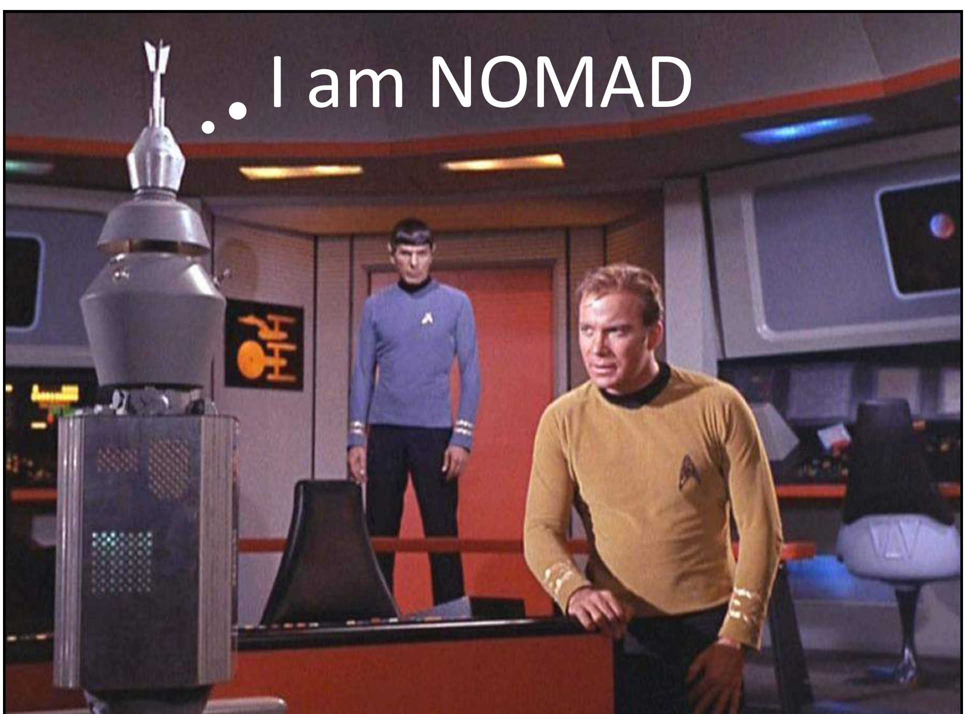
Data Mining and Machine Learning

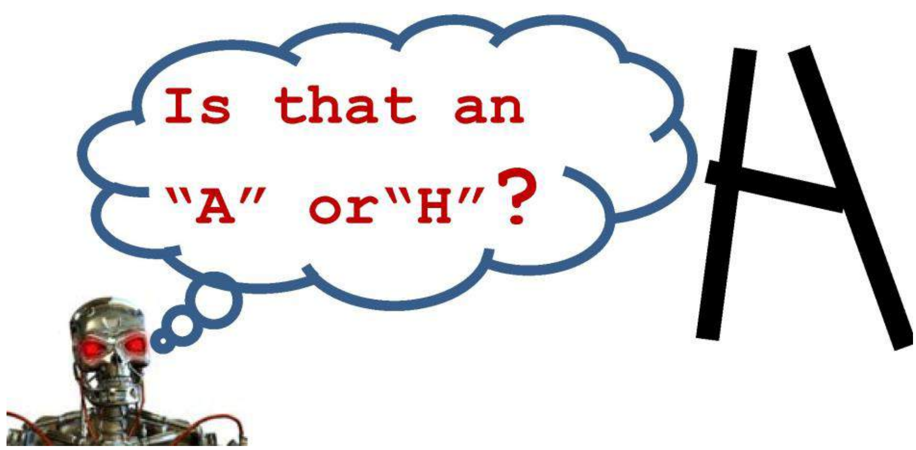
Erich Seamon
University of Idaho
www.webpages.uidaho.edu/erichs
erichs@uidaho.edu

University of Idaho NKN 1



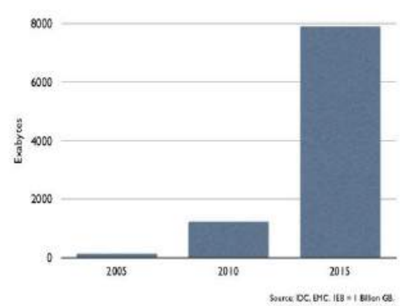
Data Mining and Machine Learning Outline

- Outlining the data mining and machine learning paradigm
  - Growth of data
- What is data mining & knowledge discovery?
  - Knowledge discovery process
  - Types of data mining
- Machine learning: an aspect of data mining
  - What is machine learning
  - Training vs. testing
  - Supervised vs. unsupervised vs. reinforced
  - Types of Algorithms
- Machine learning examples



