Confluence of Data Science and AI

S. Lakshmivarahan
School of Computer Science
University of Oklahoma
Norman, OK- 73019, USA

Data Science: Early beginnings in Astronomy

- Much of what we know in physical sciences had their origins in Astronomy - with observations of celestial objects
- Thanks to the Herculean efforts of pioneers:
- Copernicus (1473-1543)
- Galileo (1544-1642)
- Kepler (1571-1630)
- Newton (1643-1727) are note worthy among others

Early observations suggest existence of simple laws of nature

- Observations collected over decades were meticulously analyzed to formulate new laws of nature: Examples include
- Heliocentric system
- The laws of Kepler
- Law of gravitation by Newton
- Newton's laws

Note: Within the context of physical sciences these are some of the earliest examples of **data mining**

What is data mining (DM)?

- DM is the process extracting the structure or patterns that are inherent in the data/ observations
- These patterns give clues about the data generating process
- Goal of DM is to understand the data generating process- from data to model – an inverse problem
- Since the motion of celestial objects inherently followed certain laws, early pioneers with their hard work and ingenuity could discover the laws that laid the foundation of the physical sciences and engineering as we know today

Renewed interest in DM with the Abundance of data

- It is estimated that the volume of data collected doubles in every three years -thanks to computers, large scale storage device, communication and sensor technologies
- Today interest in DM include - Physical sciences, Biological sciences, Space exploration, All of Engineering, Environmental Sciences, Medical Sciences, Economics, Finance, Banking and Commerce, Sports and recreation, Governments at all levels, to name a few
- More about DM a bit later – back to early astronomy

Development of models

- The newly found Newton's laws along with the concurrent developments in Calculus by Newton (1643-1727) and Leibnitz(1646-1716)and others naturally lead to the development of dynamic models to describe the motion of planets around the sun
- With the availability of models, the potential for forecast or prediction became very clear

Beginnings of DA

- Gauss (1777-1855) (when he was only 24 years old) using the known models of his time, takes up the challenging problem of predicting when the celestial object called ***Ceres*** will reappear on the telescope
- By combining the model and the observations, he then created the "**first assimilated model**" to make an accurate prediction of the time and location of reappearance of the lost object
- Data assimilation involves estimation of the unknown parameters of the model in question- an inverse problem

Gauss in 1801 laid the foundation for DA

- This work leads the development of the method of least squares which continues to be the work horse for the data assimilation industry today
- By this time, he has also invented the notion of distribution observational errors following the bell-shaped curve which we now call as the normal or Gaussian distribution
- Gauss made a successful prediction of the bearing and the time of reappearance of the last object with great accuracy – a forward problem

What is DA? – Fusion of model with data

- Models are general descriptions of the underlying physical processes in question.
- Data/observations reveal all the secrets of or the truth about the process that model tries to capture
- By combining models and data, we can get a specialized model called the assimilated model
- This assimilated model is a good tool for creating forecast or prediction
- One of the standard tools for the fusion of model and data is based on the method of least squares.

Goal of DM-DA is to generate good forecast: Examples: Model + data

- Predict the path of a hurricane, tornado- use of several models + data collected from satellite, radar and special planes that fly into the eye
- From the crime scene data, reconstruct the case – CSI, Miami
- NTSB – estimate the causes of failure using the data from the debris
- Predict the potential tax revenues so that a Government can develop its budget for the next year
- Medical diagnosis – from symptoms to the cure

Direct and inverse problems

Input => Model = > Output

Model + input: Compute the output – forward problem

Input and output: Identify/estimate the model – inverse problem

Model and output: Find input/parameters generate the specified output – inverse problem

Data to model: Data Mining – Inverse problem

Time Series Analysis – model, estimate, predict

Three Pillars of Data Science

- Data Mining (DM) – Inverse problem – from data to model – Remember Kepler
- Data Assimilation (DA) – Inverse problem of estimation of parameters – Remember Gauss
- Prediction (P) is a forward problem
- DM, DA and P form a continuum and form the three pillars of DS

Early Indian Astronomy

The four yugas: Krita, Treta, Dvapara and Kali

Ramayana in Treta yuga – about 7,000 years ago

Mahabharata in Dvapara yuga – about 5000 years ago

Lord Rama's horoscope: 5 planets in exalted position – Valmiki Ramayana

- Sun is in Aries & exalted
- Moon is in Cancer – in his own home
- Jupiter is in Cancer & is exalted
- Saturn is in Libra & is exalted
- Mars is in Capricorn & is exalted
- Venus is in Pisces & is exalted
- Rahu is in Sagittarius
- Ketu is in Gemini

