Fall 2022 ECEN 748
Data Mining & Analysis

Nick Duffield

Department of Electrical & Computer Engineering

Texas A&M Institute of Data Science

Organization

- **Instructor:** Nick Duffield
- **Contact:** duffieldng@tamu.edu; (979) 845-7328
- **Class notes:** Canvas
- **Homework:** Canvas
- **Class times:** MW, 5:45-7:00pm, BLOC 166
- **Office hours:** MW, 4:15-5:15pm, BLOC 227F

Grading

- Components of the grade
  - Assignments: 60%, roughly every two weeks
  - Class Test: 15%, Wed, October 19, 2022, 5:45 p.m. – 7:00 p.m, BLOC 166
  - Final Exam: 25%, Fri, December 9, 2022, 7:30 a.m. - 9:30 a.m.
- Discussion of homework assignments is encouraged
- Homework must be executed independently, copying not allowed.
- Assignments must be typeset and submitted on time to receive full credit.
- No late assignments will receive full credit unless justified by an official document (e.g., doctor's note)
- 10% penalty for each 24hr hour period

Course texts and materials

- **Primary text**
  - [ZM] Data Mining and Machine Learning: Fundamental Concepts and Algorithms, M. Zaki & W. Meira, Jr. Cambridge University Press, 2020, https://dataminingbook.info, Online and through the TAMU Library
- **Secondary text**
  - [MMDS] Mining of Massive Datasets, J. Leskovec, A. Rajaraman, & J, Ullman, http://mmds.org/
- **Assignments**
  - Selected examples based on problems in [ZM] and [MMDS]
    - Typically selecting, applying or implementing computation
      - Some algebra manipulations to reinforce view “under the hood”
    - Solutions include numerical or analytical results, code snippets, reasoning narrative

Course Description

- **Overview of data mining, integrating related concepts from machine learning and statistics.**
- **Fundamental topics including exploratory data analysis, pattern mining, clustering, classification, and regression, with applications to scientific and online data.**

Learning outcomes

- **Acquire knowledge of foundations and application of methods in data mining and data analysis.**
- **Prepare students to use methods and tools of data science in research, whether focused on methods or on applications.**
- **Upon the completion of the course, the student should be able to:**
  - Conduct exploratory data analysis including visualization and summarization
  - Apply selected unsupervised machine learning methods to data analytical problems
  - Apply selected supervised machine learning methods to data analytic problems
  - Select appropriate machine learning method applicable to common problem types
  - Understand usage rationale, underpinnings, and limitations of machine learning methods
  - Apply common libraries and tools to data analytical problems

Course Topics

Week	Topic	Required Reading
1	Data and Attributes	ZM Chapters 2, 3
2	Dimensionality Reduction	ZM Chapter 7
3	Frequent Itemset Mining & Association Rules	ZM Chapter 8, MMDS Chap. 6
4	Representative Clustering	MMDS, Chapter 7
5	Gaussian Mixture Clustering & EM Method	ZM Chapter 13
6	Hierarchical Clustering	ZM Chapter 14
7	Density Estimation & Density-Based Clustering	ZM Chapter 15
8	Bayesian & Nearest Neighbor Classification	ZM Chapter 18
9	Decision Tree Classification	ZM Chapter 19
10	Graphs, Pagerank & Search	MMDS Chapter 5
11	Recommendation Systems	MMDS Chapter 9
12	Linear and Logistic Regression	ZM Chapters 23 & 24
13	Support Vector Machines	MMDS Chapter 11, ZM Chapter 21
14	Perceptrons & Networks	MMDS Chapter 11, ZM Chapter 25
Class Test: Wednesday, October 19, 2022, 5:45 p.m. – 7:00 p.m. Final Exam: Friday, December 9, 2022, 7:30 a.m. - 9:30 a.m.		

