



Summarization

- i.e., ‘characterization’ or ‘generalization’
- maps data into subsets with associated simple descriptions
- extracts or derives representative information
- succinctly characterizes the contents

Example

- criteria used to compare universities by the U.S. News & World Report
- the average SAT or ACT score
- estimates the intellectual level of the student body



Summarization Types

1. Computing the mean, **variance**, and other moments (e.g., **skewness**, **kurtosis**)
2. Extremes (min/max)
3. Estimating the number of distinct elements in a set
4. Counting the number of elements and finding frequently occurring ones
5. Calculating order statistics (e.g., median)



Association Rules

- seek to uncover relationships among data
- association rules (AR) are models that identify specific types of data associations
- AR often used in retail sales to find items that are frequently purchased together
 - aka *market basket analysis*
- also used to predict the failure of telecommunication switches

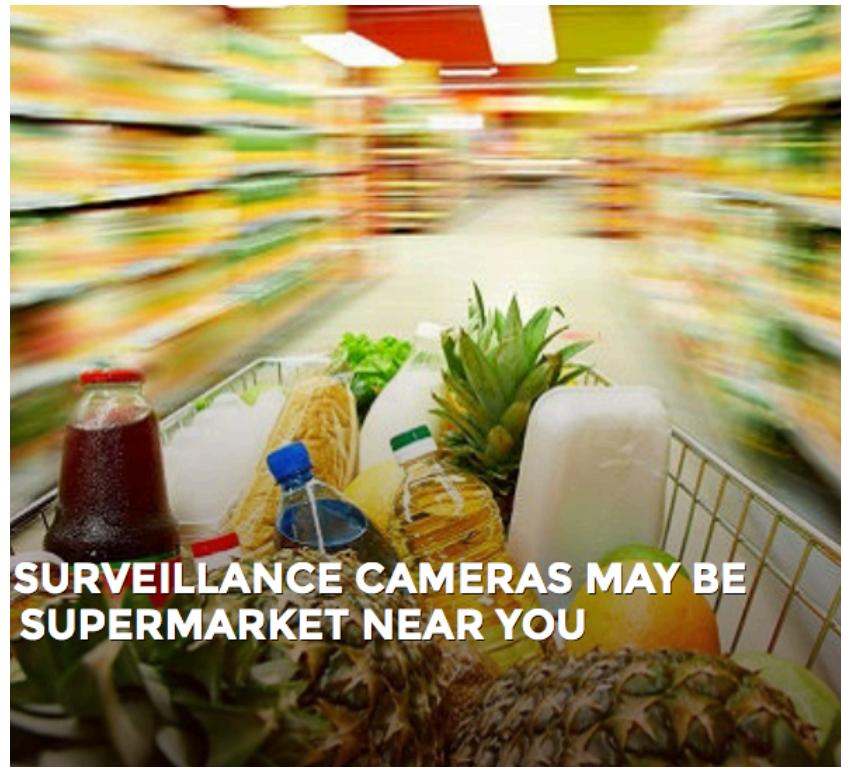
Example

- grocery retailer to decide whether to put bread on sale
- retailer generates AR that show what products are often purchased with bread
 - 60% of the time pretzels are bought
 - 70% of the time jelly is also bought
- ∴, opt to place bread, pretzels, jelly in the same aisle near each other
- AR help determine which products are frequently bought together
 - used for effective advertising, marketing, and inventory control



Association Rule Algorithms

1. A-priori
2. Eclat (Equivalence Class Transformation)
3. FP-growth (Frequent Pattern)



<http://www.inquisitr.com/1106993/store-shelf-surveillance-cameras-may-be-coming-to-a-supermarket-near-you/>



Sequence Discovery (SD)

- used to find sequential patterns in data, based on a time sequence of actions
- unlike market basket analysis, which requires items be purchased at the same time, in SD the **items are purchased over time in some order**