Great Astronomers of Kaliyuga

- Aryabhata (476-550) –Patna –Gupta Dynasty
  - Earth is spherical, revolves around the sun and 365 days in a year
- Varahamira (505-587) –Ujjain, MP
- Brahmagupta (598-668)-Bhinmal, Rajasthan
- Baskara (598-680) – Sourastra, Gujarat
- A whole host of scholars who have made lasting contributions to Astronomy and Mathematics

Models in DS

- Based on causality vs correlation
- Explicit vs implicit
- Dynamic- ODE/PDE vs. static
- Deterministic vs. stochastic
- Linear s nonlinear
- Discrete time vs continuous time
- Discrete vs continuous space

Observations/Data

- Time Series –Daily max temperature in a city
- Spatial data- Current level COVID infection across each country in the world on a given day
- Saptio-temporal- Monthly rain fall in each state capital
- Data Matrix: X: $[x_1 , x_2 , x_3 \ldots x_n] , x_i \in R^d$
- Represents n points in d- dimensional space- n and/or d large
- Example: n member ensemble prediction –Forecast covariance
- Row vector L or Y of size n associated with columns of X – labels in supervised learning or response in least squares

DS methods – ML/AI

- Statistical estimation theory
- Kalman filtering – assimilate and forecast sequentially
- Multivariate regression analysis (1800's) – Statistics
- Data reduction using PCA (1940), ICA (1990) - Statistics
- Classification using clustering (1950's), ANN (1950's), Pattern Recognition (1950's), SVM (1980's), DNN (1980)
- Association rules, Decision trees (1960) Probabilistic networks (1990's) –AI, Machine learning
- Genetic algorithms – Discrete optimization
- Random field - Spatial data analysis

Summary

- DM seeks to uncover the structure or patterns in the data so that we can summarize the intrinsic properties of the Data generating process as a model
- The goal of DA is to fit the model to data.
- The fitted model has better predictive power
- DM, DA and P: parts of a continuum
- DM, DA and P : Three pillars in the development of a Predictive Science/Data Science

Evolution of AI- a historical view

- Growth of AI is closely tied to development of digital computers
- 1951 – Mauchly and Eckert -Stored program digital computers – ENIAC
- 1954-John Backus at IBM – FORTRAN, Compliers
- 1956-First OS
- High level programming languages: LISP, C, C++

Early days – 1950's-70's

- Alan Turing (1950): If humans using logical reasoning process information and make decisions, why can't machine do the same?
- A. Samuel (1950) at IBM developed the first program to play checkers. Developed a new search technique called alpha-beta pruning
- A. Newell, H. Shaw and H. Simon in mid 1950, developed "Logical Theorist" a program to mimic problem solving skills of humans
- The term AI was coined at the first workshop at Dartmouth organized by J. McCarthy and M. Minsky in 1954
- E. Feigenbaum introduced Expert systems in 1970's

Two approaches to AI

- Two competing approaches to AI
  - The ability to manipulate symbols by machines based on formal rules of logic became a reality- Automatic theorem proving, Game playing programs
  - Biologically inspired approach to mimic neurons in brains-ANN/DNN
  - Ultimate challenge in learning is the ability to “generalize”
  - Note: Today AI – largely refers to this biologically inspired approach

History of NN

- 1943- McCulloch and Pitts – proposed a network of switching elements to mimic neurons in the human brain
- 1957- F. Rosenblatt introduced Perceptron – the first Neural network with a single neuron
- 1971- A. Ivakhnenko and G. Lapa - 8 layered first deep net
- Feed forward, multi-layered design with back propagation - Werbos 1975
- 1979-Fukushima used a hierarchical multilayered NN – first design of CNN to extract features and recognize patterns

Development of PAC

- 1984; L. Valiant introduced the “The Theory of Learnable” called PAC theory – provably approximately correct
- Introduction of VC-dimension by V. N. Vapnik and A. Ya. Cheronenkis (1971) in connection uniform convergence of relative frequencies to their probabilities
- 1989 A. Blumer et al. Learnability and VC-dimension – related Valiant’s work with VC-dimension
- Using this relation between PAC and VC-dimension, analysis of NN got a boost -from being purely empirical approach to quantitatively precise framework
- By estimating the VC-dimension of NN, estimate the rate at which NN can “learn” and “generalize”

Evolution of AI

- 1989 DNN became a reality – Yann LeCun et. al developed a DDN to recognize handwritten zip codes
- 1997 - IBM Deep Blue - A chess playing computer program that defeated the world champion Gary Kasparov
- Turing award for Y. Bengio, G. Hinton and Yann LeCun for their development of DNN
- GPU by NVIDA with its parallel processing environ now plays an important role in various implementations of AI related projects capable of handling large volumes of data