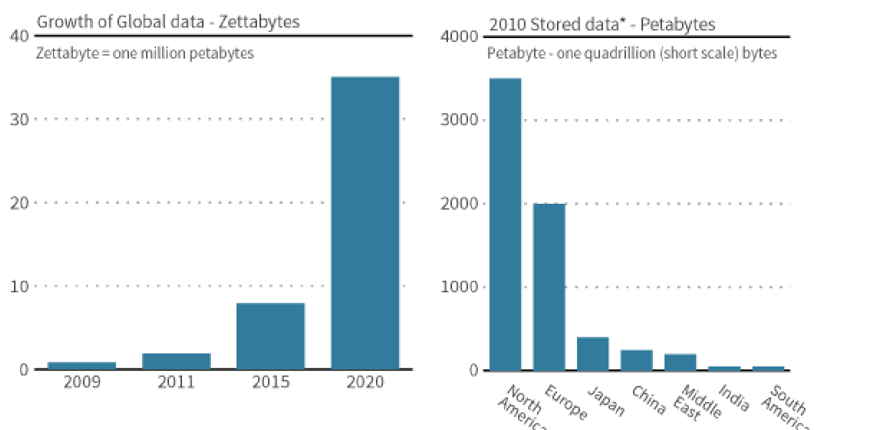
Data Growth in 2015

- Walmart handles 1M transactions per hour
- Google processes 24PB of data per day
- AT&T transfers 30PB of data per day
- 90 trillion emails are sent per year
- World of Warcraft uses 1.3PB of storage



Worldwide Data Growth at 7.9EB/Yr in 2015

Big data market is estimated to grow 45% annually to reach $25 billion by 2015



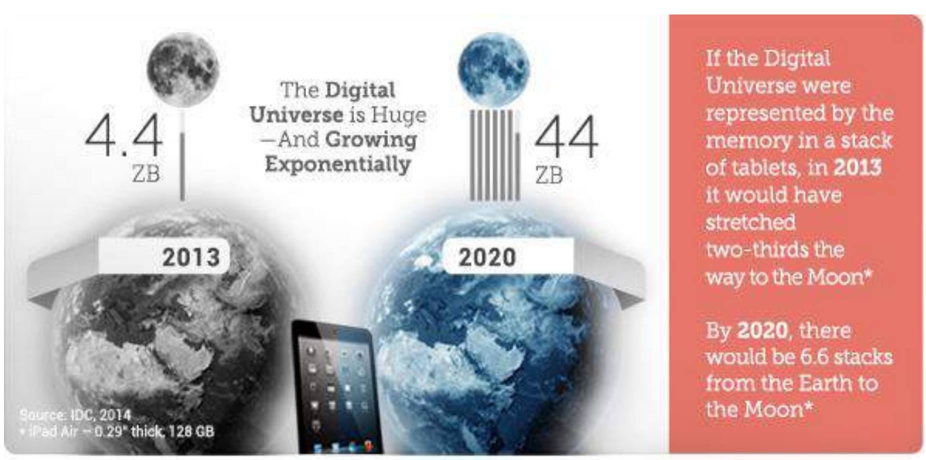
*greater than

Sources: Nasscom -CRISIL GR&A analysis

REUTERS

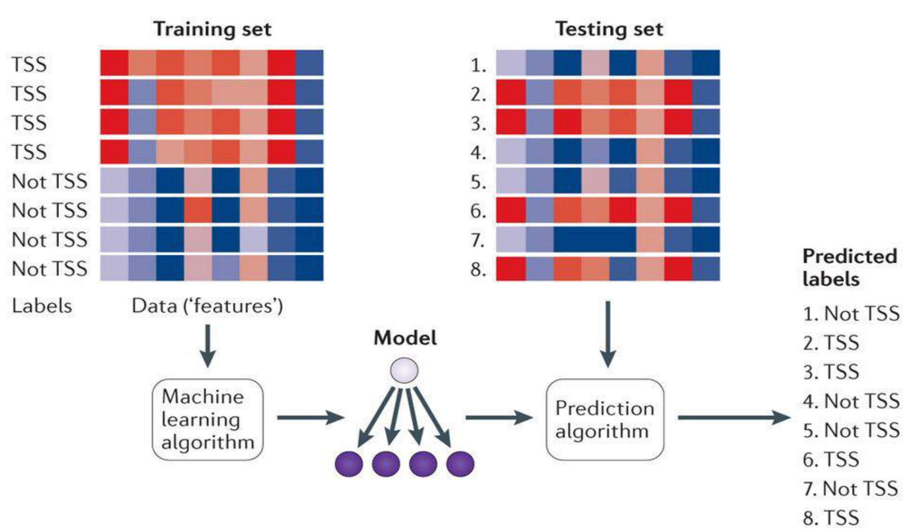
Understanding Data

- In many ways, our abilities to comprehend incomplete, disparate, or fragmented data is much more important to the discussion than the growth of data itself (King, et al 2015).
- Algorithms that allow us to gain knowledge from this incomplete data are the key.



