Data Science

A quick overview of Machine Learning, Data Mining, Information Retrieval and Complex Networks

2016 / 09

BIO

Researcher II at IPEA

MSc student at PESC/UFRJ

Research interests: Data Mining & Machine Learning in Dynamic Environments, Complex Networks, Big Data for Social Development

https://github.com/raulsenaferreira

http://lattes.cnpq.br/7007150957758256

Summary

Theory and experiments:

- Information Retrieval
- Network Science (Complex networks)
- Data Mining
- Machine Learning

Calling for research partnership

Information Retrieval

Information retrieval is a sub-field of computer science that deals with the automated storage and retrieval of documents. [1]

Applications: Web Search engines, public libraries …

Fields: Music information retrieval [2], text retrieval [3], temporal IR [4] ...

Metrics: Precision and recall, Precision @ K, DCG, NDCG, TF-IDF ...

Databases: CSTNews corpus [14], Cystic fibrosis [15], Gutenberg Corpus [16], Wikipedia [17]

Information Retrieval

Precision: Fraction of retrieved documents that are relevant to the query

Recall: Fraction of the documents that are relevant to the query that are successfully retrieved

F1: (Precision * Recall) / (Precision + Recall)

Precision @ K: Top K relevant results on the first search results page

TF-IDF: (Term frequency & Inverse document frequency) = tf * idf

NDCG: (DCG / IDCG)

Information Retrieval

Important libraries and techniques:

Porter Stemmer [6]

Apache Lucene [8]

Natural Language Toolkit (NLTK) [10]

Apache Solr [18]

Elastic (Former ElasticSearch) [19]

Information Retrieval

Stemmer: Morphological root of the word

Ex: Fishing => Fish

Friendlies => Friendly

There are many stemming algorithms. Porter is one of them.

Snowball: Small string processing language for creating stemming algorithms [20]

Information Retrieval

NLTK: Helps analyze concordance, similarity, context, count, frequency distribution, and many others aspects about a text

Solr or Elastic Search: Built over lucene search engine, ready-to-use out of box. Adds more power to applications based on lucene

Information Retrieval

Books and resources to study Information Retrieval:

NLTK book [5]

Introduction to Information Retrieval [7]

Vector Spaces [9]

PyLucene 4.0 (in 60 seconds) tutorial [10]

Preparing the test environment

1. Download repository:
    a. https://github.com/raulsenaferreira/Systems-Engineering
    b. $ mkdir data_science_venv
    c. $ cd data_science_venv

P.S.: You can try use “$ pip install -r requirements.txt” located at root of project. If something goes wrong, execute all steps

2. Install virtualenv tool and initialize the virtual environment:
    a. $ pip install virtualenv
    b. $ virtualenv IR_venv
    c. copy BRI folder from Systems-Engineering repository to IR_venv folder
    d. $ source IR_venv/bin/activate
    e. $ cd IR_venv/BRI/

Preparing the test environment

3. Install the requirements for your experiment

    a. $ sudo easy_install pip | $ sudo pip install -U nltk | sudo pip install -U numpy

4. Import NLTK corpora

    b. $ python2.7
    c. >>> import nltk
    d. >>> nltk.download()
    e. Select “all-corpora”

5. Save the virtual env state whenever you want (pip freeze > requirements.txt)

P.S.: For each project, README.md file contains instructions about the done work

2016 / 09

Information Retrieval

Implementing a retrieval system (in memory) following the vector model:

https://github.com/raulsenaferreira/Systems-Engineering/tree/master/BRI/Work_1

Information Retrieval

Evaluating an information retrieval model

https://github.com/raulsenaferreira/Systems-Engineering/tree/master/BRI/Work_2

2016 / 09

Information Retrieval

Doing the previous system using Lucene (pylucene)

https://github.com/raulsenaferreira/Systems-Engineering/tree/master/BRI/Work_3

2016 / 09

Information Retrieval

Challenges and current research:

Mathematical information retrieval [11]

Text-based intelligent systems [12]

Intelligent information retrieval [13]

References

1. Frakes, W. B., & Baeza-Yates, R. (1992). Information retrieval: data structures and algorithms.
2. http://www.cs.bu.edu/~snyder/cs591/Handouts/MusicRetrievalSurvey.pdf
3. https://en.wikipedia.org/wiki/Document_retrieval
4. Alonso, O., Strötgen, J., Baeza-Yates, R. A., & Gertz, M. (2011). Temporal Information Retrieval: Challenges and Opportunities. TWAW, 11, 1-8.
5. http://www.nltk.org/book/
6. https://github.com/sangheestyle/bisonlucene/blob/master/PorterStemmerAnalyzer.py
7. http://www-nlp.stanford.edu/IR-book/
8. http://lucene.apache.org/pylucene/
9. http://www.facom.ufu.br/%7Ealbertini/1sem2013/ori/slides/07vector-site.pdf

References

10. http://graus.nu/blog/pylucene-4-0-in-60-seconds-tutorial/

11. Schubotz, M., Youssef, A., Markl, V., & Cohl, H. S. (2015, August). Challenges of Mathematical Information Retrievalin the NTCIR-11 Math Wikipedia Task. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 951-954). ACM.

12. Jacobs, P. S. (2014). Text-based intelligent systems: Current research and practice in information extraction and retrieval. Psychology Press.

13. Lewis, David D. "Learning in intelligent information retrieval." Machine Learning: Proceedings of the Eighth International Workshop. 2014.

References

14. http://conteudo.icmc.usp.br/pessoas/taspardo/sucinto/cstnews.html

15. http://people.ischool.berkeley.edu/~hearst/irbook/cfc.html

16. http://www.nltk.org/book/ch02.html

17. https://en.wikipedia.org/wiki/Wikipedia:Database_download

18. http://lucene.apache.org/solr/

19. https://www.elastic.co/

20. http://snowballstem.org/

Network Science

Network science (also known as Complex networks) is the study of the collection, management, analysis, interpretation, and presentation of relational data [1]

Applications: Crime prediction [17], Vulnerable communities detection [18], Virus spreading prediction, Game Theory, Recommender Systems ...

Fields: Neuroscience [6], Public Health [5], Web Search Ranking [7], Network Security [8], Recommender Systems [9], Social Networks [12]...

Network Science

Metrics: Power Law [13], degree correlation [15], Page Rank [16], Betweeness, Closeness, Assortativity ...

