Introduction to Machine Learning

Ranga Raju Vatsavai, Ph.D.
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)

Feb. 25-27, 2019

What is Machine Learning?

- ML is a branch of AI that provide systems the ability to automatically learn and improve from experience without being explicitly programmed

What is Machine Learning?

- More formally: (Tom Mitchell)
- Study of algorithms that
  - improve their performance P
  - at some task T
  - with experience E
- So, a well-defined learning program includes: <P, T, E>

What is Data Mining?

- Many Definitions
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

NC STATE UNIVERSITY

What is data science?

- Data Science is the extraction of knowledge from large volumes of data that are structured or unstructured, which is a continuation of the field data mining and predictive analytics, also known as knowledge discovery and data mining (KDD). -- Wikipedia

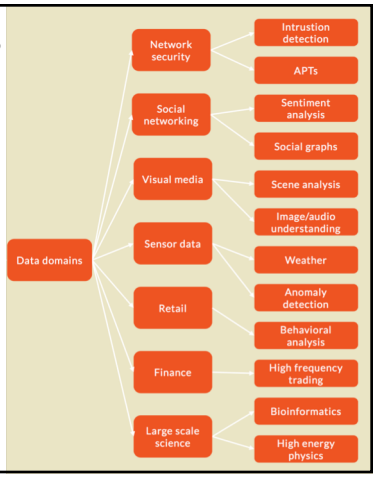
NC STATE UNIVERSITY

Data Taxonomy

- Structured (retail, financial, bioinformatics, geodata)
  - Structured data conforms to a database model, which is largely characterized by the various fields that data belongs to (name, address, age and so forth), and the data type for each field (numeric, currency, alphabetic, name, date, address).
- Unstructured (images, video, sensor data, web pages)
  - refers to information that either does not have a pre-defined data model or is not organized in a predefined manner
- Semi-structured (web logs, email, documents)
  - It is a type of structured data, but lacks a strict structure imposed by an underlying data model.

NC STATE UNIVERSITY

Data Sources

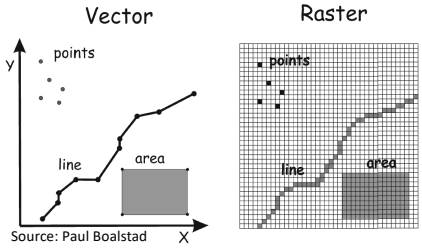


Source: Big data taxonomy (Cloud Security Alliance)

NC STATE UNIVERSITY

Data Sources

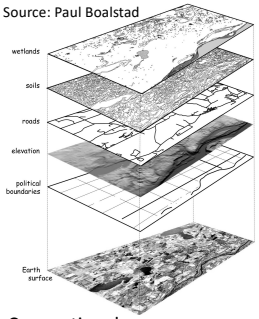
- Information about the locations and shapes of geographic features and relationships between them
  - Vector
  - Raster



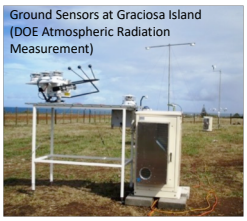
Source: Paul Boalstad

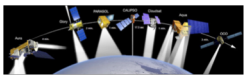
Data Sources

Source: Paul Boalstad



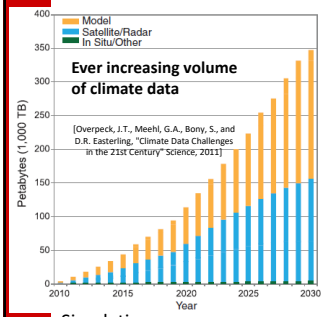
Conventional





Remote Sensing

Data Sources



Simulations

OpenStreetMap
- 1.4M users
- 3.6B uploaded GPS pnts.
- 2.0B nodes
- 0.2B ways

Twitter: 230M active user accounts, 5K tweets/s,
flickr: 87M users, 8B photos, 1.8B geotagged
facebook: 1.2B users, 250B photos, 17B geotagged

User Generated

Big Data

- A lot happens in a minute:
  - YouTube users upload 48 hours of new video
  - Instagram users share 3,600 new photos
  - Brands and organizations on Facebook receive 34,722 "likes"
  - Over 100,000 tweets are sent

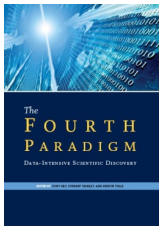
For current statistics, check "Data Never Sleeps": https://www.domo.com/learn/infographic-data-never-sleeps

Data guided discovery - A new paradigm

"... data-intensive science [is] ...a new, fourth paradigm for scientific exploration." - Jim Gray