Data Growth and Machine Learning

- Machine Learning is used when
  - A pattern exists
  - We cannot pin it down mathematically
  - We have data on it
- Learning techniques are preferred because:
  - They reduce time and cost
  - Produce results that are comparable to mining an entire data set



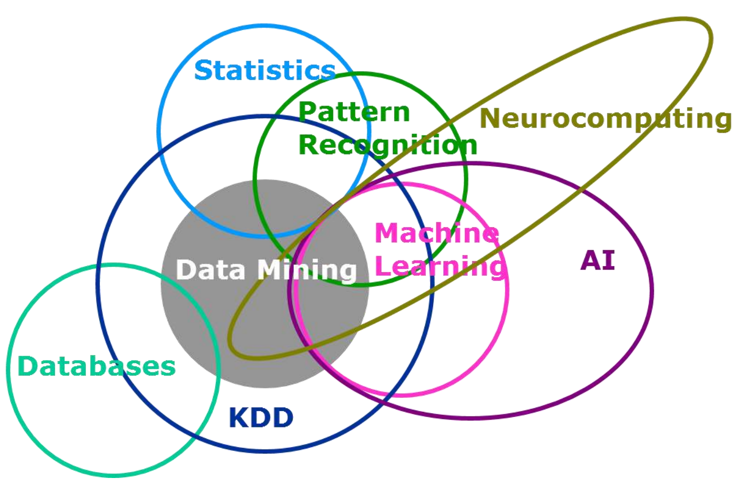
Nature Reviews | Genetics

Data Mining vs. Machine Learning

- Machine learning tends to be focused on performing a known task, whereas data mining is about the search for hidden nuggets of information.
- For instance, you might use machine learning to teach a robot to drive a car, whereas you would utilize data mining to learn what type of cars are the safest

Machine learning algorithms are virtually a prerequisite for data mining but the opposite is not true. In other words, you can apply machine learning to tasks that do not involvedata mining, but if you are using data mining methods, you are almost certainly using machine learning. (Lantz, 2013)

University of Idaho NKN 11



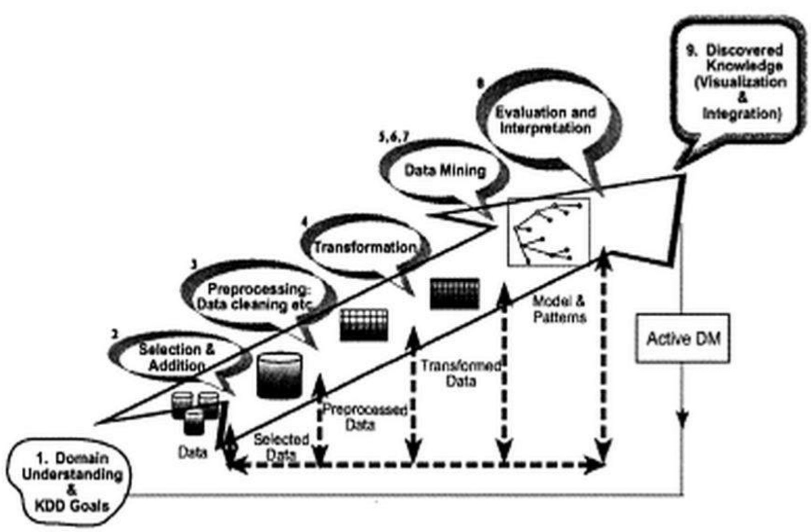
Guthrie, 2014 "Looking backwards, looking forwards: SAS, data mining and machine learning." http://blogs.sas.com/content/subconsciousmusings/2014/08/22/looking-backwards-looking-forwards-sas-data-mining-and-machine-learning/

University of Idaho NKN 12

Data Mining and Knowledge Discovery

- Fawley (1992) defines data mining as "the process of analyzing data from different perspectives and summarizing it into useful information". *Data mining is typically considered a core step of the knowledge discovery process.*
- Abu-Mostafa (2013) additionally terms data mining as "…a practical field that focuses on finding patterns, correlations, or anomalies in large relational databases".