Databases: Stanford Large Network Dataset Collection [19], AWS Public Datasets [20], UCI Network Data Repository [21], Barabasi's Research Group data sets [22]

Network Science

The degree distributions are all plotted

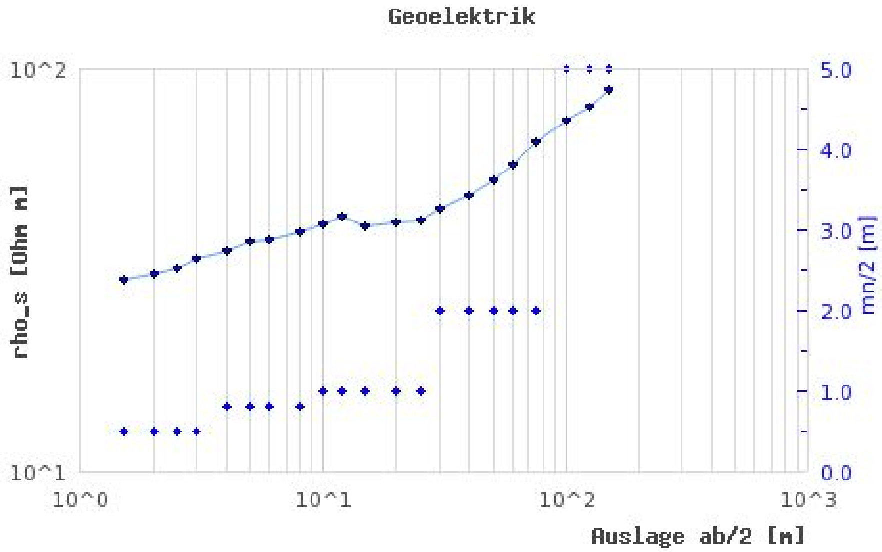
on a double logarithmic scale,

often called a log-log plot.

The main reason is that when nodes

with widely different degrees coexist,

a linear plot is unable to display them all.



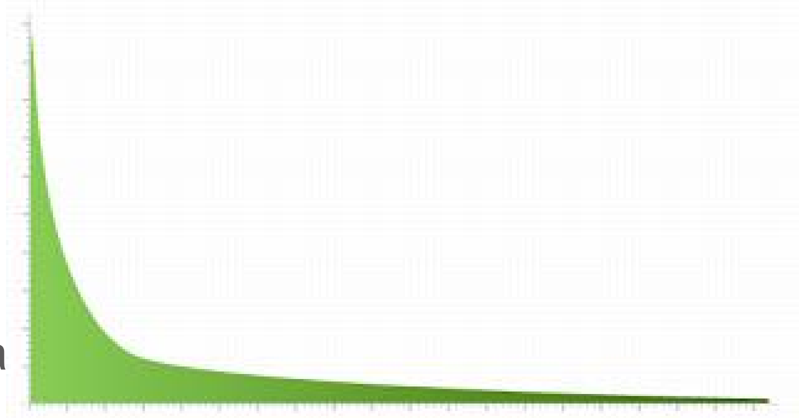
Network Science

Power Law (long tail) => p(k) ~ k⁻ᵞ

20% of the population is 80% popular

Ex: The speeds of cars on a highway,

the weights of apples in a store, air pressure, sea

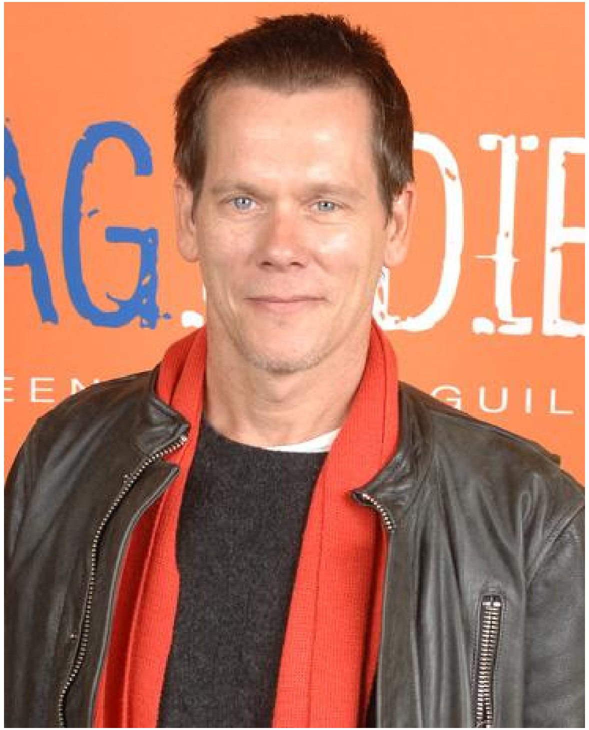


www, protein interactions …

Scale-Free Networks: Network whose degree distribution follows a power law. Hubs are bigger than other nodes.

Network Science

Small World: "6 degrees of Kevin Bacon" [27]



Network Science

Small World: Milgran experiment => 6 degrees of separation

Milgram asked the participants to record in the package each step of the path, and the mean number of hops of completed paths was about 5.9. This led to the popularization of the idea that there are no more than about 6 steps between each pair of people in the world.

Ex: Facebook has an 3.5 degrees of separation

Do you want to know your degree of separation? (i have 3.37 degrees)

https://research.facebook.com/blog/three-and-a-half-degrees-of-separation/

Network Science

Scale-free networks are robust against fails

Scale-free networks are week against directed attacks

http://barabasi.com/networksciencebook/chapter/8#robustness

Network Science

Metric:

Node degree = # of degrees

Clustering coefficient = Probability to form triangles inside the graph

Many algorithms to **measure the importance** of nodes. Ex:

Page Rank (Google)

HITS

Network Science

Metrics:

Assortativity => Nodes tend to link with another node with similar degree

Varies between -1 and 1

-1 = Node tend to connect with other nodes with different node degree

1 = Node tend to connect with other nodes with equal node degree

Network Science

Metrics:

Assortativity => Nodes tend to link with another node with similar degree

Varies between -1 and 1

-1 = Node tend to connect with other nodes with different node degree

1 = Node tend to connect with other nodes with equal node degree

Network Science

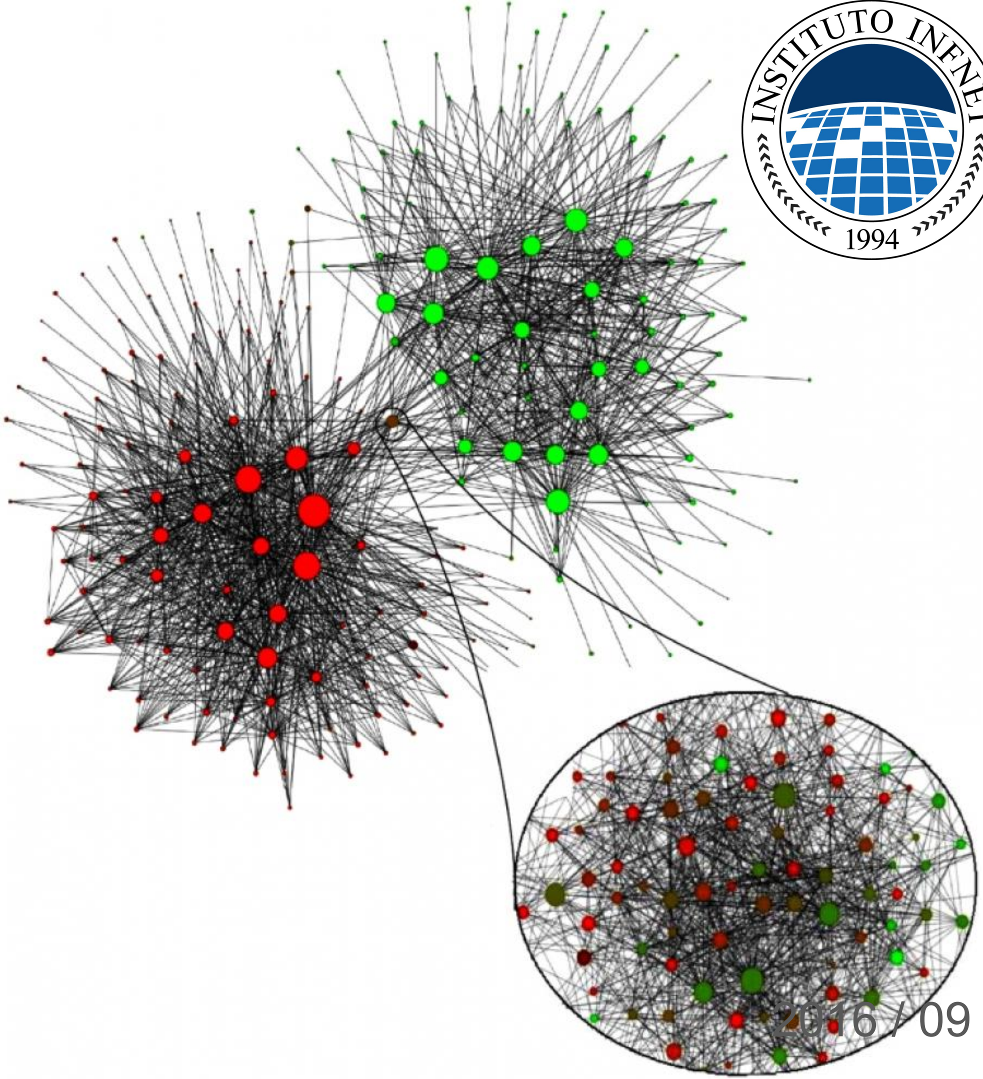
Communities:

Communities extracted from the call

pattern of the consumers

of the largest Belgian

mobile phone company



Network Science

Communities:

Plays a important role in big data problems

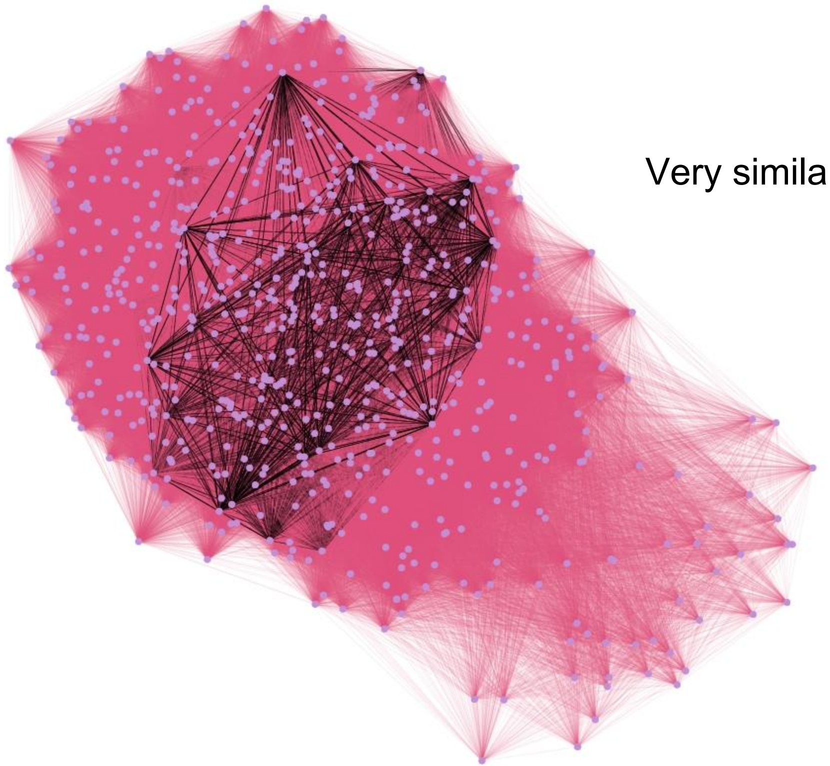
Communities can say most things about a network.

Helps to show hidden patterns inside a network.

Stochastic block models are the most commom techniques to find communities

My recent experiment tries to find risk social profiles (people murdered by narcotraffic reasons) inside a network made by missing people

Network Science



Very similar structures !

Pink edges = missing people in Rio de Janeiro and São Paulo;
Black edges = common profile of people murdered by narcotraffic

Stochastic block model with K = 3

Network Science

Network science theory can be applied to knowledge discovery problems

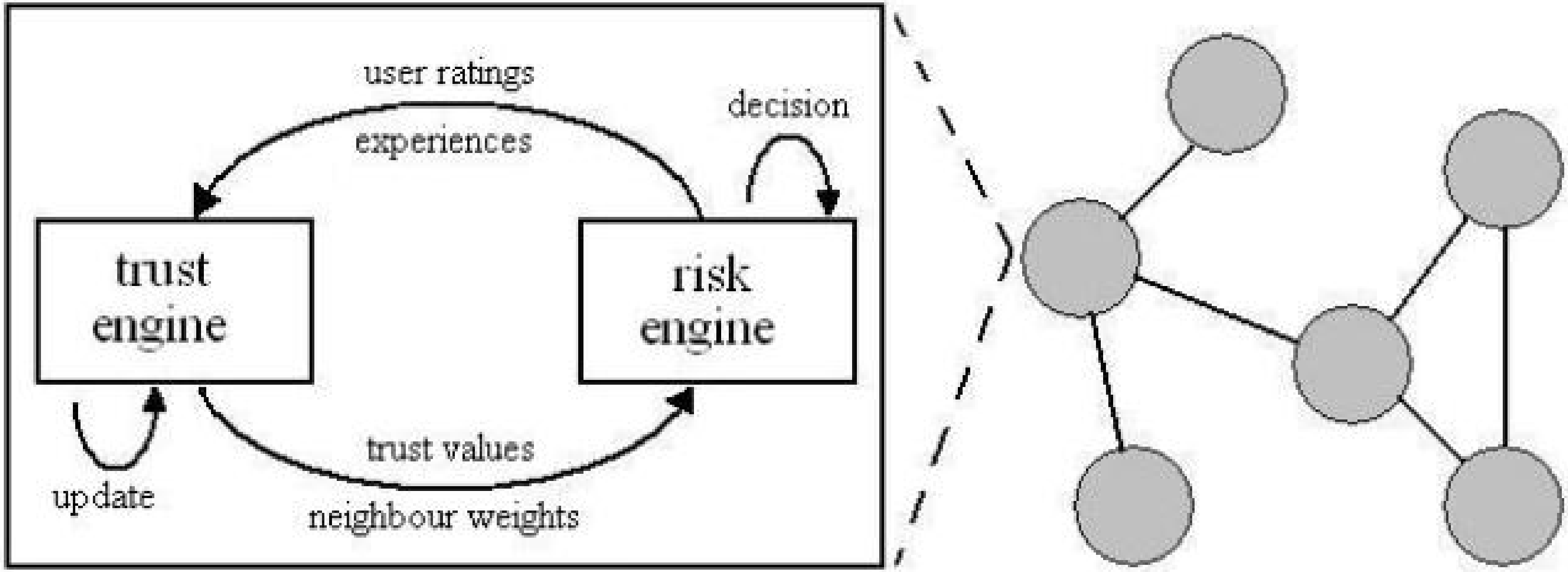
Also can be applied to machine learning problems