Example

- webmaster analyzes web log to see how users access their webpages
- goal is to determine what sequence of pages are frequently accessed
- learns that 70% of users of page A follow one of these behavior patterns
 - $\langle A, B, C \rangle$, $\langle A, D, B, C \rangle$, $\langle A, E, B, C \rangle$
- \therefore , a link is added from page A to page C



Sequence Discovery Algorithms

1. GSP (Generalized Sequential Pattern)
2. SPADE (Sequential Pattern Discovery using Equivalence Classes)
3. FreeSpan
4. PrefixSpan
5. MAPres



Anomaly (or Outlier) Detection

- identify unusual data
- that might be interesting or
- errors that require further investigation

Example

- intrusion detection in networks
- discover consistent and useful patterns of system features that describe program and user behavior
- use relevant system features to compute classifiers that can recognize anomalies and known intrusions
- cs.columbia.edu/~wenke/papers/usenix/usenix.html



Anomaly (or Outlier) Detection

- In data mining, **anomaly detection** (or **outlier detection**) is the identification of items, events or observations which **do not conform to an expected pattern** or other **items** in a dataset.^[1]
- Typically the anomalous items will translate to some kind of problem, such as bank fraud, a structural defect, medical problems or errors in a text.
- Anomalies are also referred to as outliers, noise, deviations, and exceptions.^[2]

https://en.wikipedia.org/wiki/Anomaly_detection



Anomaly (or Outlier) Detection

- In the context of **abuse and network intrusion detection**, interesting objects are often **not rare objects**, but unexpected *bursts* in activity.
- This pattern does not adhere to the common statistical definition of an outlier as a rare object, and many outlier detection methods (in particular unsupervised methods) will fail on such data, unless it has been aggregated appropriately.
- Instead, a **cluster analysis algorithm** may be able to detect the microclusters formed by these patterns.^[3]

https://en.wikipedia.org/wiki/Anomaly_detection



Anomaly/Outlier Detection Algorithms

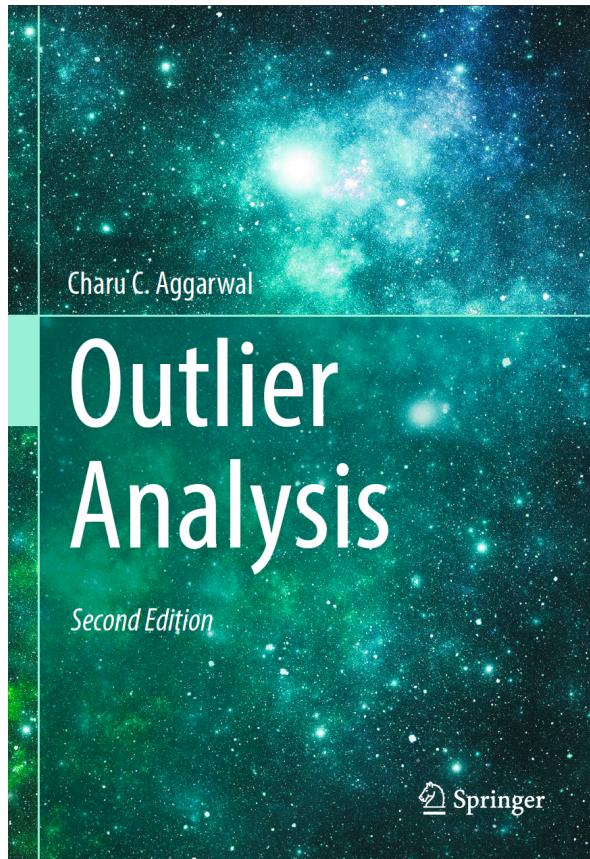


Table 1.1: Classification methods and their unsupervised analogs in outlier analysis