University of Idaho NKN 13



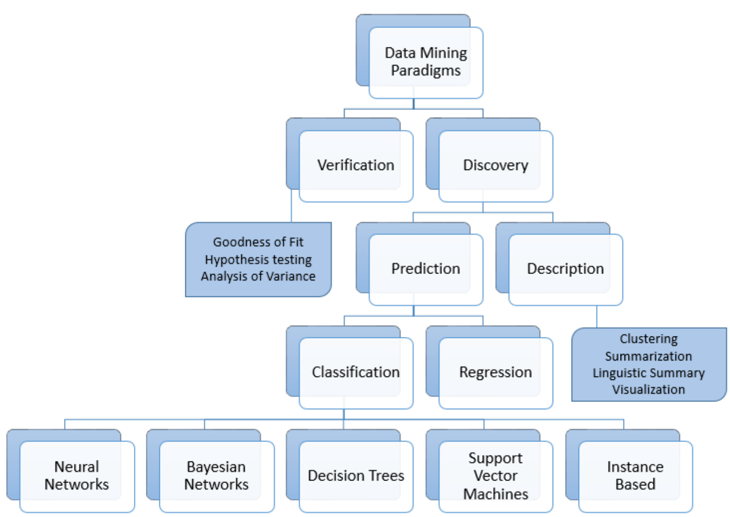
Nine steps that define the data mining/knowledge discovery process (Maimon, Rokach, 2006)

University of Idaho NKN 14

Components of Data Mining

- Machine Learning can be considered a sub-component of Data Mining (Rokach, 2014)
- Data Mining approaches can be divided into Discovery and Verification Systems
- Machine Learning falls under the Discovery area

University of Idaho NKN 15



University of Idaho NKN 16

Supervised and Unsupervised Learning

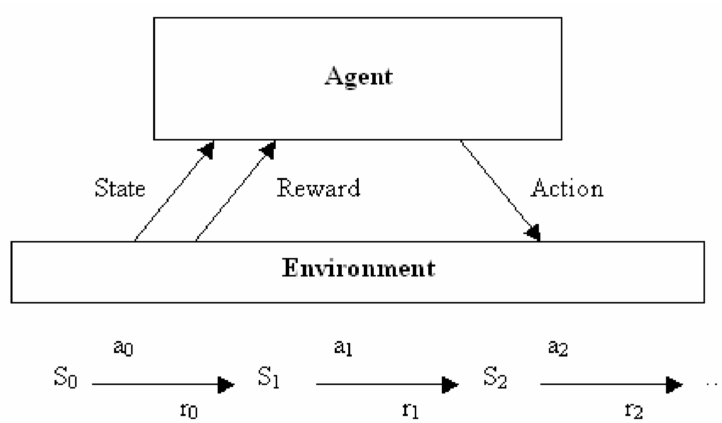
- Supervised Learning discovers patterns in data that related data attributes with a class. These patterns are then used to predict values of the class in future data instances.
- Unsupervised Learning is where data have no class. The intention of unsupervised learning is to explore the data to find its inherent structure, using various statistical methods

Machine Learning Algorithms

	Unsupervised	Supervised
Continuous	• Clustering & Dimensionality Reduction ○ SVD ○ PCA ○ K-means	• Regression ○ Linear ○ Polynomial • Decision Trees • Random Forests
Categorical	• Association Analysis ○ Apriori ○ FP-Growth • Hidden Markov Model	• Classification ○ KNN ○ Trees ○ Logistic Regression ○ Naive-Bayes ○ SVM

Reinforcement Learning

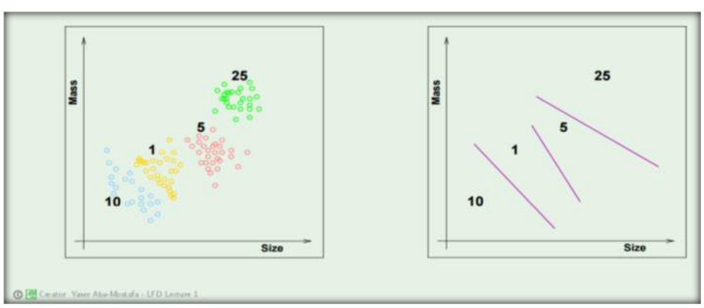
- Reinforcement learning is particularly well suited to problems which include a long-term versus short-term reward trade-off.
  - robot control,
  - telecommunications,
  - backgammon and checkers (Sutton and Barto 1998, Chapter 11).
- Monte Carlo Methods are sometimes used
  - Monte Carlo integration
  - Numerical optimization/iterative simulation

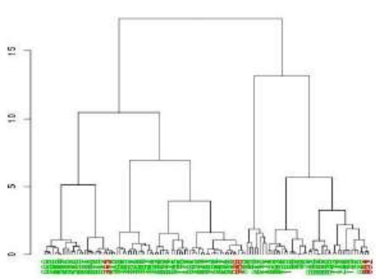


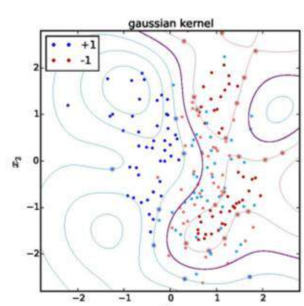
Goal: learn to choose actions that maximize:
$r_0 + \gamma r_1 + \gamma^2 r_2 + \ldots$, where $0 \leq \gamma < 1$