Ex:

Complex network **+** collaborative filtering = Trust-based Collaborative Filtering

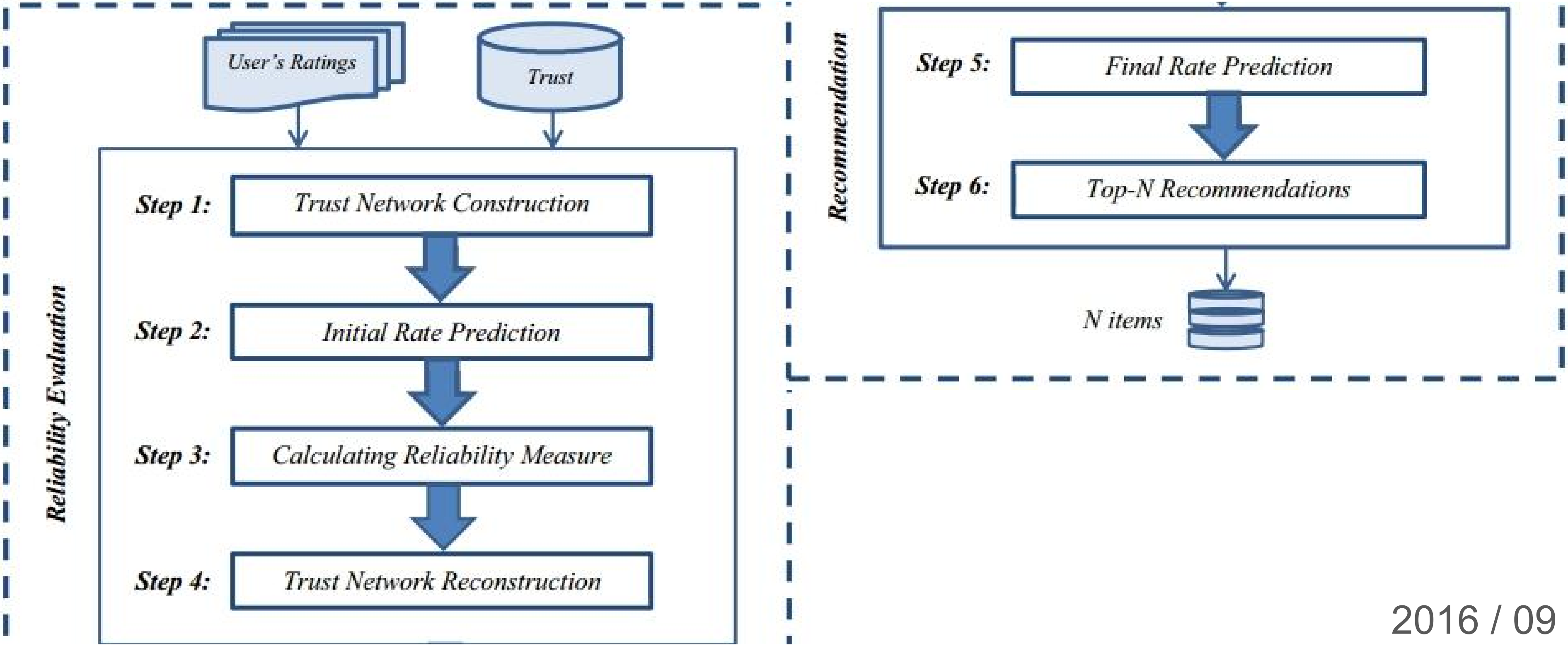
Network Science

Trust-Based Collaborative Filtering



Source: [Carbone et al.,2003]

Network Science



Network Science

Important Libraries/Techniques:

GraphTool [2]

Stanford Network Analysis Project (SNAP) [3]

Powerlaw.py [14]

Network Science

Books and resources:

Network Science by Albert-Lásló Barabási [12]

Connected (book) [4]

Statistical mechanics of complex networks [10]

Scale Free Networks [11]

Network Science

Investigating some networks (dolphins, political books, political blogs)

https://github.com/raulsenaferreira/Systems-Engineering/blob/master/Redes%20Complexas/dolphin.py

Network Science

Network Science applied to Missing People phenomenon

https://github.com/raulsenaferreira/Systems-Engineering/blob/master/Redes%20Complexas/generateGraph.py

Network Science

Challenges and current research:

The current and future challenges [23]

Using graph theory to understand the brain [24]

Fingerprinting in wireless networks [25]

References

1. Brandes, Ulrik, et al. "What is network science?." Network Science 1.01 (2013): 1-15.
2. https://graph-tool.skewed.de/
3. http://snap.stanford.edu/
4. http://www.connectedthebook.com/
5. http://www.ted.com/talks/nicholas_christakis_the_hidden_influence_of_social_networks
6. http://www.ted.com/talks/sebastian_seung
7. Massimo Franceschet, PageRank: Standing on the Shoulders of Giants. Communications of the ACM, Vol. 54 No. 6 (2011)
8. R. Albert, H. Jeong, A.-L. Barabási, Error and attack tolerance of complex networks, Nature 406, 378-482 (2000).

References

9. O'Donovan, John, and Barry Smyth. "Trust in recommender systems." Proceedings of the 10th international conference on Intelligent user interfaces. ACM, 2005.

10.http://www.barabasilab.com/pubs/CCNR-ALB_Publications/200201-30_RevModernPhys-StatisticalMech/200201-30_RevModernPhys-StatisticalMech.pdf

11.http://www3.nd.edu/~networks/Publication%20Categories/01%20Review%20Articles/ScaleFree_Scientific%20Ameri%20288,%2060-69%20(2003).pdf

12. http://barabasi.com/networksciencebook/

13. Duncan J. Watts, Six Degrees: The Science of a Connected Age. W. W. Norton & Company, 2003.

References

14.https://github.com/raulsenaferreira/Systems-Engineering/tree/master/Redes%20Complexas/powerlaw-1.3.5

15.http://ocw.mit.edu/courses/engineering-systems-division/esd-342-network-representations-of-complex-engineering-systems-spring-2010/readings/MITESD_342S10_ntwk_metrics.pdf

16. Page, Lawrence, et al. "The PageRank citation ranking: bringing order to the web." (1999).

17. http://efavdb.com/predicting-san-francisco-crimes/

References

18. Ferreira, R. S. “Complex Networks Applied to Missing People Problem: A case study about the missing people phenomenon in Brazil” 2016 (to appear)

19.http://snap.stanford.edu/data/

20. https://aws.amazon.com/datasets/

21. http://networkdata.ics.uci.edu/index.php

22. http://www3.nd.edu/~networks/resources.htm

23. Holder, Lawrence B., et al. "Current and Future Challenges in Mining Large Networks: Report on the Second SDM Workshop on Mining Networks and Graphs." ACM SIGKDD Explorations Newsletter 18.1 (2016): 39-45.

References

24. Mears, David, and Harvey B. Pollard. "Network science and the human brain: Using graph theory to understand the brain and one of its hubs, the amygdala, in health and disease." Journal of neuroscience research (2016).