McKinsey Global Institute

Big data: The next frontier for innovation, competition, and productivity



WIRED MAGAZINE: 16.07

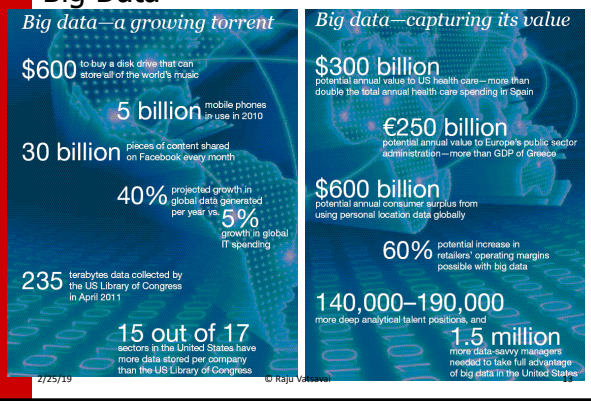
SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

McKinsey Global Institute – Report on Big Data

June 2011

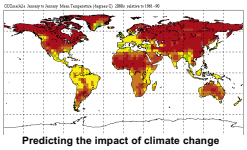


Motivating Applications

- Smart Assistants
  - Google Assistant
  - Alexa
  - Siri
- Smart Routing
  - Google Maps
  - Uber
- Ratings
  - Netflix, Yelp, …

Great Opportunities to Solve Society's Major Problems

Improving health care and reducing costs

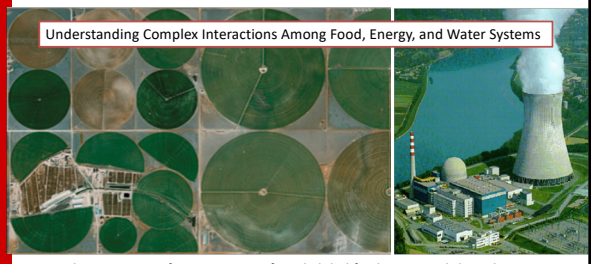


Predicting the impact of climate change

Finding alternative/ green energy sources

Reducing hunger and poverty by increasing agriculture production

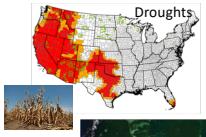
Great Opportunities to Solve Society's Major Problems

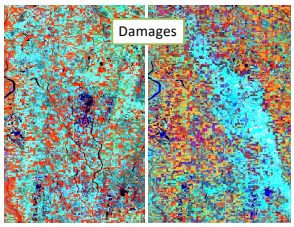


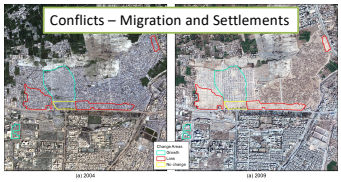
- Agriculture accounts for 70 percent of total global freshwater withdrawals
- Food production and supply chain consumes about 30 percent of total energy consumed globally
- Fast forward: 2050
  - 60% more food is needed; 50% energy consumption; 10% increase of water usage in agriculture

