
CSE 40647/60647
"Data Science" Fall 2017:
Introduction to Data Mining

Chapter 1. Introduction

Meng Jiang

CSE 40647/60647 Data Science Fall 2017

Introduction to Data Mining

Course History

- Taught in the University of Illinois at Urbana-Champaign, Summer 2017
- Course project: Data Science for Data Science
  - Can we mine data-science publications with data-driven technologies for data-scientific knowledge discover and technology exploration?

data

method

problem

Example: Knowledge Discovery

AAAI'06

Efficient L_1 Regularized Logistic Regression

Su-In Lee, Honglak Lee, Pieter Abbeel and Andrew Y. Ng

Computer Science Department
Stanford University
Stanford, CA 94305

Abstract

L_1 regularized logistic regression is now
machine learning: it is widely used for r
tion problems, particularly ones with ma
regularized logistic regression requires s
optimization problem. However, stand
for solving convex optimization proble
well enough to handle the large datase
in many practical settings. In this pap
an efficient algorithm for L_1 regularized
sion. Our algorithm iteratively approxin
tive function by a quadratic approximati
point, while maintaining the L_1 constrai
ation, it uses the efficient LARS (Least
sion) algorithm to solve the resulting
quadratic optimization problem. Our th
show that our algorithm is guaranteed to
global optimum. Our experiments show
rithm significantly outperforms standard
solving convex optimization problems. Moreover, our algorithm outperforms four previously published algorithms that were specifically designed to solve the L_1 regularized logistic regression problem.

	Dataset	Problem	Method
Paper-AAAI'06		convex optimization	logistic regression; L1 regularization

in a continuously differentiable objective. These (and similar) observations have led to various algorithms that try to exploit the structure in the optimization problem and solve it more efficiently.

Example: Knowledge Discovery

CS229

Evaluating the effectiveness of regularized logistic regression for the Netflix movie rating prediction task

Adam Sadovsky
sadovsky@cs.

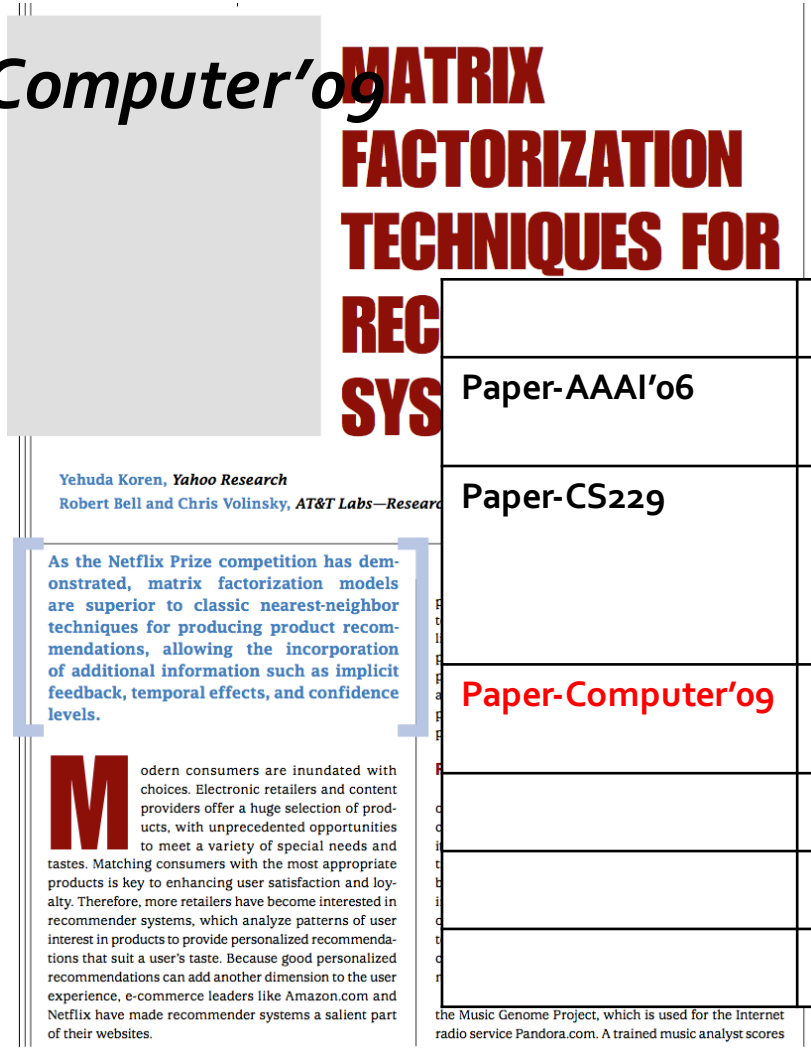
Xing Chen

1 Introduction

Netflix Prize is a comp
will rate movies he or
Computer Science De
developing and testing
clustering, various typ
prediction methods th
approximately 0.93 ro
regularization impose
all features; for variou
results in zero weight
models (such as Mark

