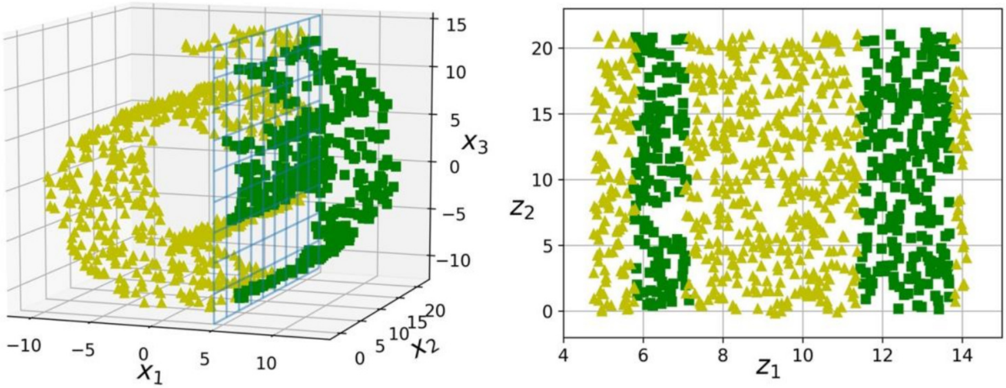
Preprocessing for supervised learning

Dimensionality reduction



Preprocessing for supervised learning

Need not be a good idea — perils of working blind!



Summary

Machine Learning

- Supervised learning
  - Build predictive models from historical data
- Unsupervised learning
  - Search for structure
  - Clustering, outlier detection, dimensionality reduction

If intelligence were a cake, unsupervised learning would be the cake, supervised learning would be the icing on the cake, . . .

Yann Le Cun, ACM Turing Award 2018