Supervised Model	Unsupervised Analog(s)	Type
k -nearest neighbor	k -NN distance, LOF, LOCI (Chapter 4)	Instance-based
Linear Regression	Principal Component Analysis (Chapter 3)	Explicit Generalization
Naive Bayes	Expectation-maximization (Chapter 2)	Explicit Generalization
Rocchio	Mahalanobis method (Chapter 3) Clustering (Chapter 4)	Explicit Generalization
Decision Trees Random Forests	Isolation Trees Isolation Forests (Chapters 5 and 6)	Explicit generalization
Rule-based	FP-Outlier (Chapter 8)	Explicit Generalization
Support-vector machines	One-class support-vector machines (Chapter 3)	Explicit generalization
Neural Networks	Replicator neural networks (Chapter 3)	Explicit generalization
Matrix factorization (incomplete data prediction)	Principal component analysis Matrix factorization (Chapter 3)	Explicit generalization

Ranking

comparing items

Web search

The image shows a Google search results page. The search bar at the top contains the query "learning to rank". Below the search bar, a dropdown menu lists several suggested search terms: "learning to rank", "learning to rank for information retrieval", "learning to rank using gradient descent", and "learning to rank tutorial". To the right of these suggestions is a blue "I'm Feeling Lucky »" button. A magnifying glass icon is located to the right of the search bar.

Search

Web

[Learning to rank - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Learning_to_rank
Learning to rank or machine-learned ranking (MLR) is a type of supervised or semi-supervised machine learning problem in which the goal is to automatically ...
Applications Feature vectors Evaluation measures Approaches

[Yahoo! Learning to Rank Challenge](#)
learningtorankchallenge.yahoo.com/
Learning to Rank Challenge is closed! Close competition, innovative ideas, and fierce determination were some of the highlights of the first ever Yahoo!

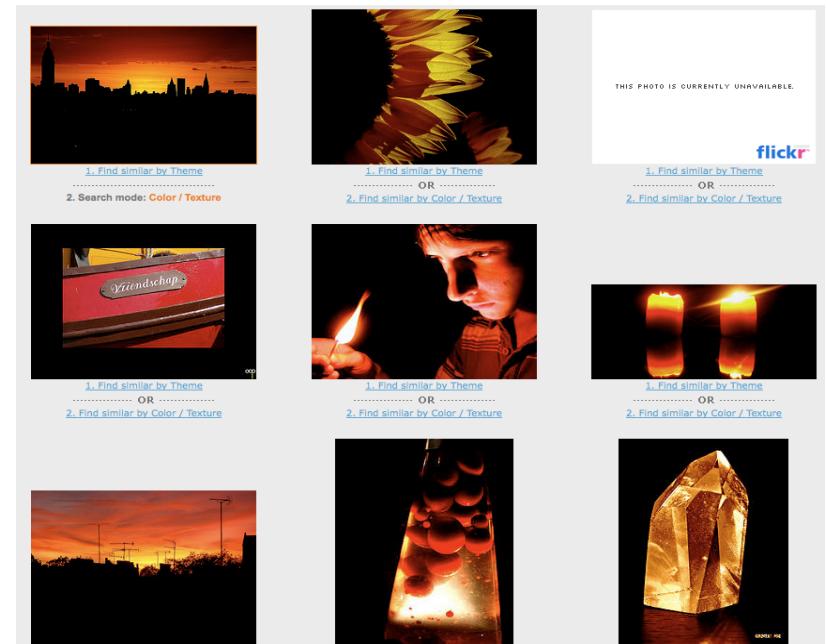
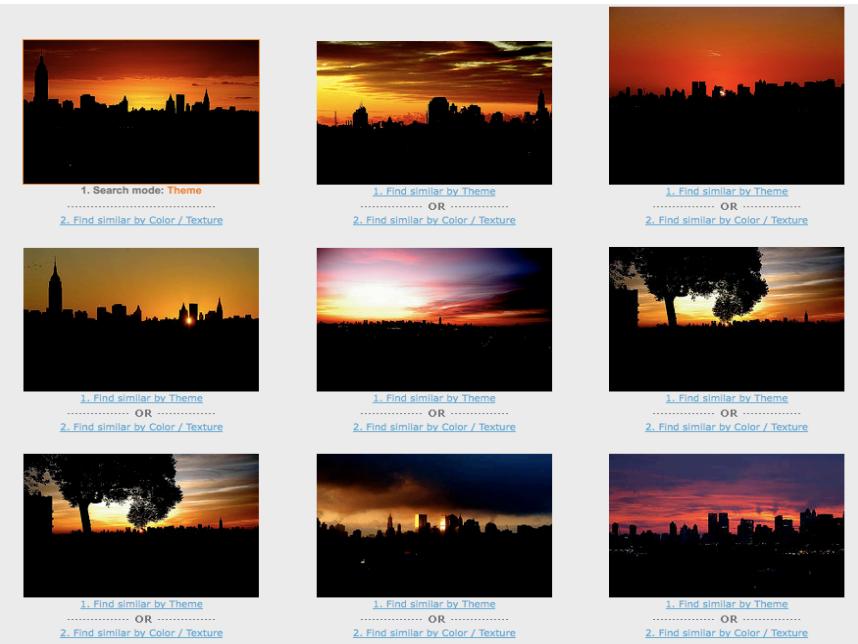
[\[PDF\] Large Scale Learning to Rank](#)
www.eecs.tufts.edu/~dsculley/papers/large-scale-rank.pdf
File Format: PDF/Adobe Acrobat - Quick View
by D Sculley - Cited by 24 - Related articles
Pairwise learning to rank methods such as RankSVM give good performance, ... In this paper, we are concerned with learning to rank methods that can learn on ...