Supervised Learning

- Classification
  - KNN (K nearest neighbor)
    - Can be used in regression as well
    - Classification determined by K nearest neighbors which is most common.
    - Lazy learning – function is approximated localy and computation is deferred until classification
  - Decision Trees
    - Classification and regression approaches
    - Data mining trees are on data, not the decision. Output classification tree can be used for decision
    - Random forest and bagging methods output tree results
    - Varying decision tree algorithms: CART, CHAID, C4.5, ID3
  - Logistic Regression
  - Naïve-Bayes (spam, text filtering)
  - Support Vector Machines (SVM)
    - Classification and regression approaches
    - Non-probabilistic binary linear classifier

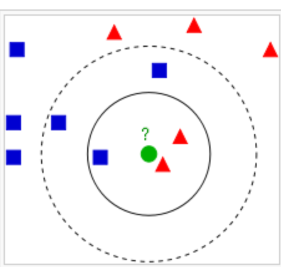




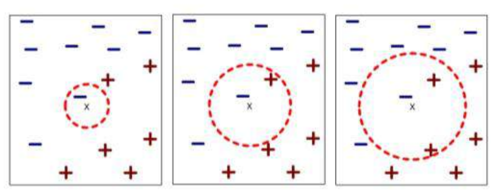


Supervised Learning (con'd)

- Classification
  - KNN (K nearest neighbor)
    - Can be used in regression as well
    - Classification determined by K nearest neighbors which is most common.
    - Lazy learning – function is approximated localy and computation is deferred until classification
  - Decision Trees
    - Classification and regression approaches
    - Data mining trees are on data, not the decision. Output classification tree can be used for decision
    - Random forest and bagging methods output tree results
    - Varying decision tree algorithms: CART, CHAID, C4.5, ID3
  - Logistic Regression
  - Naïve-Bayes (spam, text filtering)
  - Support Vector Machines (SVM)
    - Classification and regression approaches
    - Non-probabilistic binary linear classifier

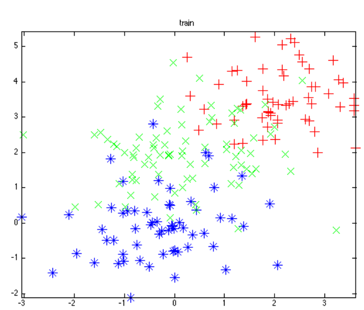


Example of *k*-NN classification. The test sample (green circle) should be classified either to the first class of blue squares or to the second class of red triangles. If *k* = 3 (solid line circle) it is assigned to the second class because there are 2 triangles and only 1 square inside the inner circle. If *k* = 5 (dashed line circle) it is assigned to the first class (3 squares vs. 2 triangles inside the outer circle).



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

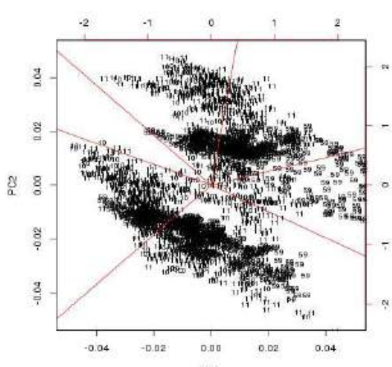
K-nearest neighbors of a record x are data points that have the k smallest distance to x

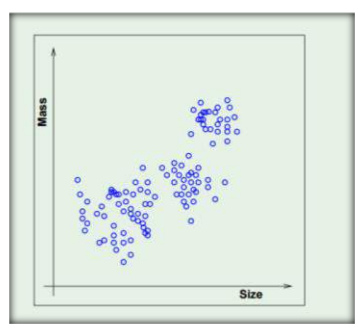


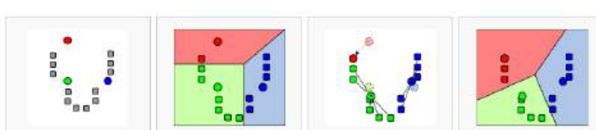
Unsupervised Learning

- Clustering and Dimensionality Reduction
  - SVD – Singular Value Decomposition. If you have two variables, one is humidity index and another one is probability of rain, then their correlation is so high, that the second one does not contribute with any additional information useful for a classification or regression task. The eigenvalues in SVD help you determine what variables are most informative, and which ones you can do without.
  - Principal Components
  - K-means
- Association Analysis
  - Apriori
  - FP-Growth
- Hidden Markov (related to Hoeffding's Inequality)

