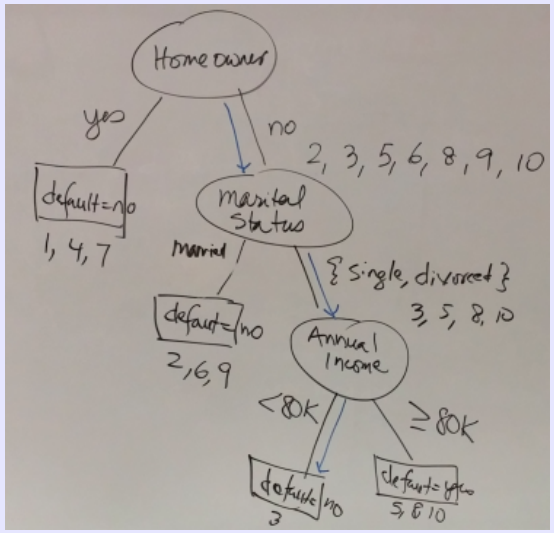
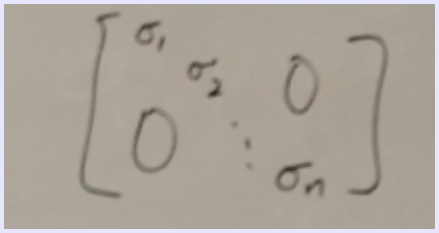
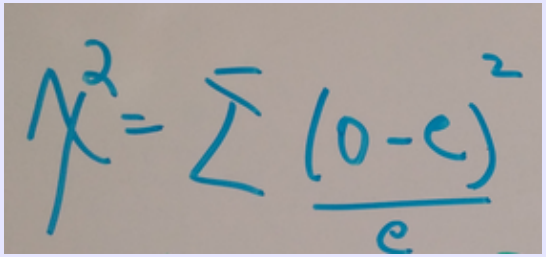
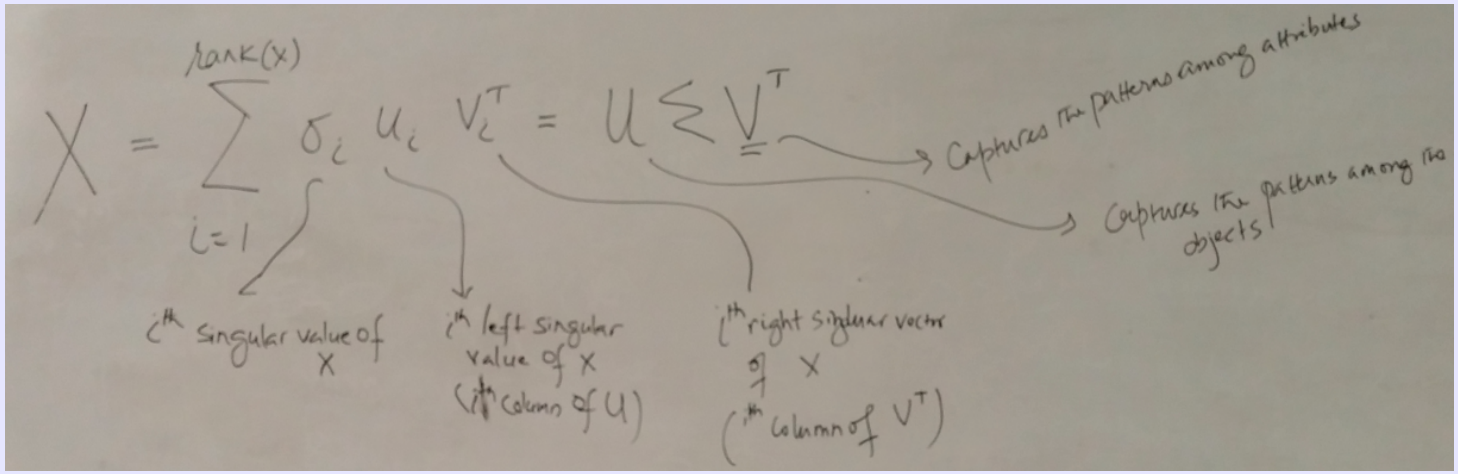
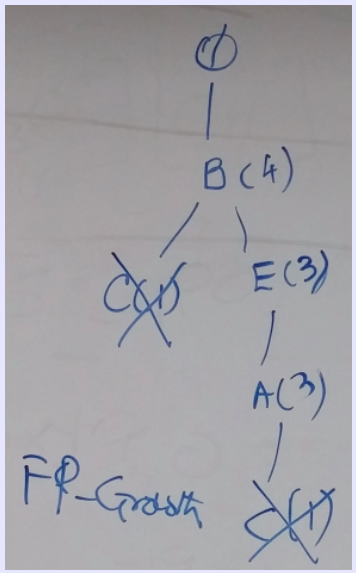
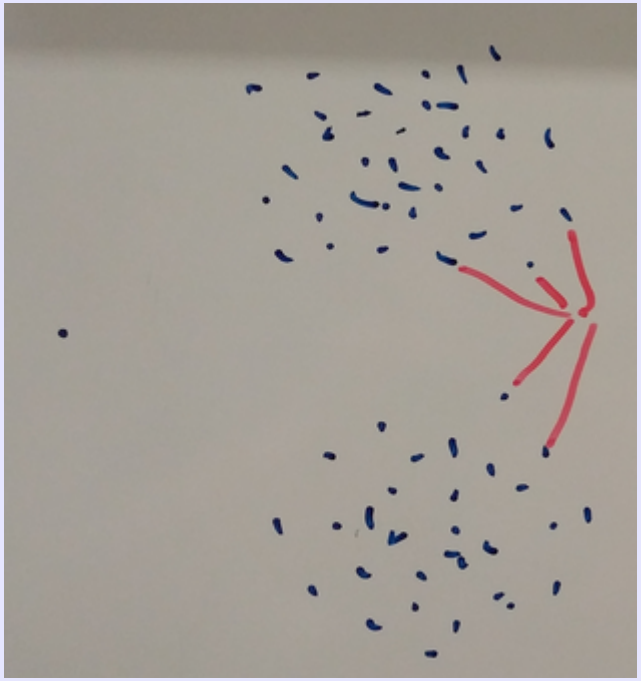