About me

- **Joined Texas A&M in August 2014**
- **Worked for 18 years in AT&T Labs-Research in New Jersey**
  - Research in data science for communications networks
- **Previously Asst. Professor in Europe**
- **Undergrad/PhD in Physics and Mathematical Physics**
- **Research Interests:**
  - Measurement and analysis of communications networks
    - Measurements in software defined networks
    - Network security / cybersecurity
  - Data Science
    - Streaming algorithms
    - Statistical Inference, Machine Learning
    - Applications in Transportation, Urban Science Geosciences, Agriculture, …
- **Director of Texas A&M Institute of Data Science**
  - https://tamids.tamu.edu

A Quick Tour of the Course

A quick tour of the course

- **What is driving the recent growth and abundance of data?**
- **What does data look like?**
- **What questions are we trying to answer?**
- **Exploratory data analysis**
- **Dealing with high dimensional data**
- **Finding patterns in the data: clustering**
- **Learning and classification**

Drivers for recent rapid growth in data

- **Instrumentation of the internet & internet-based services**
  - Operational data from online social networks
    - Billions of users; trillions of connections;
  - Search logs at search providers such as Google
  - Internet Service Providers
    - Traffic measurements: #bytes, #packets between network endpoint pairs
  - Internet of Things
    - Huge increase in number of connected endpoints
  - Retail transactions
    - Online retailers, movie purchases

Drivers for recent rapid growth in data (2)

- **Increased Sensing in Science, Health, Engineering**
  - Satellite Imaging
    - Land use, elevation, slope, soil type & moisture, vegetation, radiance
  - Unmanned Aerial Vehicles
    - Agricultural imaging down to scale of individual plants
  - Weather & Climate
    - Wind velocity, precipitation, humidity, temperature, …
      - Crowdsourcing (Internet connected weather stations)
    - Ocean sensing: temperature, currents
  - Transportation
    - Location from apps, cell-towers, vehicle state from smart cars
  - Health & Medicine
    - Genetic sequencing, phenotypes, clinical records, treatments, outcomes, environment & economics, lifestyle
  - Energy and Utilities
    - Seismological data from oil exploration
    - Infrastructure sensors, household smart meters

General properties of data

- **Most data in this course is in form of a set of records**
  - E.g. Data matrix D:
    - Each row = record containing values of d of attributes
    - Such as measured values from n individuals in a survey
  - E.g. each record is a list of items
    - E.g. list of items purchased in a transaction

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

t	$\mathbf{i}(t)$
1	*ABDE*
2	*BCE*
3	*ABDE*
4	*ABCE*
5	*ABCDE*
6	*BCD*

Transaction Database

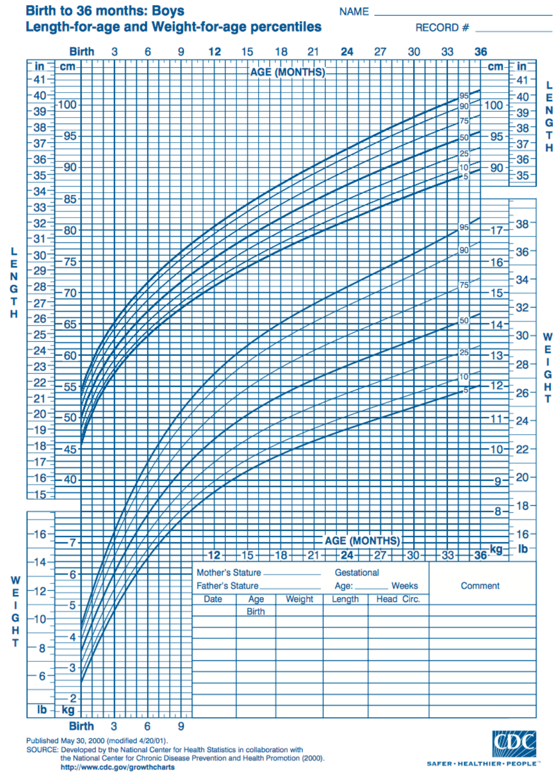
- **These are examples of structured data**
  - Organized data, e.g. in database with specified field formats
    - Information readily searched / retrieved / inserted
  - Contrast with **unstructured data**
    - E.g. audio files, images, videos, free text

TEXAS A&M UNIVERSITY
Department of Electrical
& Computer Engineering

TEXAS A&M
Institute of
Data Science

How to make model of the data?