PCA







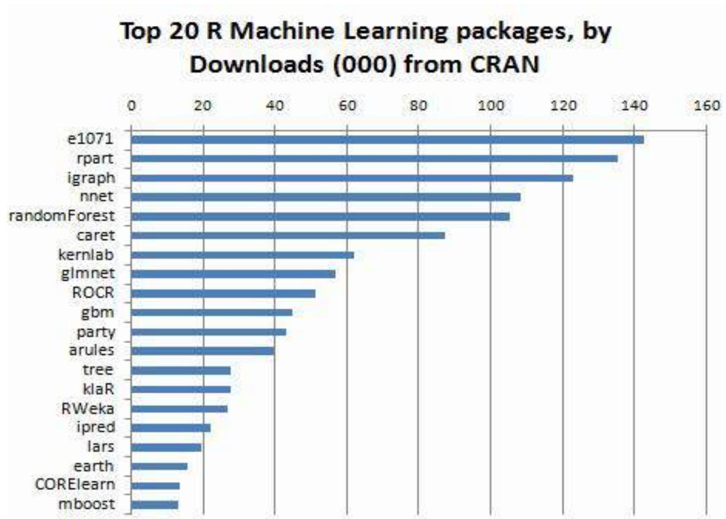
1. k initial "means" (in this case k=3) are randomly generated within the data domain (shown in color).

2. k clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the k clusters becomes the new mean.

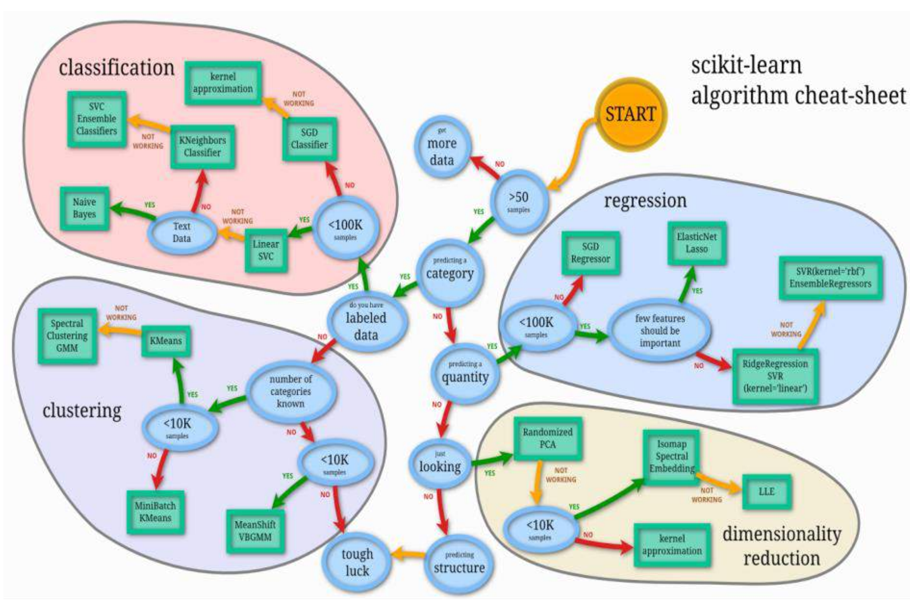
4. Steps 2 and 3 are repeated until convergence has been reached.

K Means



20 popular Machine Learning R packages by analyzng the most downloaded R packages from Jan-May 2015. (Kdnuggets – Geethika,2015)
http://www.kdnuggets.com/2015/06/top-20-r-machine-learning-packages.html

University of Idaho NKN 27

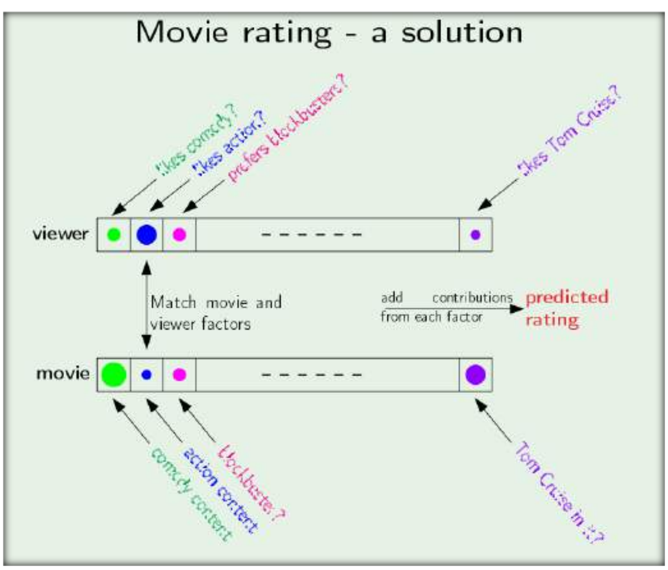


University of Idaho NKN 28

Examples

- Retail: Data drives prices and recommendations
- Marketing: Market sales and recommendations
- IT Management: IT operational intelligence
- Customer Management: Customer insight
- Operations: Automated response
- Public Safety: Crime hot spot/COMSTAT
- Medical diagnosis
- Climate modeling and downscaling