[Microsoft Learning to Rank Datasets - Microsoft Research](#)
research.microsoft.com/en-us/projects/mslrl/
We release two large scale datasets for research on learning to rank: L2R-WEB30k with more than 30000 queries and a random sampling of it L2R-WEB10K ...

[LETOR: A Benchmark Collection for Research on Learning to Rank ...](#)
research.microsoft.com/~letor/
This website is designed to facilitate research in Learning TO Rank (LETOR). Much information about learning to rank can be found in the website, including ...

Due to Prof. David Sontag, NYU

Given image, find similar images



<http://www.tiltomo.com/>

Due to Prof. David Sontag, NYU

Slide 93

Collaborative Filtering

Due to Prof. David Sontag, NYU

Slide 94

Recommendation systems

David's Amazon.com | Today's Deals | Gift Cards | Sell | Help

Daily Lightning Deals
Back-to-School Savings
Shop now

Hello, David
Your Account | Try Prime | Cart | Wish List

Shop by Department | Search Books | Go | Your Amazon.com | Your Browsing History | Recommended For You | Amazon Betterizer | Improve Your Recommendations | Your Profile | Learn More

Your Amazon.com > Recommended for You > Books > Subjects > Science & Math > History & Philosophy

Just For Today | Browse Recommended

Recommendations | History & Philosophy | History of Science | Philosophy of Biology | Philosophy of Medicine

These recommendations are based on Items you own and more.

view: All | New Releases | Coming Soon

1. **Causality: Models, Reasoning and Inference**
by Judea Pearl (September 14, 2009)
Average Customer Review: ★★★★★ (10)
In Stock
List Price: \$60.00
Price: \$32.49
61 used & new from \$28.00
 I own it Not interested Rate this item
Recommended because you purchased Probabilistic Graphical Models and more (Fix this)

2. **The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century**
by David Salsburg (May 1, 2002)
Average Customer Review: ★★★★★ (26)
In Stock
List Price: \$18.99
Price: \$13.88
81 used & new from \$9.00
 I own it Not interested Rate this item
Recommended because you added The Theory That Would Not Die to your Wish List (Fix this)