25. Xu, Qiang, et al. "Device fingerprinting in wireless networks: challenges and opportunities." IEEE Communications Surveys & Tutorials 18.1 (2016): 94-104.

26. Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman. "Power-law distributions in empirical data." SIAM review 51.4 (2009): 661-703.

27. https://en.wikipedia.org/wiki/Six_Degrees_of_Kevin_Bacon

Data Mining

Data mining is the process of discovering interesting patterns and knowledge from a large amounts of data [1]

Applications: Customer segmentation, disease patterns, fraud detection, news categorization, market basket, bio informatics ...

Fields: Anomaly detection, Association rule mining, Clustering, Regression, Classification, streaming mining, graph mining ...

Metrics: RMSE, MAE, AUC ...

Databases: KDNuggets data repository [2], UCI KDD archive [3]

Data Mining

Metrics:

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$ (Mean Absolute Error)

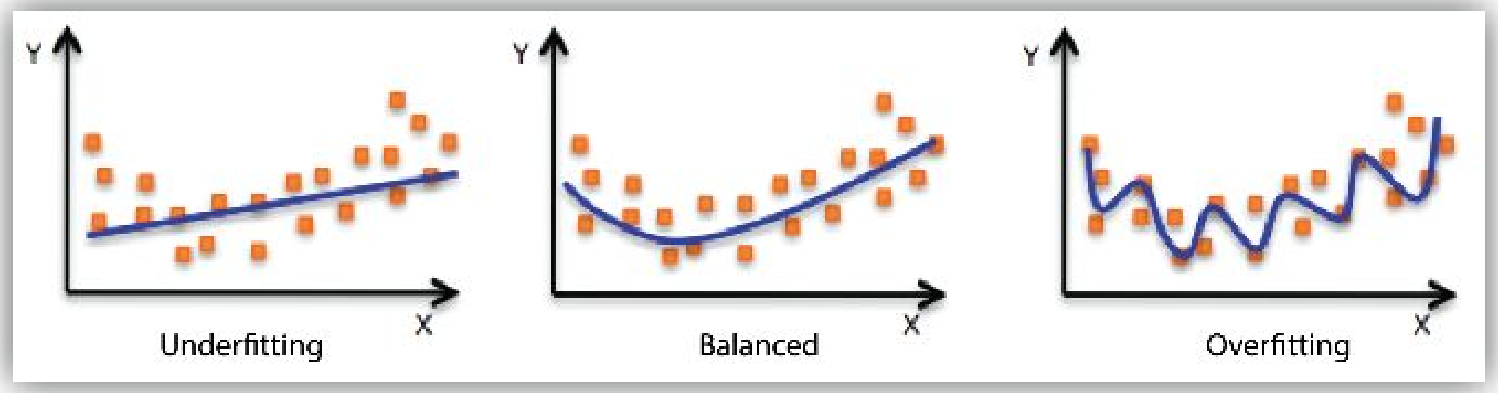
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$ (Root Mean Squared Error)

RMSE amplifies and severely punishes large errors

Data Mining

Overfitting: Too tight. Model too complex. Needs less parameters.

Underfitting: Too relaxed. Model too simple. Needs more parameters.



Source: http://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

Data Mining

Cross Validation: Measuring the predictive performance of a statistical model

K-Fold

Train with K-1 folds and test with the remaining fold. Make it interactively to all folders (K times)

Holdout

Train with a fixed part of data (⅔) and test with the remaining data (⅓)

Leave-one-out

Makes the same of K-Fold but instead a fold, it makes the interaction through all observations

Data Mining

Association rules

The most common algorithm is called Apriori

Database

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C_1

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

C_2

Itemset	Support
{1 2}	1
{1 3}*	2
{1 5}	1
{2 3}*	2
{2 5}*	3
{3 5}*	2

C_3

Itemset	Support
{2 3 5}*	2

Every subset of a frequent itemset is also frequent.

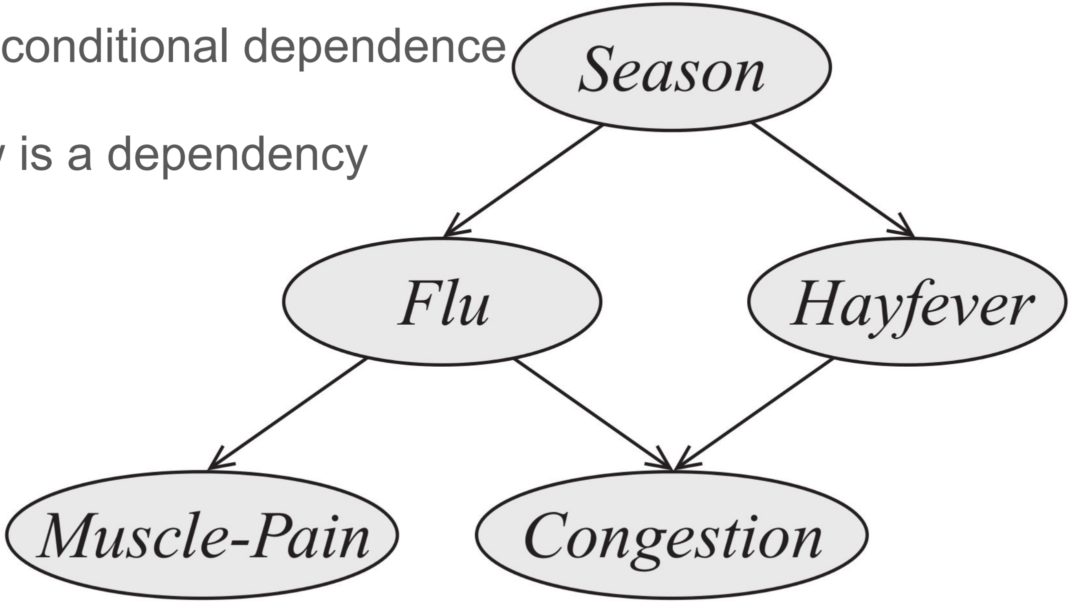
Source: http://www.saedsayad.com/association_rules.htm

Data Mining

Probabilistic Graphical Models (PGM)

Expresses conditional dependence

Each arrow is a dependency



Source: http://pgm.stanford.edu/Figures/Chapter1/01_01a.pdf

Data Mining

Logistic Regression

Measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a **logistic function**

Logistic function $= \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 \cdot x_{i2} + \beta_2 \cdot x_{i2} + \ldots + \beta_p \cdot x_{in}$