University of Idaho NKN 29



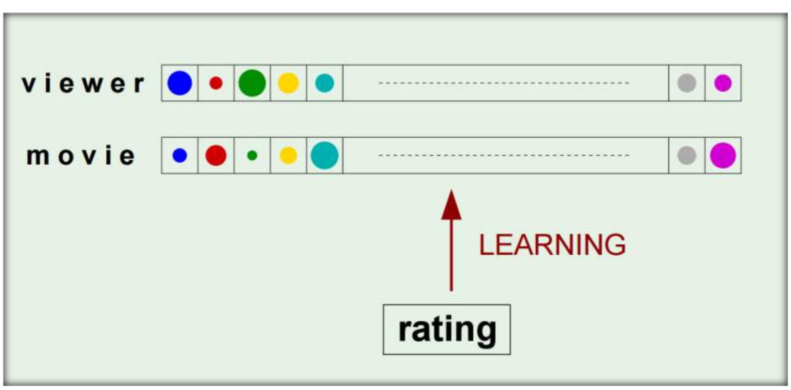
Netflix Machine Learning

University of Idaho NKN 30

Examples

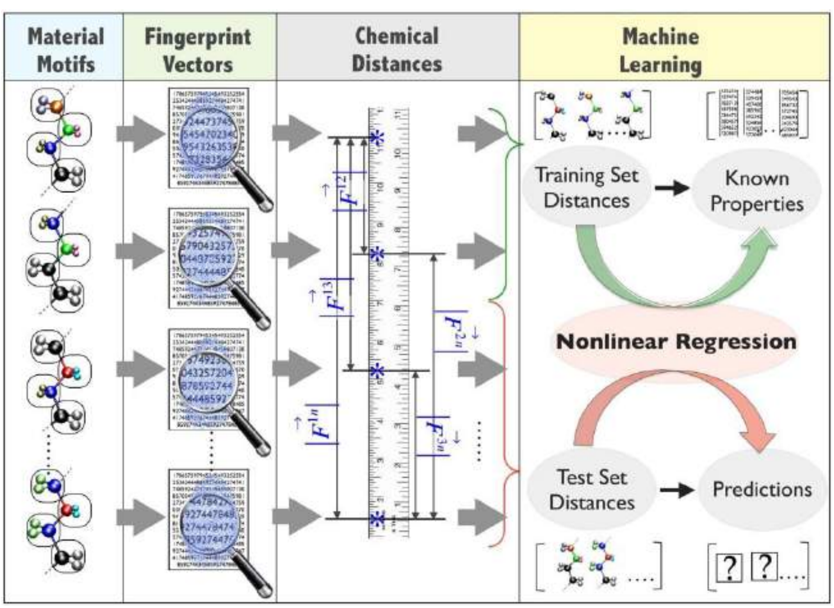
- Retail: Data drives prices and recommendations
- Marketing: Market sales and recommendations
- IT Management: IT operational intelligence
- Customer Management: Customer insight
- Operations: Automated response
- Public Safety: Crime hot spot/COMSTAT
- Medical diagnosis
- Climate modeling and downscaling

Netflix Machine Learning



Examples

- Retail: Data drives prices and recommendations
- Marketing: Market sales and recommendations
- IT Management: IT operational intelligence
- Customer Management: Customer insight
- Operations: Automated response
- Public Safety: Crime hot spot/COMSTAT
- Medical diagnosis
- Climate modeling and downscaling



Examples

- Retail: Data drives prices and recommendations
- Marketing: Market sales and recommendations
- IT Management: IT operational intelligence
- Customer Management: Customer insight
- Operations: Automated response
- Public Safety: Crime hot spot/COMSTAT
- Medical diagnosis
- Climate modeling and downscaling

Mapping Chemical properties

…learning methods may be used to establish a mapping between a suitable representation of a material (i.e., its 'fingerprint' or its 'profile') and any or all of its properties using known historic, or intentionally generated, data. The material fingerprint or profile can be coarse-level chemo-structural descriptors, or something as fundamental as the electronic charge density, both of which are explored here. Subsequently, once the profile u property mapping has been established, the properties of a vast number of new materials within the same subclass may then be directly predicted (and correlations between properties may be unearthed) at negligible computational cost, thereby completely bypassing the conventional laborious approaches towards material property determination alluded to above (Pilania, 2013)