- **Data records = (age, height, weight, …)**
  - from clinical surveys of children
- **What is relation between variables?**
  - Model distribution of height and weight as a fu
    - Used to make clinical growth charts
- **Exploratory data analysis**
  - For each age, characterize mean, variance, covariance amongst height, weight, …
  - Multivariate Gaussian model



Finding patterns in the data

- **Dataset = retail transactions, each listing items purchase**
  - These are called *itemsets*
- **What are the most frequently purchased items?**
- **Which items are frequently purchased together?**
  - Information used to advertizing, special offers, recommendations
  - If item A more often bought with item B that without, market together
- **How do we find all the *frequent itemsets* of a given size?**

The challenge of data size

- **Datasets may be inherently large**
  - Data matrix with large number n of records
    - See previous examples

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

- **Computations can have high possible size**
  - n**k possible itemsets of size k built from n objects
- **Abstract specification (“find frequent itemsets”) often needs careful implementation to compute efficiently**

Challenges of high data dimensionality

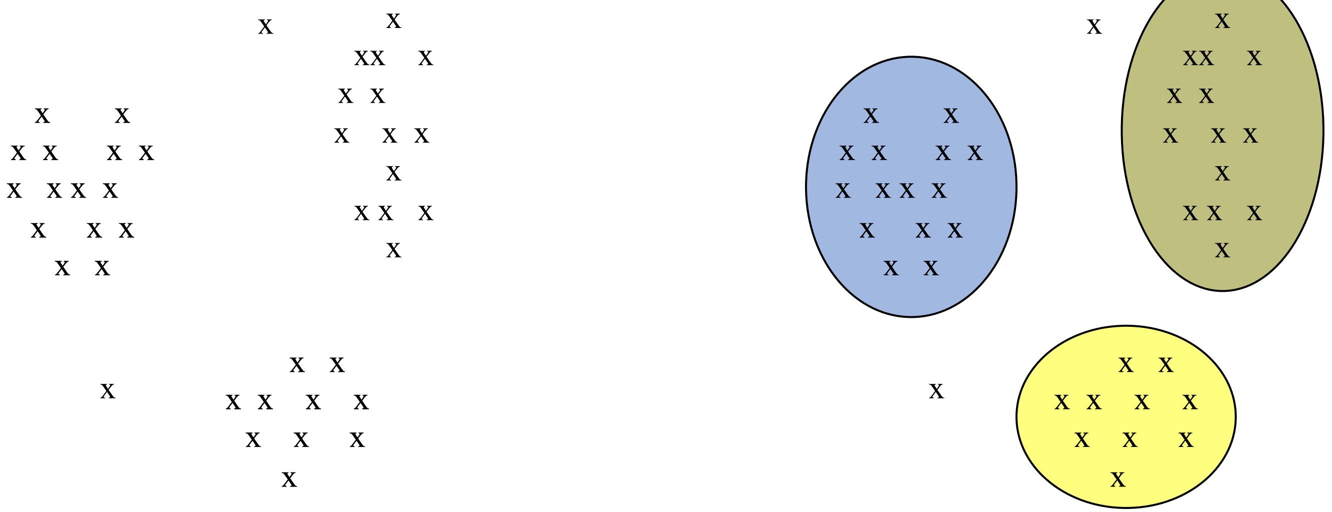
- **Data may have large number of attributes**
  - Data matrix with many columns d

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

  - Challenges
    - Models have large number of parameter
      - d(d-1)/2 variances / covariances in a Gaussian model
      - Cumbersome, more difficult to estimate well
  - Dimension reduction
    - Identify small number of attributes (or combinations of attributes) that account for most of the data variation
    - Other dimensions of the data are viewed as noise,

Clustering

- **In many data sets, the points evidently disjoin into groups**
  - Different subpopulations within data, each with own data characteristic



- **Want to automate identification of clusters in data**
- **Different notions of clustering**
  - Representative: cluster points close to a central representative
  - Density based: separated groups of contiguous points
  - Hierarchical clustering: grouping at multiple levels

