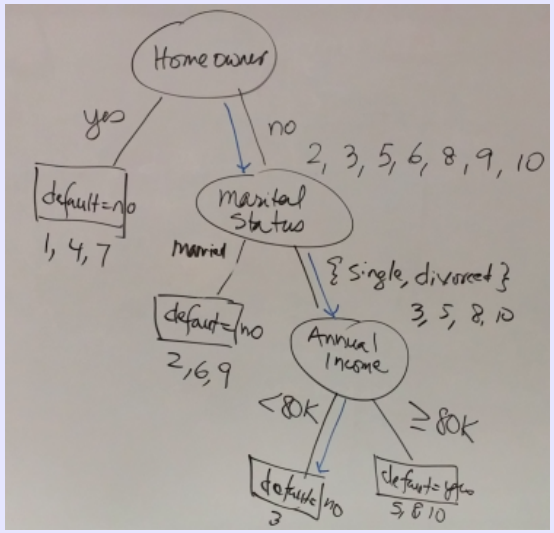

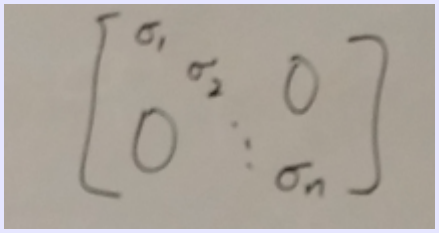

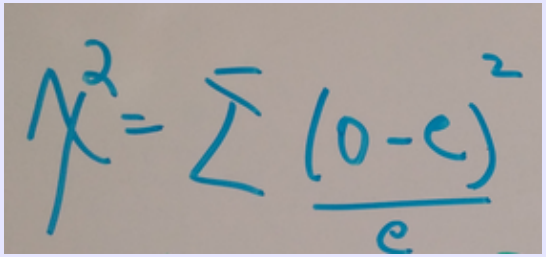

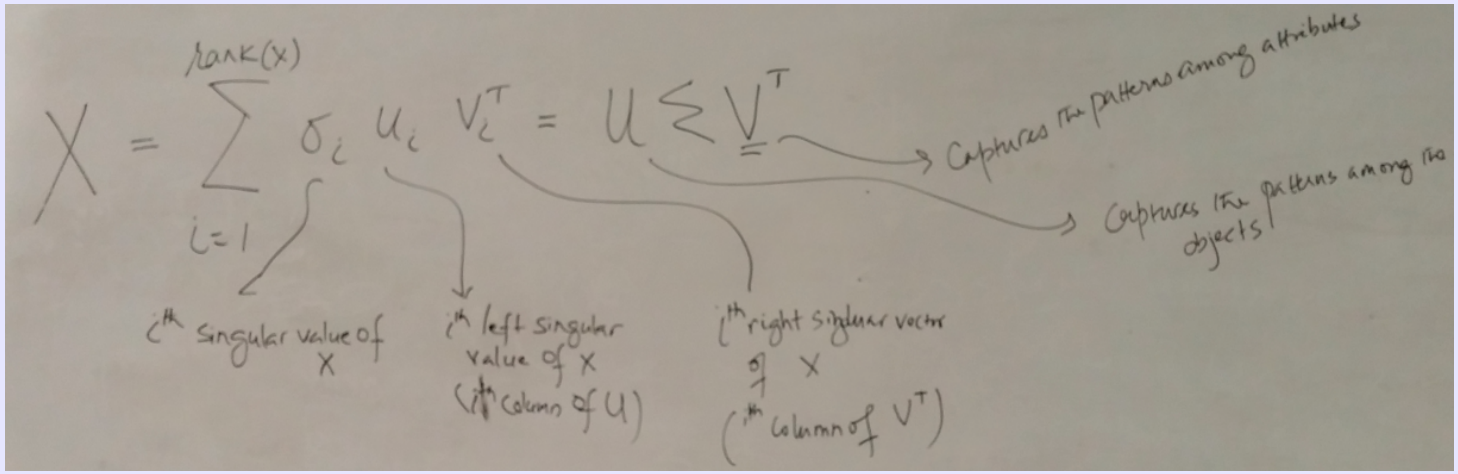

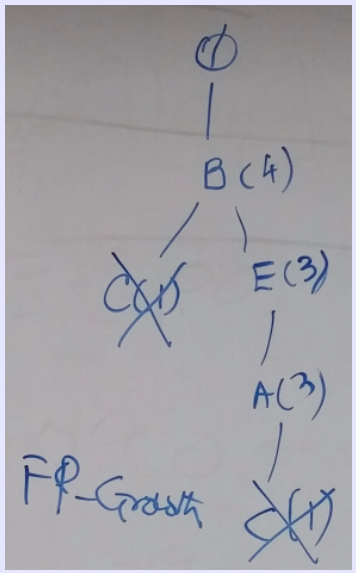

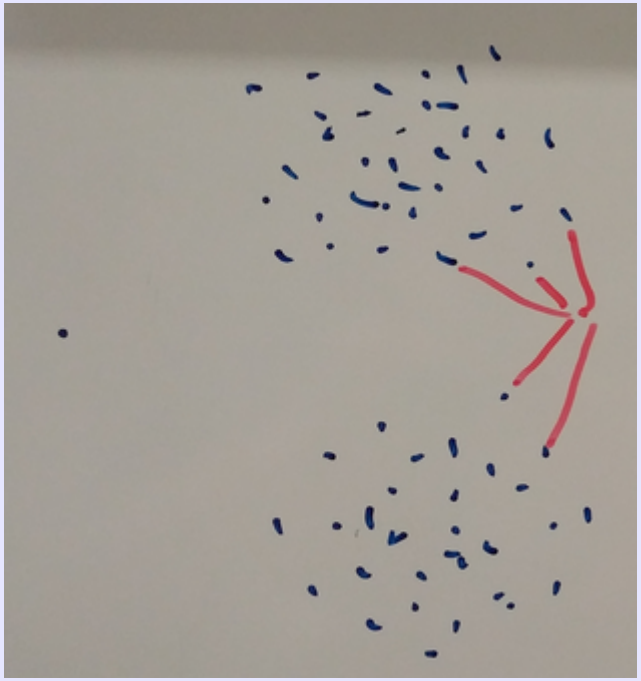

CS 422: Data Mining
Vijay K. Gurbani, Ph.D.,
Illinois Institute of Technology

Lecture 1: **Introduction**

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).

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).
- 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 2e1000 records to represent all of your data!!
- The 1 trillions records are one "gazillionth*" of 1 percent of 2e1000!

* Gazillionth = 10e-285

Morals:
- Curse of dimensionality is real
- Generalization is how we deal with combinatorial 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)
- 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/

Resources

- R Tutorials
  - **http://www.cookbook-r.com/**
  - **https://github.com/matloff/fasteR**
  - http://www.r-tutor.com/r-introduction
  - https://cran.r-project.org/doc/manuals/R-intro.pdf
- R Programming Tutorial, Derek Banas https://www.youtube.com/watch?v=s3FozVfd7q4
- Recitation on R today by TA.
- **There will be a short quiz on R soon!**

- R Development Environment:
  - R https://www.r-project.org/
  - RStudio https://www.rstudio.com
  - RStudio Cloud https://rstudio.cloud/
  - R@Illinois Tech

    ~~Available on the Fusion cluster: https://linux1.cs.iit.edu:8787 Use your normal IIT login/ password for authentication.)~~