CS 422: Data Mining
Vijay K. Gurbani, Ph.D.,
Illinois Institute of Technology

Lecture 1: **Introduction**

Introduction: Back to data mining

Is Data Mining a new field?

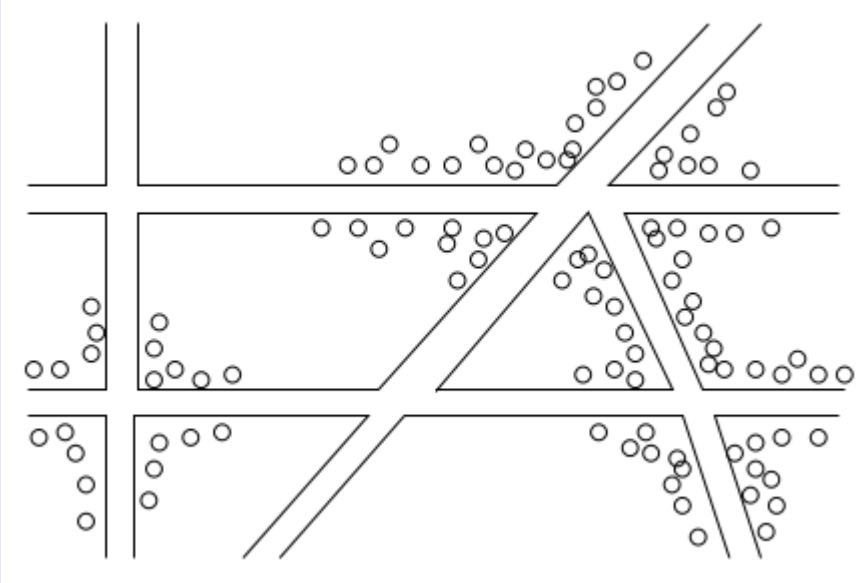


Figure source: Mining of Massive Datasets, Leskovec et al., 2014.

Introduction: Back to data mining

- Data mining vs. machine learning vs. …
  - Data mining: a cross-disciplinary field focused on discovering properties (patterns) of (large, very large) data sets.
  - How does it do this?
    - Machine learning is one approach.

Introduction: Back to data mining

- Machine learning is all around us.
  - Netflix
  - Google
  - Amazon, ...
- You've all heard the fantastic stories of self-driving cars, mood-sensing environment, why diapers and beer go together, what Target knows about you, and so on ?

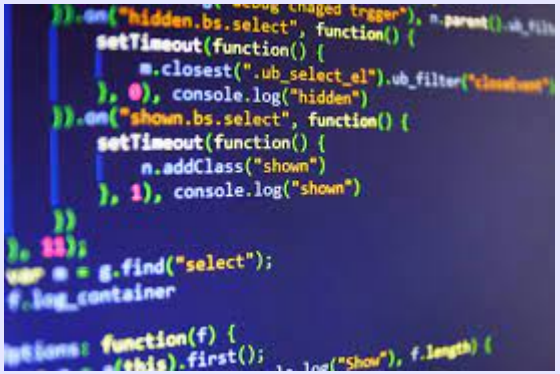
Data mining and machine learning

Computer Science

- Determinism rules.

Machine Learning

- Generalization is key.