SCIENTIFIC REPORTS

OPEN

Accelerating materials property predictions using machine learning

SUBJECT AREAS: POLYMERS, ELECTRONIC STRUCTURE, COMPUTATIONAL METHODS, ELECTRONIC PROPERTIES AND MATERIALS

Received 25 June 2013
Accepted 9 September 2013
Published 30 September 2013

The materials discovery process can be significantly expedited and simplified if we can learn effectively from available knowledge and data. In the present contribution, we show that efficient and accurate prediction of a diverse set of properties of material systems is possible by employing machine (or statistical) learning methods trained on quantum mechanical computations in combination with the notions of chemical similarity. Using a family of one-dimensional chain systems, we present a general formalism that allows us to discover decision rules that establish a mapping between easily accessible attributes of a system and its properties. It is shown that fingerprints based on either chemo-structural (compositional and configurational information) or the electronic charge density distribution can be used to make ultra-fast, yet accurate, property predictions. Harnessing such learning paradigms extends recent efforts to systematically explore and mine vast chemical spaces, and can significantly accelerate the discovery of new application-specific materials.

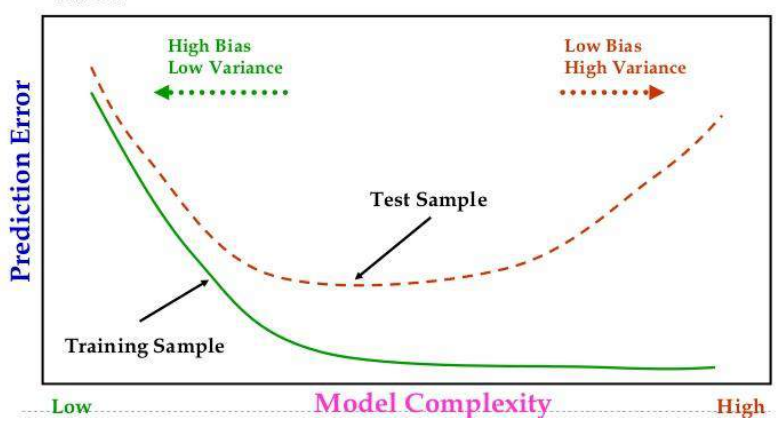
Other topics

- **Generalization/approximation tradeoffs**
- **Numerical optimization/simulation**
- **Hoeffding's Inequality**
  - In probability theory, **Hoeffding's inequality** provides an upper bound on the probability that the sum of random variables deviates from its expected value.
- **Vapnik–Chervonenkis dimension**
  - The VC dimension has utility in statistical learning theory, because it can predict a probabilistic upper bound on the test error of a classification model.
    - VC is the size of the largest finite subset of X – Shattered by H (Hypothesis space)
    - If arbitrarily large finite sets of X can be shattered by H – then VC(H) = infinity

University of Idaho NKN 37

Model Selection and Bias-Variance Tradeoff

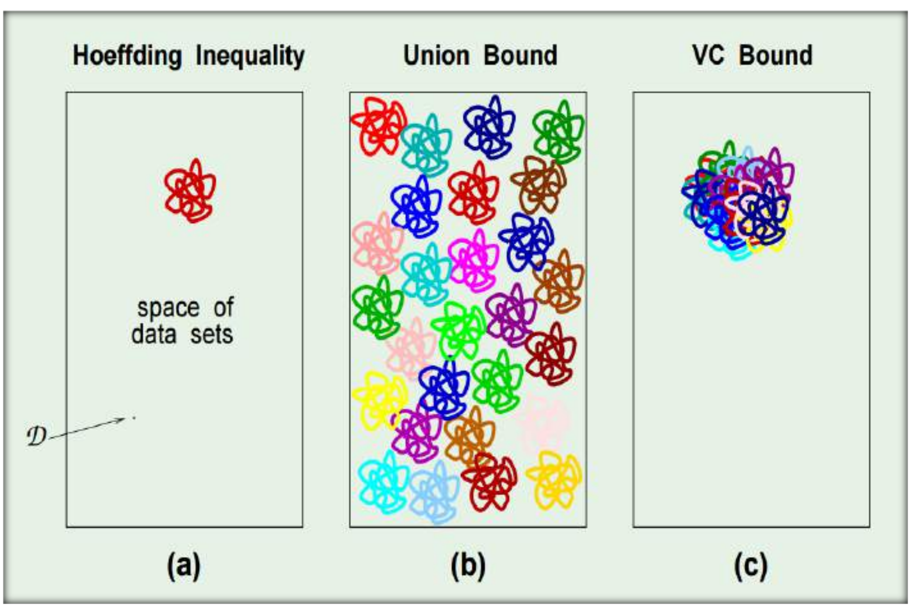
- Typical behavior of the test and training error, as model complexity is varied.



University of Idaho NKN 38

Questions

- Why use machine learning techniques?
- What is the value scientifically, financially?
- How does machine learning stack up to historical information?
- How does data mining relate to machine learning?
- Can machine learning techniques be used in everyday practice?



FINIT

FINIT