Often used with continous data

Data Mining

Nayve Bayes

Naive Bayes assumes conditional independence among discrete variables

Prior Probability

Likelihood of the evidence 'E' if the Hypothesis 'H' is true

$$P(H|E) = \frac{P(H) * P(E|H)}{P(E)}$$

Posterior Probability of 'H' given the evidence

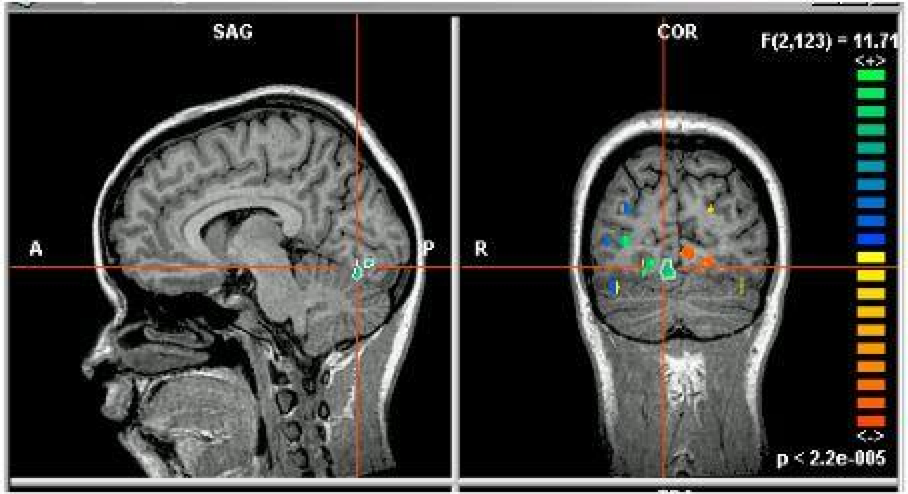
Priori probability that the evidence itself is true

Data Mining

What if we have continuous variables?

Naïve Bayes + continuous variables = **Gaussian Nayve Bayes**

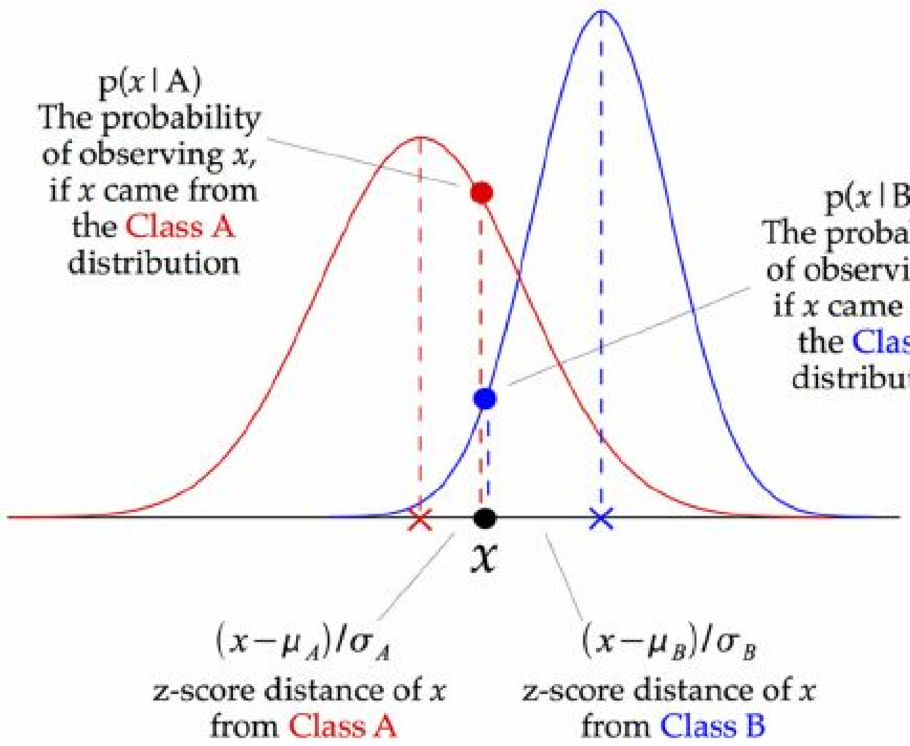
Eg., image classification: X_i is real-valued i^{th} pixel



Data Mining

Gaussian Nayve Bayes

x came from A or B ?



Data Mining

Kernel density estimation (KDE)

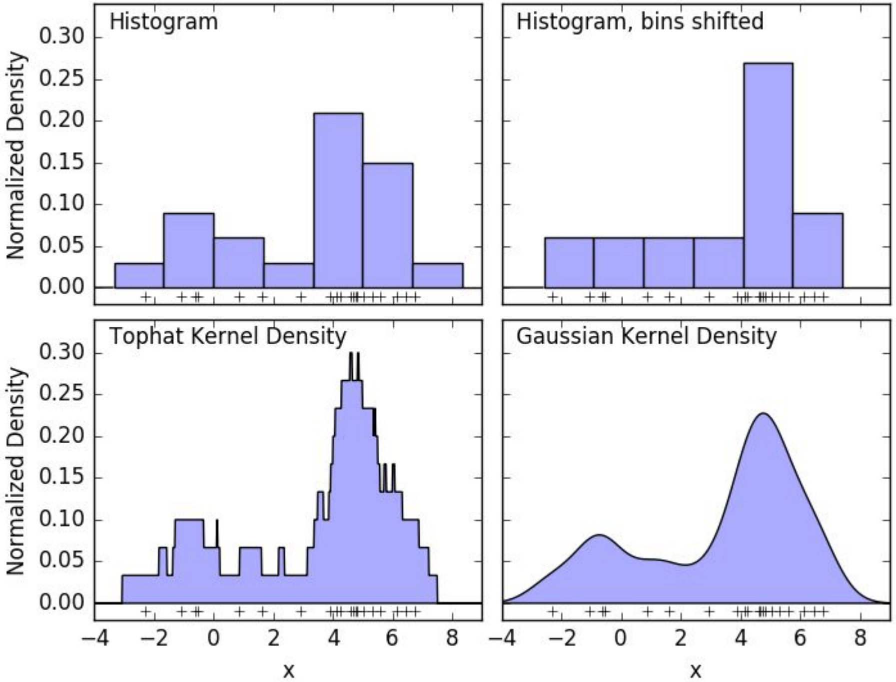
Non parametric statistical method

Has many kernel methods

Gaussian kernel is the most common

Estimates the PDF

(Probability Density Function)

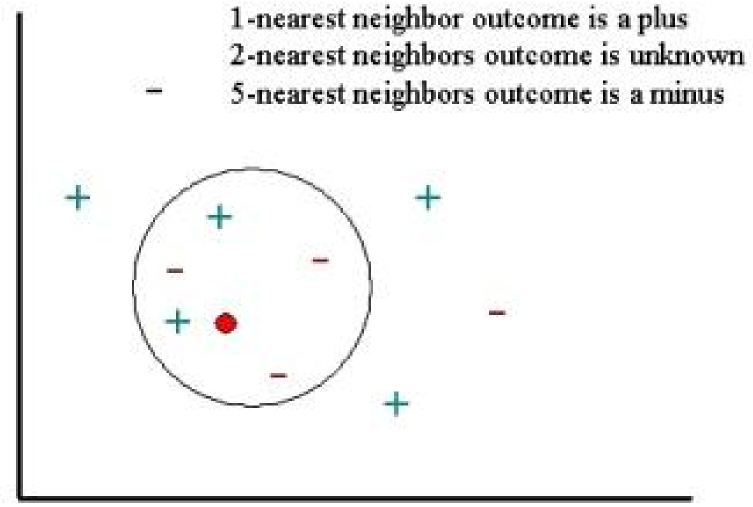


Data Mining

K-NN (K-Nearest Neighbors)