3. **The Eighth Day of Creation: Makers of the Revolution in Biology, 25th Anniversary Edition**
by Horace Freeland Judson (November 1, 1996)
Average Customer Review: ★★★★★ (10)
In stock on September 4, 2013
List Price: \$66.00
Price: \$36.09
59 used & new from \$26.95
 I own it Not interested Rate this item
Recommended because you purchased Molecular Biology of the Cell (Fix this)

4. **The Machinery of Life**
by David S. Goodsell (April 28, 2009)
Average Customer Review: ★★★★★ (41)
In Stock
List Price: \$25.00
Price: \$17.49
92 used & new from \$12.00
 Add to Cart Add to Wish List

Recommendation systems

Machine learning competition with a \$1 million prize

Leaderboard

Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	The Ensemble	0.8553	10.10	2009-07-26 18:38:22
2	BellKor's Pragmatic Chaos	0.8554	10.09	2009-07-26 18:18:28
Grand Prize - RMSE <= 0.8563				
3	Grand Prize Team	0.8571	9.91	2009-07-24 13:07:49
4	Opera Solutions and Vandelay United	0.8573	9.89	2009-07-25 20:05:52
5	Vandelay Industries!	0.8579	9.83	2009-07-26 02:49:53
6	PragmaticTheory	0.8582	9.80	2009-07-12 15:09:53
7	BellKor in BigChaos	0.8590	9.71	2009-07-26 12:57:25
8	Dace	0.8603	9.58	2009-07-24 17:18:43
9	Opera Solutions	0.8611	9.49	2009-07-26 18:02:08
10	BellKor	0.8612	9.48	2009-07-26 17:19:11
11	BigChaos	0.8613	9.47	2009-06-23 23:06:52
12	Feeds2	0.8613	9.47	2009-07-24 20:06:46
Progress Prize 2008 - RMSE = 0.8616 - Winning Team: BellKor in BigChaos				
13	xiangliang	0.8633	9.26	2009-07-21 02:04:40
14	Gravity	0.8634	9.25	2009-07-26 15:58:34
15	Ces	0.8642	9.17	2009-07-25 17:42:38
16	Invisible Ideas	0.8644	9.14	2009-07-20 03:26:12
17	Just a guy in a garage	0.8650	9.08	2009-07-22 14:10:42
18	Craig Carmichael	0.8656	9.02	2009-07-25 16:00:54
19	J Dennis Su	0.8658	9.00	2009-03-11 09:41:54
20	acmehill	0.8659	8.99	2009-04-16 06:29:35
Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell				
Cinematch score on quiz subset - RMSE = 0.9514				



Due to Prof. David Sontag, NYU



Communities and Conferences

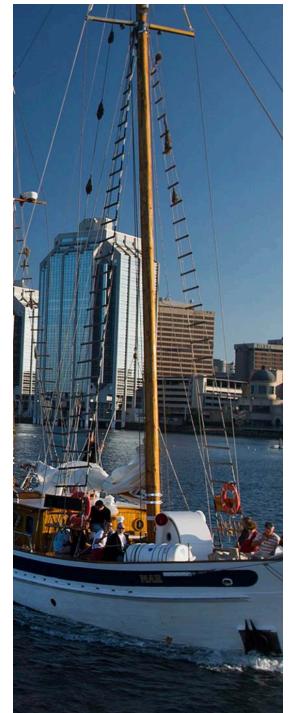
1. ACM [SIGKDD](#) (Know. Disc./DM)
2. ACM [SIGMOD](#) (Mgmt. of Data)
3. [ACM CIKM](#) (Info. Know. Mgmt.)
4. [IEEE ICDE](#) (Data Engineering)
5. [VLDB](#) (Very Large Data Bases)
6. ASA [SLDM](#) (Stat. Learning / Data Mining)
7. INFORMS [Data Mining Section](#)
8. [SIAM Data Mining](#)

KDD 2017

Halifax, Nova Scotia - Canada

August 13 - 17, 2017

KDD 2017 is a premier interdisciplinary conference bringing together researchers and practitioners from data science, data mining, knowledge discovery, large-scale data analytics, and big data.