	Dataset	Problem	Method
Paper-AAAI'06		convex optimization	logistic regression; L1 regularization
Paper-CS229	netflix	rating prediction	logistic regression; L1 regularization; L2 regularization; matrix factorization

Example: Knowledge Discovery


Computer'09

MATRIX FACTORIZATION TECHNIQUES FOR REC SYS

Yehuda Koren, *Yahoo Research*
Robert Bell and Chris Volinsky, *AT&T Labs—Researc*

As the Netflix Prize competition has demonstrated, matrix factorization models are superior to classic nearest-neighbor techniques for producing product recommendations, allowing the incorporation of additional information such as implicit feedback, temporal effects, and confidence levels.

Modern consumers are inundated with choices. Electronic retailers and content providers offer a huge selection of products, with unprecedented opportunities to meet a variety of special needs and tastes. Matching consumers with the most appropriate products is key to enhancing user satisfaction and loyalty. Therefore, more retailers have become interested in recommender systems, which analyze patterns of user interest in products to provide personalized recommendations that suit a user's taste. Because good personalized recommendations can add another dimension to the user experience, e-commerce leaders like Amazon.com and Netflix have made recommender systems a salient part of their websites.

the Music Genome Project, which is used for the Internet radio service Pandora.com. A trained music analyst scores

	Dataset	Problem	Method
Paper-AAAI'06		convex optimization	logistic regression; L1 regularization
Paper-CS229	netflix	rating prediction	logistic regression; L1 regularization; L2 regularization; matrix factorization
Paper-Computer'09	netflix	rating prediction	matrix factorization; regularization

Example: Knowledge Discovery

WWW'07

Review Spam Detection

Nitin Jindal and Bing Liu
Department of Computer Science
University of Illinois at Chicago
851 South Morgan Street
Chicago, IL 60607-7053
nitin.jindal@gmail.com, liub@cs.uic.edu

ABSTRACT

It is now a common practice for e-con
their customers to write reviews of
purchased. Such reviews provide valua
on these products. They are used by
opinions of existing users before deci
They are also used by product manufa
of their products and to find competiti
about their competitors. Unfortunately,
also gives good incentive for *spam*, w
or malicious negative opinions. In this
to study review spam and spam dete
knowledge, there is still no reported stu

Categories and Subject Des

H.3.3 [**Information Storage and Reti**
and Retrieval – *Information filtering*.

General Terms: Experimentatio

Keywords: Product reviews, revie

	Dataset	Problem	Method
Paper-AAAI'06		convex optimization	logistic regression; L1 regularization
Paper-CS229	netflix	rating prediction	logistic regression; L1 regularization; L2 regularization; matrix factorization
Paper-Computer'09	netflix	rating prediction	matrix factorization; regularization
Paper-WWW'07	amazon	review spam detection; classification	logistic regression

Example: Technology Exploration

AAAI'12

Discovering Spammers in Social Networks

Yin Zhu[†], Xiao Wang[⋆], Erheng Zhong[†], Nanthan N. Liu[†], He Li[⋆], Qiang Yang[†]
[†]Hong Kong University of Science and Technology, Hong Kong
[⋆]Renren Inc., China
{yinz,ezhong

Abstract

As the popularity of the social me
evidenced in Twitter, Facebook and
spamming activities also picked up
variety. On social network sites, spa
guise themselves by creating fake acc
ing normal users' accounts for person
from the spammers in traditional syst
and email, spammers in social media
mal users and they continue to chang
strategies to fool anti-spamming syste
to the privacy and resource concerns,
dia websites cannot fully monitor a
users, making many of the previous ap
topology-based and content-classifica
ods, infeasible to use. In this paper,
pervised Matrix Factorization method

	Dataset	Problem	Method
Paper-AAAI'06		convex optimization	logistic regression; L1 regularization
Paper-CS229	netflix	rating prediction	logistic regression; L1 regularization; L2 regularization; matrix factorization
Paper-Computer'09	netflix	rating prediction	matrix factorization; regularization
Paper-WWW'07	amazon	review spam detection; classification	logistic regression
Paper-AAAI'12	renren	social spam detection; classification	???