Evolution of AI

- Today AI is everywhere: Heath care, Education, Manufacturing, Law, Finance, Politics

- Note: Lawyers solve the inverse problem: Knowing what I know, how to develop the case to reach a desired outcome for the client

Concurrent development in Robotics

- Interdisciplinary area of Robotics – Mechanical, Electrical, Computer Engineering, Computer vision and tools from AI
- Great advances have been made
  - Medical surgery
  - Driverless cars

Two modes of Learning in AI

- Supervised / Learning with a teacher (may be probabilistic)
  - A finite set of input-output – identify the system – inverse problem
  - Given n labelled samples (x_i, l_i), train ANN to classify the input samples
  - Given n pairs (x_i, y_i), express y as a function of x - Least squares theory of model building
  - Note: Ultimate goal relates to the ability to generalize. PAC theory is helpful in this assessment

Two modes of Learning in AI

- Unsupervised/Learning without a teacher
  - Given a finite set of data, build model, estimate and predict – two inverse and one direct problem
  - Automatically generate features and use them to model and/or classify
  - DNN – image analysis, speech/speaker recognition
  - Time Series Analysis – listen to the data using the autocorrelation
  - Reinforcement learning and Markov decision process

Rose smells the same by whatever name you call

1) AI, Expert systems, Knowledge discovery,

2) M/L : Classification, Pattern Recognition- SVM, ANN/DNN

3) Multivariate Data Analysis, Data Mining, Statistical Learning

4) Data Science, Big Data Analytics

5) Data centric AI

CS vs DS

- Growth in Computer Science was driven by advances in
  - Discrete mathematics, Theory of formal language, Algorithms and complexity theory and hard-ware technology

- Growth in Data Science will be propelled by advances
  - Statistics, Probability theory, Applied mathematics, Large scale inverse problems, Numerical simulation, Domain specific knowledge

Applications of AI/DNN

- Automatic speaker/speech recognition
- Image recognition-Handwritten
- Visual art processing – style, period,
- Natural Language processing
- Drug discovery/toxicology
- Customer relationship
- Recommender system
- Bioinformatics- discovery of gene functional relationship
- Image restoration

Confluence: DS and AI

- Both deal with solving a variety of inverse problems specific to various domains – role and importance of importance of domain knowledge is fundamental in the pursuit of DS
- Both deal big data -Training DNN needs large volumes data. PAC theory provides a lower bound on the number of samples to guarantee a certain level of performance
- Note: AI is part of DS

Algorithmic decision making-challenges

- DNN based approach lacks theory – why it works?
- Long way from building causality theory
- DNN is a step towards realizing strong AI- to create machines indistinguishable from humans, but not all-encompassing solution
- Transparency is needed to gain confidence
- Democratizing AI ? – Open access subject to malicious attacks
- DNN based decisions are sensitive perturbations in the input, malicious changes in the components of the architecture

Reading list

- B. G. Lindsay, J. Kettenring and D. O. Siegmund (2004) “A report on the future of Statistics”, Statistical Science, Vol 19, 387-413
- IMS Presidential Address: “Let us own data science” IMS Bulletin, Vol 43, Issue 7, 2014
- J. Fan, F. Han and H. Liu (2014) “Challenges of Big Data analysis”, National Science Review, Vol 1, 293-314
- D. Donoho (2017) 50 Years of Data Science, Journal of Computational and Graphical Statistics, 26:4, 745-766, DOI:10.1080/10618600.2017.1384734

References

- A. Blum, J. Hopcroft, and R. Kannan (2020) **Foundations of Data Sciences**, Cambridge University Press

- J. M. Lewis, S. Lakshmivarahan and S. K. Dhall (2006) **Dynamic Data Assimilation: a least squares approach**, Cambridge University Press

- V. Shikhman and D. Muller (2021) **Mathematical Foundations of Big Data Analytics,** Springer

- M. Vidyasagar (2003) **Learning and generalization**, Springer Verlag

References-continued

- I. Rish and G. Ya. Grabarnik (2015) **Sparse Modeling; Theory, Algorithms and Applications,** CRC Press

- **M Wainwright (2019) High-Dimensional Statistics: A non-asymptotic viewpoint, Cambridge University Press**

- J. Wang (2012) **Geometric Structure of High-Dimensional Data and Dimensionality** R**eduction**, Springer

- R. Vershynin (2020**) High Dimensional Probability Theory: an introduction with applications in Data Science**, Cambridge University Press
- B. Abu-Salih, et. al. (2021) **Social Big Data Analytics**, Springer