Great Opportunities to Solve Society's Major Problems

- Diseases to disasters to conflicts

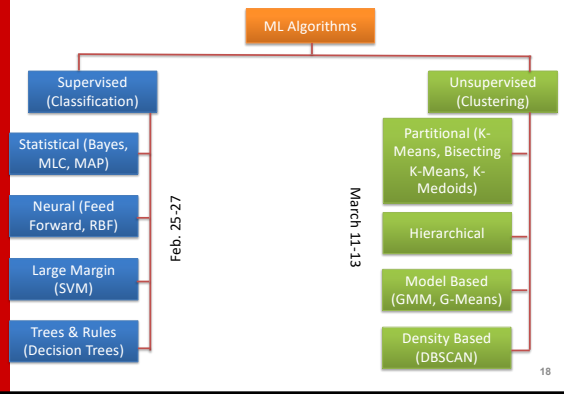






2/25/19

Taxonomy of Algorithms



18

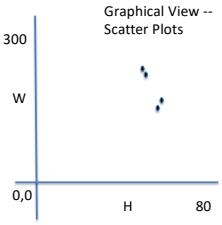
Classification Overview

Tabular View

Object	Height	Weight	Class
1	65	250	Y
2	66	240	Y
3	72	175	N
4	68	160	N

Original Data Space

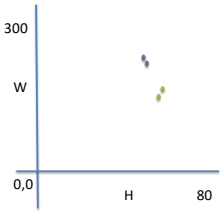
Graphical View -- Scatter Plots



Feature Space

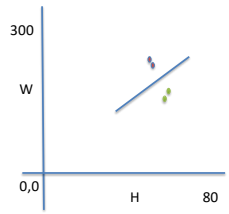
Classification Overview

Object	Height	Weight	Class
1	65	250	Y
2	66	240	Y
3	72	175	N
4	68	160	N



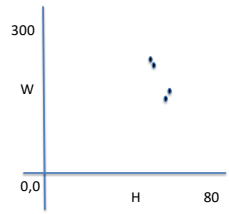
Classification Overview

Object	Height	Weight	Class
1	65	250	Y
2	66	240	Y
3	72	175	N
4	68	160	N



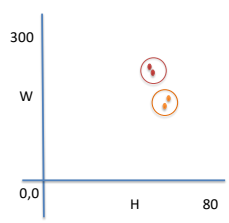
Clustering Overview

Object	Height	Weight	Class
1	65	250	-
2	66	240	-
3	72	175	-
4	68	160	-



Clustering Overview

Object	Height	Weight	Class
1	65	250	-
2	66	240	-
3	72	175	-
4	68	160	-



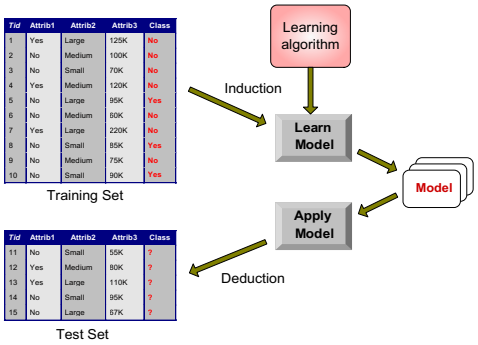
Classification: Definition

- Given a collection of records (training dataset)
  - Each record is characterized by a tuple $(\boldsymbol{x},y)$, where $\boldsymbol{x}$ is the attribute set and y is the class label
    - $\boldsymbol{x}$: attribute, predictor, independent variable, input
    - y: class, response, dependent variable, output
- Task:
  - Learn a model that maps each attribute set $\boldsymbol{x}$ into one of the predefined class labels y

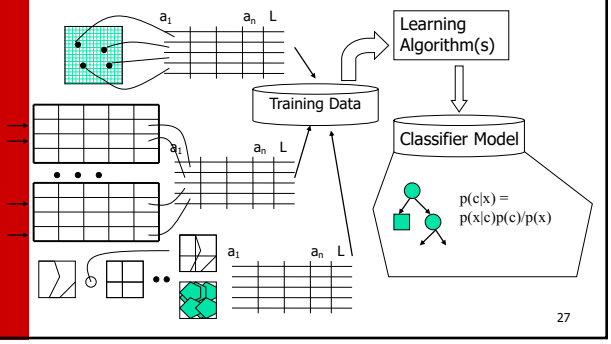
Examples of Classification Task

Task	Attribute set, x	Class label, y
Categorizing email messages	Features extracted from email message header and content	spam or non-spam
Identifying tumor cells	Features extracted from MRI scans	malignant or benign cells
Cataloging galaxies	Features extracted from telescope images	Elliptical, spiral, or irregular-shaped galaxies

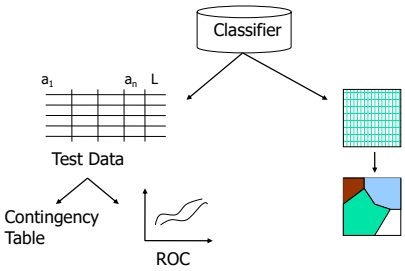
General Approach for Building Classification Model



Classifier Model Building

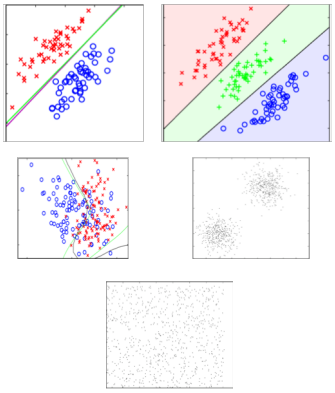


Classifier Model Usage

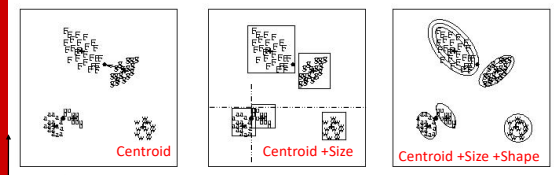


Why So Many Methods?

- Variety of data
- Variety of similarity measures
- Variety of decision surfaces (linear vs. nonlinear)
- Dimensionality
- Noise
- Classes (Binary vs. Multiclass, Multiclass vs. Multilabel
- Semantic
- Single instance vs. multiple instance



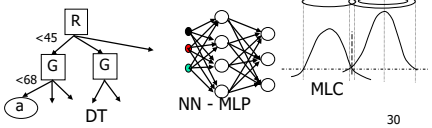
Why So Many Methods?



Min-distance to Means

Parallelepiped classification

x = [10,20]



30