Example: Technology Exploration

AAAI'12

Discovering Spammers in Social Networks

Yin Zhu[†], Xiao Wang[⋆], Erheng Zhong[†], Nanthan N. Liu[†], He Li[⋆], Qiang Yang[†]
[†]Hong Kong University of Science and Technology, Hong Kong
[⋆]Renren Inc., China
{yinz,ezhong

Abstract

As the popularity of the social me
evidenced in Twitter, Facebook and
spamming activities also picked up
variety. On social network sites, spa
guise themselves by creating fake acc
ing normal users' accounts for person
from the spammers in traditional syst
and email, spammers in social media
mal users and they continue to chang
strategies to fool anti-spamming syste
to the privacy and resource concerns,
dia websites cannot fully monitor a
users, making many of the previous ap
topology-based and content-classifica
ods, infeasible to use. In this paper,
pervised Matrix Factorization method

	Dataset	Problem	Method
Paper-AAAI'06		convex optimization	logistic regression; L1 regularization
Paper-CS229	netflix	rating prediction	logistic regression; L1 regularization; L2 regularization; matrix factorization
Paper-Computer'09	netflix	rating prediction	matrix factorization; regularization
Paper-WWW'07	amazon	review spam detection; classification	logistic regression
Paper-AAAI'12	renren	social spam detection; classification	matrix factorization

Example: Technology Exploration

AAAI'14

Online Social Spammer Detection

Xia Hu, Jiliang Tang, Huan Liu
Computer Science and Engineering, Arizona State University, USA
{xiahu, jiliang.tang, huan.liu}@asu.edu

Abstrac

The explosive use of social medi
platform for malicious users, kno
overwhelm normal users with unv
tive way for social spammer detec
based on content and social netwo
social spammers are sophisticated
system with fast evolving content a
social spammers continually chan
patterns to avoid being detected. S
makes it easier for social spamme
ence and pretend to be normal use
a large number of "human" friend
isting anti-spamming systems bas
to quickly respond to newly eme
social spammer detection. In thi
eral optimization framework to c
network information for social sp
vide the solution for efficient onli
tal results on Twitter datasets cor
efficiency of the proposed framew

	Dataset	Problem	Method
Paper-AAAI'06		convex optimization	logistic regression; L1 regularization
Paper-CS229	netflix	rating prediction	logistic regression; L1 regularization; L2 regularization; matrix factorization
Paper-Computer'09	netflix	rating prediction	matrix factorization; regularization
Paper-WWW'07	amazon	review spam detection; classification	logistic regression
Paper-AAAI'12	renren	social spam detection; classification	matrix factorization
Paper-AAAI'14	twitter	social spam detection; convex optimization	matrix factorization; regularization

Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!

What is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?

About Me: Data Mining

- Dr. Meng Jiang (www.meng-jiang.com)

B.S. and Ph.D.

Visiting Ph.D.

Postdoc Researcher

Assistant Professor

Visiting Researcher

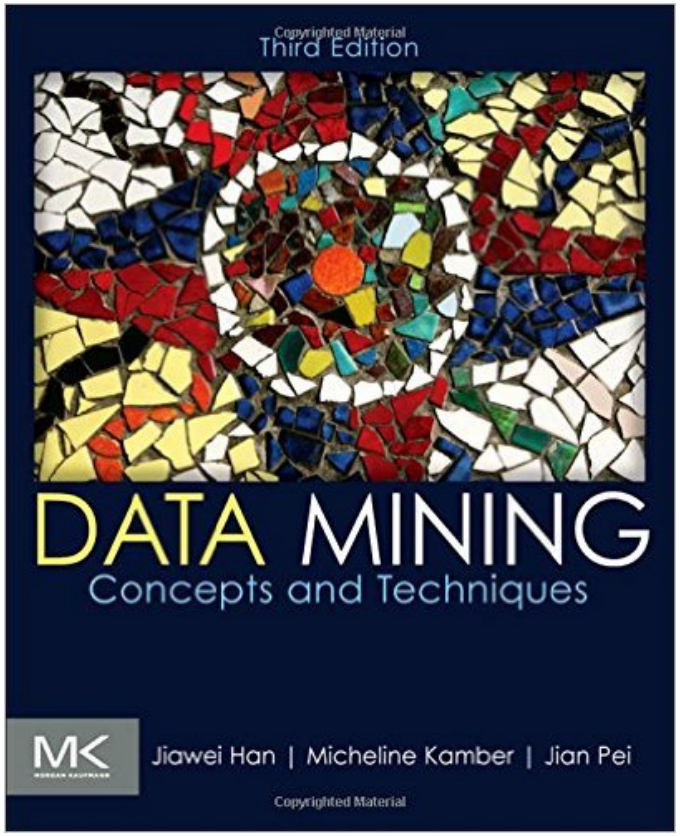
Visiting Researcher

TODO

- Official description
- Course information (hours, credits)
- Prerequisites
- Syllabus and schedule
- Class time and location
- Office hours
- Teaching assistant
- TA office hours
- Piazza
- Compass
- Course work and grading
- Assignments
- Exams
- Project