K neighbors near from the data votes to

determine what the class of the unlabeled data



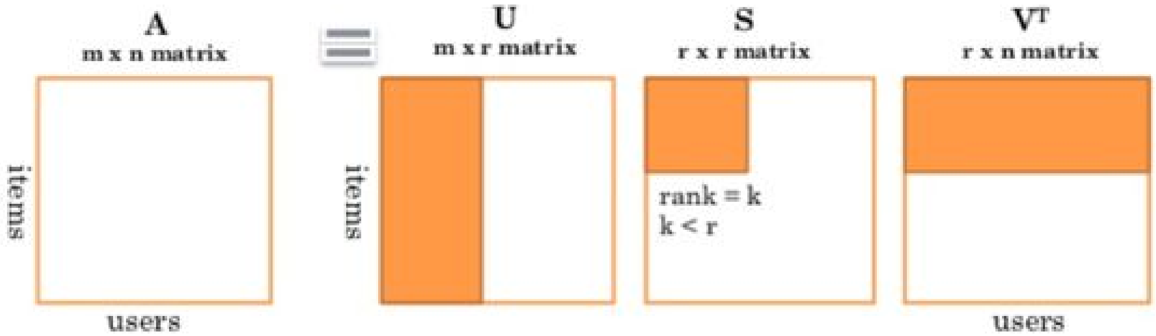
Source: http://www.statsoft.com/textbook/k-nearest-neighbors

Data Mining

Dimensionality reduction techniques:

SVD (Singular Value Decomposition)

$$A = U \times S \times V^T$$

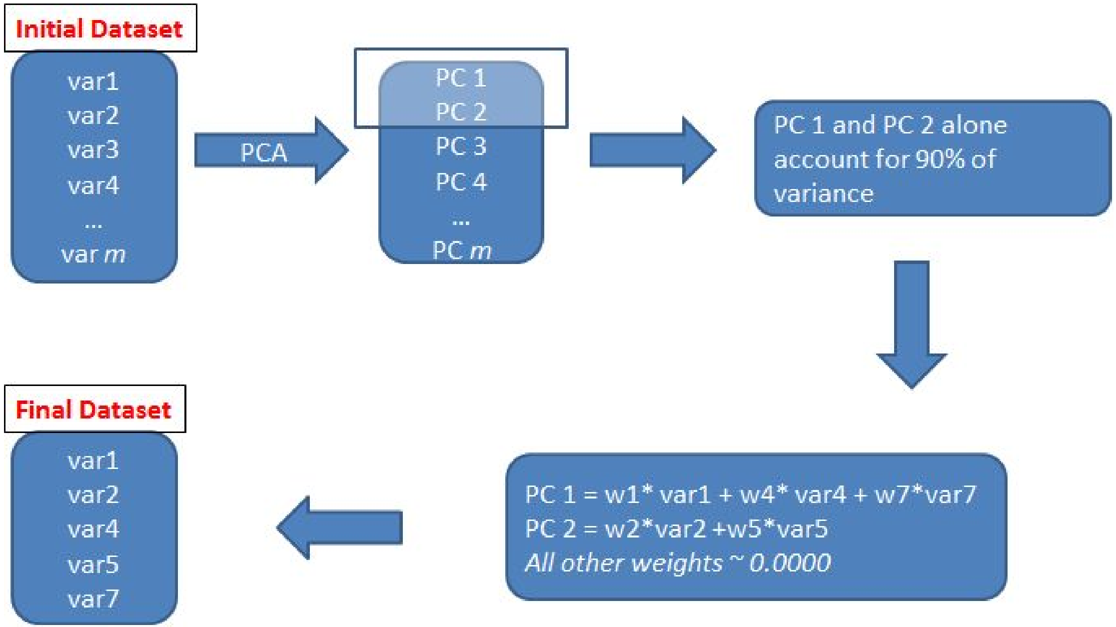


$$A_k = U_k \times S_k \times V_k^T$$

Data Mining

Dimensionality reduction techniques:

PCA (Principal Component Analysis)



Data Mining

Nayve Bayes, Gaussian Nayve Bayes, K-NN, PGM, Statistical recommendation

https://github.com/raulsenaferreira/Systems-Engineering/tree/master/Data%20Mining/Tests

2016 / 09

Data Mining

Dimensionality reduction and Clustering

https://github.com/raulsenaferreira/Systems-Engineering/tree/master/Data%20Mining/Work_1

2016 / 09

Data Mining

Challenges and current research:

IoT Big Data Stream Mining [4]

Data mining and machine learning in cybersecurity [5]

References

1. Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011
2. http://www.kdnuggets.com/datasets/index.html
3. http://kdd.ics.uci.edu/
4. De Francisci Morales, Gianmarco, et al. "IoT Big Data Stream Mining." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.
5. Dua, Sumeet, and Xian Du. Data mining and machine learning in cybersecurity. CRC press, 2016.

Machine Learning

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed [1]

Applications: Help students in online education, classify diseases, recommend movies, recognize faces, predicting city traffic [9]

Fields: Classification, Regression, Supervised learning, Unsupervised learning, Semi-supervised learning

Metrics: RMSE, MAE, NDCG, AUC ...

Databases: UC Irvine Machine Learning Repository [2], Movie Lens [3], Kaggle datasets [4], Epinions [13]

Machine Learning

Classifiers

Kind of problems: text categorization, fraud detection, optical character recognition, market segmentation, natural-language processing, machine vision []

Machine Learning

Binary classifiers

0 or 1 (Yes or No) classification problems

Multi-class classifiers

More than two states, ex:

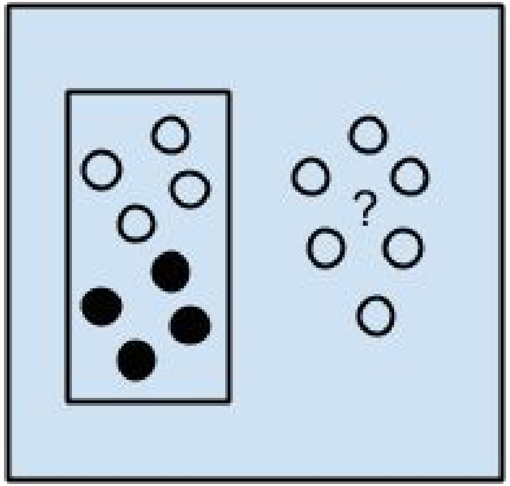
Predicting weather condition (rain, snow, tornado, heat...)