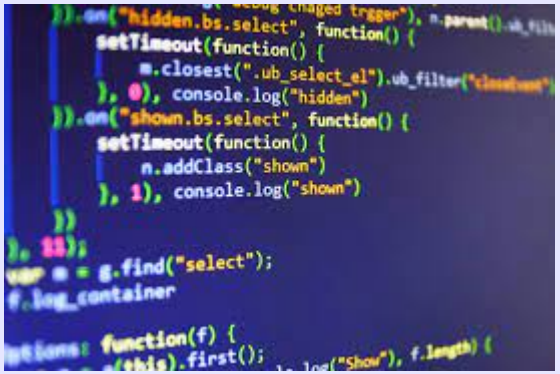
Data mining and machine learning

Computer Science

- Determinism rules.
- Errors not tolerated.

Machine Learning

- Generalization is key.
- Errors part of the landscape.



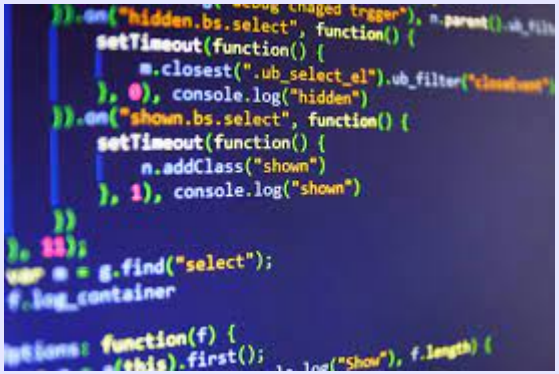
Data mining and machine learning

Computer Science

- Determinism rules.
- Errors not tolerated.
- Algorithms do not learn.

Machine Learning

- Generalization is key.
- Errors part of the landscape.
- Algorithms learn (backprop, genetic programming).



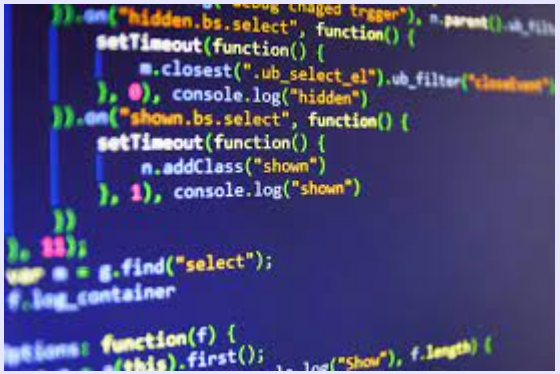
Data mining and machine learning

Computer Science

- Determinism rules.
- Errors not tolerated.
- Algorithms do not learn.
- Program(Data) => Output. Program most important artifact.

Machine Learning

- Generalization is key.
- Errors part of the landscape.
- Algorithms learn (backprop, genetic programming).
- Data(Program) => Model => Ouput. Data most important artifact.



Data mining and machine learning

- The Machine Learning Problem: *Generalizing to cases we have not seen before.*
- But, can’t we simply see all or most of the data?

Data mining and machine learning

- The Machine Learning Problem: *Generalizing to cases we have not seen before.*
- But, can’t we simply see all or most of the data?
- Suppose: You have data that consists of 1,000 Boolean fields, and you have 1,000,000,000,000 records in a database.
- How much insight do these 1 trillion records represent?

Data mining and machine learning

- The Machine Learning Problem: *Generalizing to cases we have not seen before.*
- But, can't we simply see all or most of the data?
- Suppose: You have data that consists of 1,000 Boolean fields, and you have 1,000,000,000,000 records in a database.
- How much insight do these 1 trillion records represent?
- Theoretically, you will need 2^1000 records to represent all of your data!!
- How big is 2^1000?
  - 2^1000 ≈ 1.07^301 ≈ 10^301.
  - 10^18 grains of sand in the world.
  - 10^22 stars in the observable universe.
  - 10^78 atoms in the observable universe.

- The 1 trillions records are one "gazillionth*" of 1 percent of 2^1000!
  * Gazillionth = 10e-285
- Morals:
  - Curse of dimensionality is real
  - Generalization is how we deal with comb-inatorial explosion!

Resources

- Conferences in data mining and machine learning (non-exhaustive)
  - ACM KDD (Knowledge Discovery and Data Mining), http://www.kdd.org
  - ICML (International Conference on Machine Learning)
  - ACM CIKM (International Conference on Information and Knowledge Management)
  - SDM (SIAM International Conference on Data Mining)
  - NeurIPS
- Journals (non-exhaustive)
  - IEEE Transactions on Pattern Analysis and Machine Intelligence
  - ACM Transactions on Knowledge Discovery from Data
  - IEEE Transactions on Knowledge and Data Engineering

Resources

- Useful general Internet resources on data mining and machine learning:
  - Kaggle (https://www.kaggle.com)
  - Kdnuggets (https://www.kdnuggets.com)
  - https://machinelearningmastery.com
  - https://towardsdatascience.com