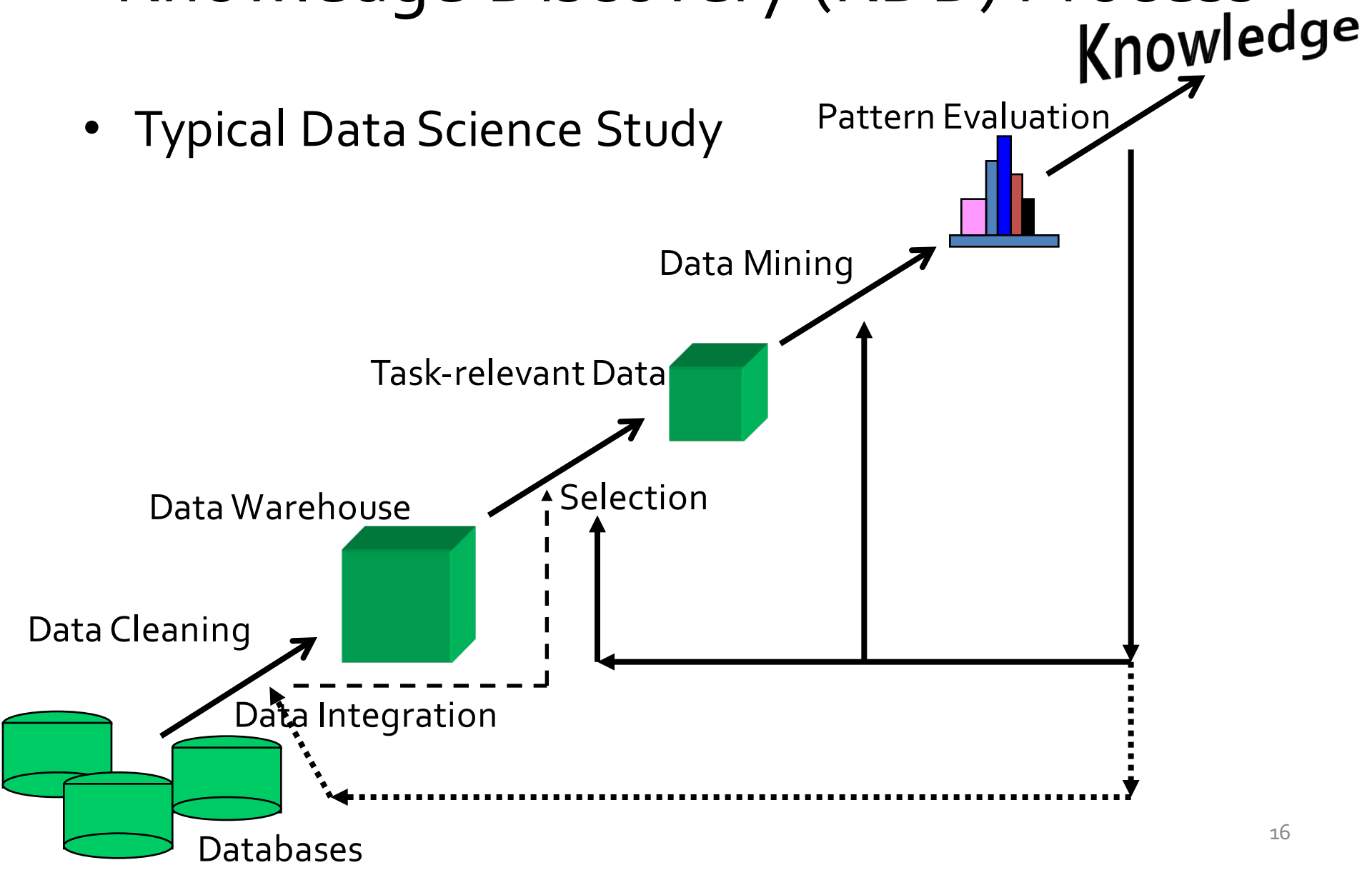
Textbook

- Jiawei Han, Micheline Kamber and Jian Pei, Data Mining: Concepts and Techniques (3rd ed), Morgan Kaufmann, 2011