Machine Learning

Supervised Learning

classification

regression



Supervised Learning Algorithms

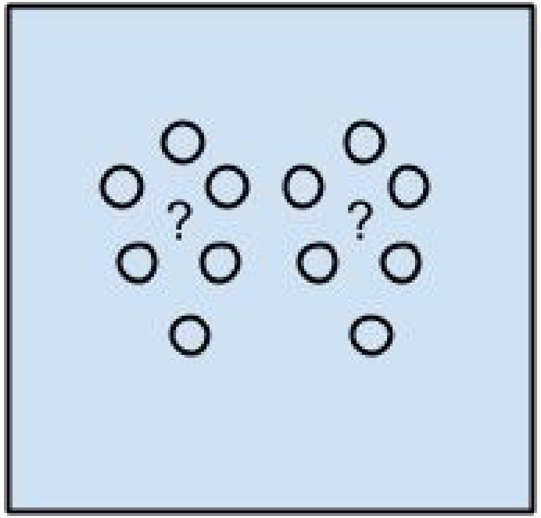
Machine Learning

Unsupervised Learning

clustering

dimensionality reduction

association rule learning



Unsupervised Learning
Algorithms

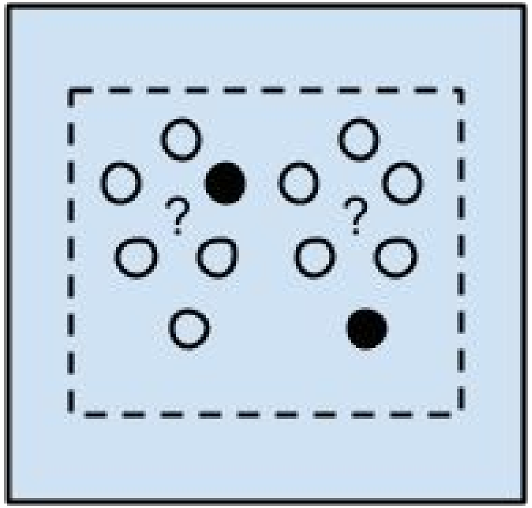
Machine Learning

Semi-supervised learning

classification

regression

(Assuming unlabeled data)



Semi-supervised
Learning Algorithms

Machine Learning

Ensemble methods [15]

Use many algorithms to predict, producing multiple models and combining them achieving improved results.

Majority Voting, Weighted Voting, Simple Averaging, Weighted Averaging

Bagging

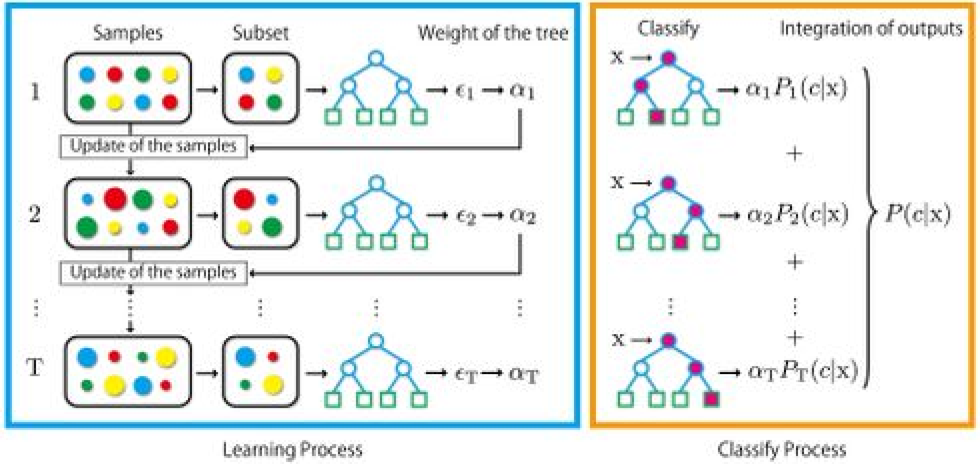
Boosting

Machine Learning

Bagging

Uses bootstrap technique: Statistical method for estimating a quantity from a data sample

Bootstrap Aggregation = **Bagging => Random Forests**



Machine Learning

Boosting

Family of algorithms which converts weak learner to strong learners

AdaBoost

Gradient Tree Boosting

XGBoost

Machine Learning

Active learning

Most used algorithms Uncertainty Sampling, QBC

Basically cut the cardinality of the data

Reduce the computational effort to train the model

Machine Learning

Deep Learning algorithms

Convolutional Neural Networks

Often applied for video classification (face recognition, image representation) [10]

Recurrent Neural Networks

Applied to speech recognition tasks (NLP problems, translation) [11]

Stacked Denoising Autoencoders

Applied to extract representative features for learning tasks [12]

Machine Learning

Some experiments:

1. Simple recommending (by user, item and global average)
   a. https://github.com/raulsenaferreira/Systems-Engineering/blob/master/TEBD%20VI/secondList.jl
2. K-NN and Improved Regularized SVD recommenders
   a. https://github.com/raulsenaferreira/Systems-Engineering/blob/master/TEBD%20VI/thirdList.jl
   b. https://github.com/raulsenaferreira/Systems-Engineering/tree/master/Data%20Mining/Recommender
3. Kaggle competitions
   a. https://github.com/raulsenaferreira/Kaggle/tree/master/Animal_Shelter
   b. https://github.com/raulsenaferreira/Kaggle/tree/master/Titanic_Comptetition
4. Collaborative Filtering
   a. https://github.com/raulsenaferreira/Recsys.jl

2016 / 09

Machine Learning

Challenges and current research:

Explaining Deep learning models [5]

Learning in non-stationary environments (concept drift) [6]

Efficient Transfer Learning algoritms [7]

Transductive / Semi-Supervised Learning in stream data [8]

References

1. http://whatis.techtarget.com/definition/machine-learning
2. http://archive.ics.uci.edu/ml/
3. http://grouplens.org/datasets/movielens/
4. https://www.kaggle.com/datasets
5. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why Should I Trust You?": Explaining the Predictions of Any Classifier." arXiv preprint arXiv:1602.04938 (2016).
6. Sugiyama, Masashi, and Motoaki Kawanabe. Machine learning in non-stationary environments: Introduction to covariate shift adaptation. MIT Press, 2012.

References

7. Lu, Jie, et al. "Transfer learning using computational intelligence: a survey." Knowledge-Based Systems 80 (2015): 14-23.

8. Marian Puscas, Mihai, et al. "Unsupervised tube extraction using transductive learning and dense trajectories." Proceedings of the IEEE International Conference on Computer Vision. 2015.

9. http://research.ibm.com/cognitive-computing/machine-learning-applications/index.shtml#fbid=_tr0VGk0FtR

10. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

References

11. Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, 2013.

12. Vincent, Pascal, et al. "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." Journal of Machine Learning Research 11.Dec (2010): 3371-3408.

13. http://www.epinions.com/

14. http://www.cs.princeton.edu/~schapire/talks/picasso-minicourse.pdf

15. Dietterich, Thomas G. "Ensemble methods in machine learning." International workshop on multiple classifier systems. Springer Berlin Heidelberg, 2000.

Interested in dig deeper?

A Fast Semi-supervised learning framework for non-stationary environments

- Concept drift
- Semi-supervised machine learning
- Non-parametric statistical methods
- Active learning
- Temporal Series
- Unlabeled data

Interested in dig deeper?

Complex networks and knowledge discovery applied to missing people problem

- Unstructured and heterogeneous data
- Data integration
- ETL processes
- Kernel methods
- Communities detection
- Graph theory

Interested in dig deeper?

Big Data processing with scalable regularized gradient boosting trees: Building a scalable support system for Early Diagnosis of Alzheimer Disease

- XGBoost algorithm improvements
- Ensemble methods in machine learning
- Mobile and cloud environments
- Data sparsity in classification problems

Get in touch

raul@ipea.gov.br

raulsf@cos.ufrj.br

raulsenaferreira@gmail.com

https://br.linkedin.com/in/raulsenaferreira

https://github.com/raulsenaferreira