Classification

- **Clustering is type of *unsupervised machine learning***
  - We have no side information on variables that may distinguish clusters
- **Classification**
  - For each data point **x** we are given a *class* y
  - Example:
    - data point **x** = (height, length) of plant leaf
    - class y = plant variety
  - Want to learn the relationship between the data values and the class
- **A *classifier* is a function that can be used to predict the class of *any* possible data point x**
- ***Supervised machine learning*:**
  - Compute a classifier from set of (data, class) values $\{(\mathbf{x}_i, y_i): i=1,\ldots,n\}$
- **Methods include:**
  - Bayesian, Support Vector Machines, Linear Perceptrons, Boosting

Graphical data mining

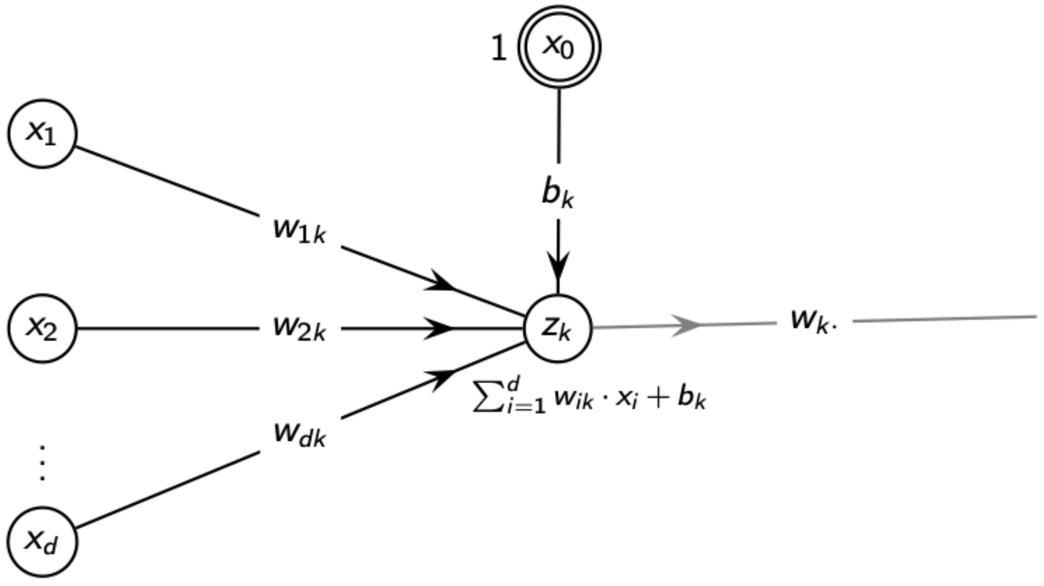
- **Many interesting dataset can be represented as graphs**
- **Example: web graph**
  - Vertex = web page
  - Directed edge (a,b) iff page a contains a hyperlink to page b
- **Web search problem: how to rank pages by importance**
  - Content based criteria:
    - is page content relevant for search, e.g. matching keywords
  - Topology based criteria:
    - Are there many hyperlinks to that page?
    - Google's PageRank algorithms based on random walk of web graph

Regression

- **Given variables X = $(X_1, X_2, \ldots, X_d)$**
  - Known as predictors, explanatory or independent variables
- **Response variable Y**
- **Can we learn a regression function f that predicts Y from X**
  - $Y = f(X_1, X_2, \ldots, X_d) + \varepsilon = f(X) + \varepsilon$
    - Here ε is random error term assumed independent of X
- **Linear regression**
  - $f(X) = \beta + \omega_1 X_1 + \ldots + \omega_d X_d = \beta + \omega . X$
- **Logistic regression**
  - Probability of binary outcome (Y = 1 or 0) follows logistic function
    - $\text{Prob}(Y = 0) = 1/(1+\exp(\omega . X))$
- **Learning problem**
  - Find the parameters $\beta, \omega_1, \ldots, \omega_d$ that best explain the obervations

Neural Networks and Learning

- **Machine Learning Inspired By Abstraction of Neural Systems**
- **From single neuron model**



- **To deep network models**