Knowledge Discovery (KDD) Process

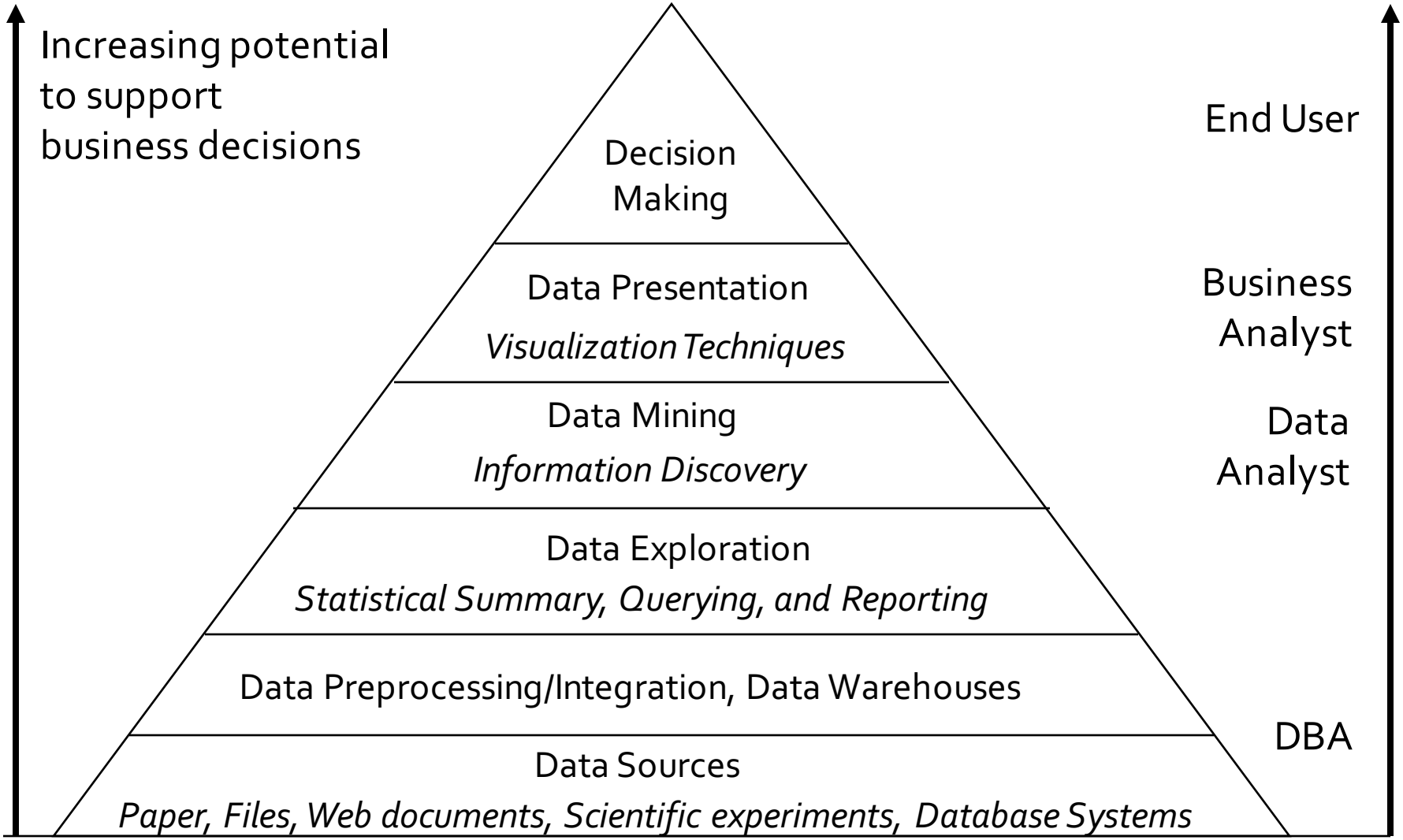
- Typical Data Science Study



Example: Web Mining

- Web mining usually involves
  - Data cleaning
  - Data integration from multiple sources
  - Warehousing the data
  - Data cube construction
  - Data selection for data mining
  - Data mining
  - Presentation of the mining results
  - Patterns and knowledge to be used or stored into knowledge-base

Example: Business Intelligence

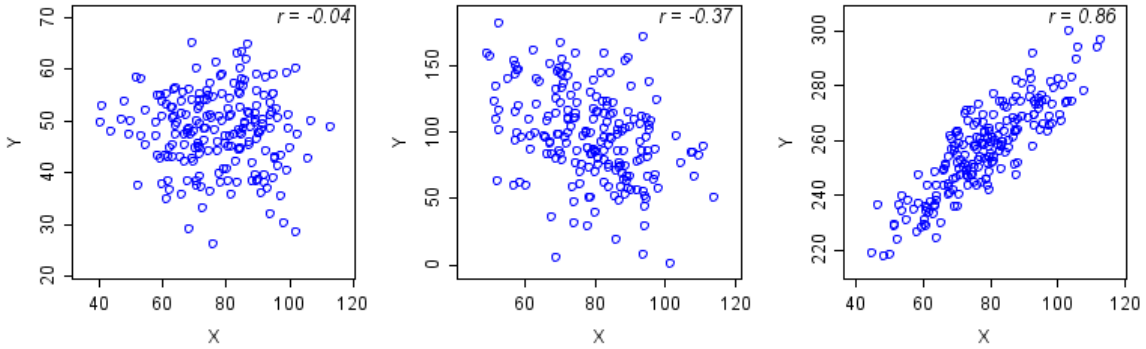


Data Mining Functions: (1) Generalization

- Information integration and data warehouse construction
- Data cube technology
- Multidimensional concept description: Characterization and discrimination

Data Mining Functions: (2) Pattern Discovery

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- A typical association rule
  - Diaper → Beer [0.5%, 75%] (support, confidence)
  - **Support:** the proportion of transactions in the dataset which contains the itemset (Diaper, Beer).
  - **Confidence:** the proportion of the transactions that contains Diaper which also contains Beer.
- **Association** and **Correlation** Analysis

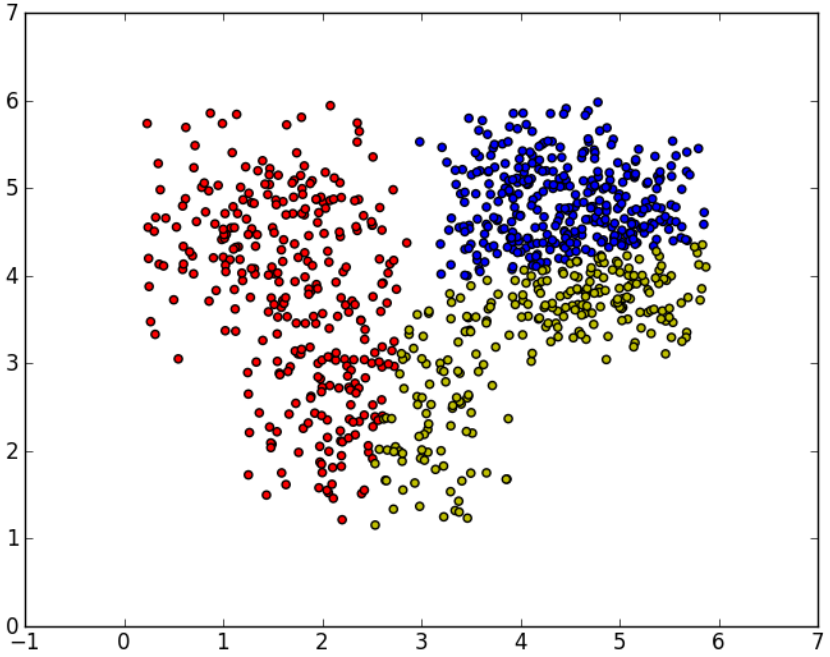


Data Mining Functions: (3) Classification

- Classification and label prediction
  - Construct models (functions) based on some **training** examples
  - Distinguish classes or concepts for future prediction
    - Ex. 1. Classify countries based on (climate)
    - Ex. 2. Classify cars based on (gas mileage)
  - Predict some unknown **class labels**
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …

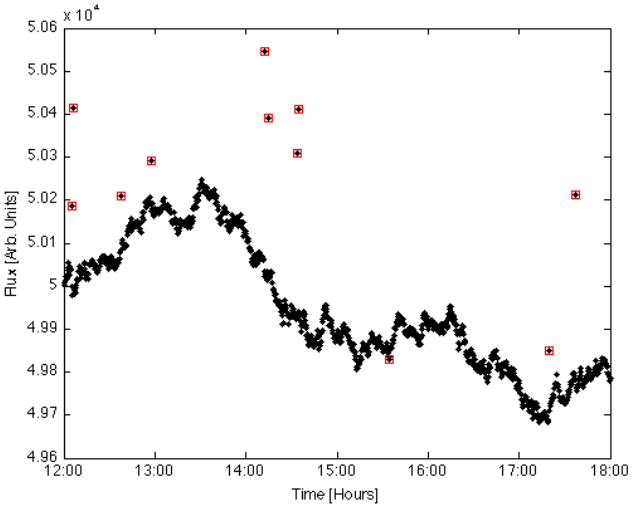
Data Mining Functions: (4) Clustering

- Unsupervised learning (i.e., class label is unknown)
- Group data to form new categories (i.e., clusters)
- Principle: Maximizing intra-cluster similarity & minimizing inter-cluster similarity



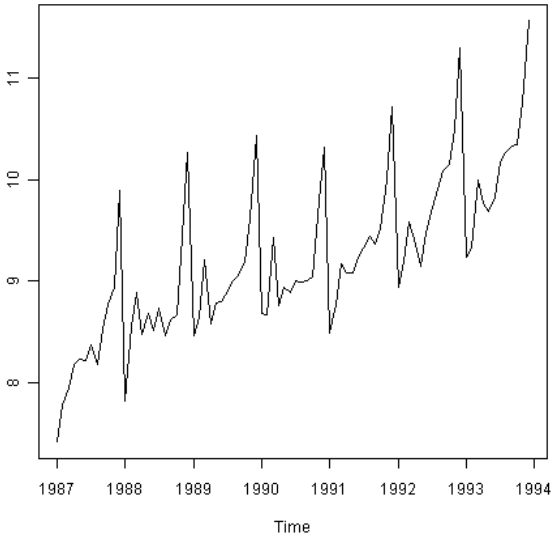
Data Mining Functions: (5) Outlier Analysis

- Outlier analysis
  - Outlier: A data object that does not comply with the general behavior of the data
  - Methods: by product of clustering or regression analysis…
  - Useful in fraud detection, rare events analysis



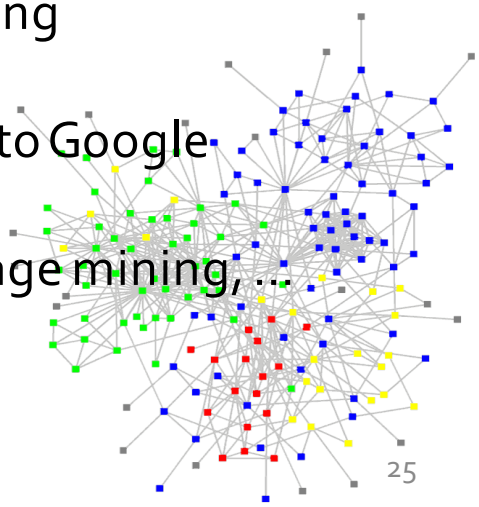
Data Mining Functions: (6) Sequential Pattern, Evolution Analysis

- Sequence, trend and evolution analysis
  - Trend, time-series, and deviation analysis
    - e.g., regression and value prediction
  - Sequential pattern mining
    - e.g., buy digital camera, then buy large memory cards
  - Periodicity analysis
  - Motifs and biological sequence analysis
    - Approximate and consecutive motifs
  - Similarity-based analysis



Data Mining Functions: (7) Structure and Network Analysis

- Graph mining
  - Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- Information network analysis
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g., author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be multiple information networks: friends, family, classmates, …
  - Links carry a lot of semantic information: Link mining
- Web mining
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
    - Web community discovery, opinion mining, usage mining, …



Applications

- Web page analysis: classification, clustering, ranking
- Collaborative analysis & recommender systems
- Basket data analysis to targeted marketing
- Biological and medical data analysis
- Data mining and software engineering
- Data mining and text analysis
- Data mining and social and information network analysis
- Scientific knowledge discovery and technology exploration

Syllabus

- Introduction
- Getting to Know Your Data: Data Description
- Getting to Know Your Data: Visualization and Measures
- Data Preprocessing: Data Cleaning and Data Integration
- Data Preprocessing: Data Reduction and Dimension Reduction
- Data Cube: Concepts and Operations
- Data Cube: Computation Methods
- Data Cube: Data Warehouse and OLAP
- Frequent Pattern Mining: Concepts and Apriori
- Frequent Pattern Mining: FP-Growth
- Frequent Pattern Mining: Pattern Evaluation
- Advanced Frequent Pattern Mining: Diverse Patterns and Constraint-based Frequent Pattern Mining
- Advanced Frequent Pattern Mining: Sequential Pattern Mining and Graph Pattern Mining
- Classification: Concepts and Decision Tree Induction
- Classification: Bayesian Classification
- Classification: Evaluation and Ensemble
- Advanced Classification: Support Vector Machines
- Advanced Classification: Neural Networks
- Clustering: Concepts
- Clustering: Partitioning Methods
- Clustering: Kernel-based Clustering
- Clustering: Density-based Clustering
- Clustering: Evaluation Methods
- Outlier Analysis: Concepts
- Outlier Analysis: Methods
- Lecture: Mining Behavioral Datasets?
- Lecture: Mining Text Datasets?

The 18 Identified Candidates

- Classification
  - √ – #1. C4.5: Quinlan, J. R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc.
  - √ – #2. CART: L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, Belmont, CA, 1984.
  - – #3. K Nearest Neighbours (kNN): Hastie, T. and Tibshirani, R. 1996. Discriminant Adaptive Nearest Neighbor Classification. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI). 18, 6 (Jun. 1996), 607-616.
  - √ – #4. Naive Bayes Hand, D.J., Yu, K., 2001. Idiot's Bayes: Not So Stupid After All? Internat. Statist. Rev. 69, 385-398.
- Statistical Learning
  - √ – #5. SVM: Vapnik, V. N. 1995. The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc.
  - – #6. EM: McLachlan, G. and Peel, D. (2000). Finite Mixture Models. J. Wiley, New York.
- Association Analysis
  - √ – #7. Apriori: Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In VLDB '94.
  - √ – #8. FP-Tree: Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without candidate generation. In SIGMOD '00.
- Link Mining
  - – #9. PageRank: Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In WWW-7, 1998.
  - – #10. HITS: Kleinberg, J. M. 1998. Authoritative sources in a hyperlinked environment. In Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, 1998.

18 Candidates (2)

- Clustering
  - √ – #11. K-Means: MacQueen, J. B., Some methods for classification and analysis of multivariate observations, in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, 1967.
  - – #12. BIRCH Zhang, T., Ramakrishnan, R., and Livny, M. 1996. BIRCH: an efficient data clustering method for very large databases. In SIGMOD '96.
- Bagging and Boosting
  - √ – #13. AdaBoost: Freund, Y. and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 1 (Aug. 1997), 119-139.
- Sequential Patterns
  - √ – #14. GSP: Srikant, R. and Agrawal, R. 1996. Mining Sequential Patterns: Generalizations and Performance Improvements. In Proceedings of the 5th International Conference on Extending Database Technology, 1996.
  - √ – #15. PrefixSpan: J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M-C. Hsu. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In ICDE '01.
- Integrated Mining
  - – #16. CBA: Liu, B., Hsu, W. and Ma, Y. M. Integrating classification and association rule mining. KDD-98.
- Rough Sets
  - – #17. Finding reduct: Zdzislaw Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Norwell, MA, 1992
- Graph Mining
  - √ – #18. gSpan: Yan, X. and Han, J. 2002. gSpan: Graph-Based Substructure Pattern Mining. In ICDM '02.

Top 10 Algorithms: Summary

- √ **#1: C4.5** (61 votes), presented by Hiroshi Motoda
- √ **#2: K-Means** (60 votes), presented by Joydeep Ghosh
- √ **#3: SVM** (58 votes), presented by Qiang Yang
- √ **#4: Apriori** (52 votes), presented by Christos Faloutsos
- **#5: EM** (48 votes), presented by Joydeep Ghosh
- **#6: PageRank** (46 votes), presented by Christos Faloutsos
- √ **#7: AdaBoost** (45 votes), presented by Zhi-Hua Zhou
- **#7: kNN** (45 votes), presented by Vipin Kumar
- √ **#7: Naive Bayes** (45 votes), presented by Qiang Yang
- √ **#10: CART** (34 votes), presented by Dan Steinberg

History

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM Data Mining (2001), (IEEE) ICDM (2001), WSDM (2008), etc.
- ACM Transactions on KDD (2007)

Venues

- Data mining and KDD (SIGKDD)
  - Conferences: ACM SIGKDD, IEEE ICDM, SIAM DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, IEEE TKDE, ACM TKDD
- Database systems (SIGMOD)
  - Conferences: ACM SIGMOD, ACM-PODS, VLDB, IEEE ICDE, EDBT, ICDT, DASFAA
  - Journals: ACM TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
  - Conferences: ICML, AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journals: Machine Learning, Artificial Intelligence, IEEE PAMI, etc.
- Web and IR
  - Conferences: SIGIR, WWW, CIKM, etc.
  - Journals: Internet and Web Information Systems
- Statistics
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization
  - Conference proceedings: CHI, ACM SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

Summary

- Data mining: Discovering interesting patterns and knowledge from massive amount of data
- A natural evolution of science and information technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, data mining, pattern evaluation, and knowledge presentation
- Data mining functionalities: generalization, association, classification, clustering, trend and outlier analysis, …

Summary

- Data mining: Discovering interesting patterns and knowledge from massive amount of data
- A natural evolution of science and information technology, in great demand, with wide applications

What you can expect from this course:
Fundamental data science theories
Basic concepts and methods for mining datasets

Not included:
State-of-the-art machine learning/artificial intelligence algorithms
Full coverage of specific skills that your start-up ideas require

References

- Charu C. Aggarwal, Data Mining: The Textbook, Springer, 2015
- E. Alpaydin. Introduction to Machine Learning, 2nd ed., MIT Press, 2011
- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. , 2011
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005 (2nd ed. 2016)
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005
- Mohammed J. Zaki and Wagner Meira Jr., Data Mining and Analysis: Fundamental Concepts and Algorithms 2014