4. Real World Challenges

Social Implications, Data Mining in Practice, Challenges



Social Implications of Data Mining

- data mining techniques are part of our everyday lives
- today we are routinely targeted with advertising
- data mining applications can derive information that was previously unknown
 - area code may correlate ethnicity
 - identify individuals using simple demographics
 - postal code, birthdate, gender



Social Implications of Data Mining

- data mining techniques are part of our everyday lives
- today we are routinely targeted with advertising
- data mining applications can derive information that was previously unknown
 - area code may correlate ethnicity
 - identify individuals using simple demographics
 - postal code, birthdate, gender
- data mining is used for fraud detection, identifying criminal suspects, predicting terrorists
 - aka profiling behaviors
- drawbacks
 - classification is imperfect
 - mistakes are made
- data mining techniques must be sensitive to privacy issues
 - i.e., not violate privacy directives



Data Mining: In Practice

- human interaction
 - requires cooperation between data mining and domain-area experts
- generated models
 - overfitting
 - models generated do not fit future states; describe noise more so than relationships
 - coping with outliers
 - including outliers may present challenges to models



Data Mining: In Practice

- human interaction
 - requires cooperation between data mining and domain-area experts
- generated models
 - overfitting
 - models generated do not fit future states; describe noise more so than relationships
 - coping with outliers
 - including outliers may present challenges to models
- interpretation of results
 - data mining output requires experts to interpret results
- visualization of results
 - helpful to communicate findings
- large datasets
 - sampling, parallelization helps
- high dimensionality
 - difficult to determine which attributes (dimensions) to use
 - apply dimensionality reduction methods



Data Mining: In Practice

- multimedia data
 - present challenges to data mining algorithms
- missing data
 - can lead to invalid results
- irrelevant data
 - some attributes may be irrelevant
- noisy data
 - attributes might be invalid or incorrect; need to be resolved



Data Mining: In Practice

- multimedia data
 - present challenges to data mining algorithms
- missing data
 - can lead to invalid results
- irrelevant data
 - some attributes may be irrelevant
- noisy data
 - attributes might be invalid or incorrect; need to be resolved
- changing data
 - datasets are not necessarily static; data mining must be rerun when there are updates
- integration
 - integration of data mining functions (e.g., into DBMS is desirable)
- disruptive
 - business practices may have to be modified to determine how to effectively use the information uncovered



5. Ten DM Challenges (Among Others)



C1. Curse of Dimensionality (COD)

Coined by R. Bellman (1961) in the context of control theory

- statistical inference increasingly more difficult as data becomes multivariate

In high dimensions ...

- nearly all data sets are sparse
 - as the dimension increases the distance between points also increases

“Statistical Data Mining,” David L. Banks, Wiley Interdisciplinary Reviews: Computational Statistics, 2, 9-25, doi: 10.1002/wics.53



C1. Curse of Dimensionality (COD)

Coined by R. Bellman (1961) in the context of control theory

- statistical inference increasingly more difficult as data becomes multivariate

In high dimensions ...

- nearly all data sets are sparse
 - as the dimension increases the distance between points also increases
- number of possible models grows combinatorially fast
 - many more possible models

“Statistical Data Mining,” David L. Banks, Wiley Interdisciplinary Reviews: Computational Statistics, 2, 9-25, doi: 10.1002/wics.53



C2. No Free Lunch (NFL) Theorem

A wide variety of techniques exist for modeling

- traditional statistical approaches, such as generalized linear models (GLMs) and linear discriminant analysis, to modern machine learning (ML) techniques, such as artificial neural networks (ANNs), support-vector machines (SVMs), Bayes (or belief) networks, and decision trees

Arising in Statistical Machine Learning the NFL Theorem states that ...

- no technique will outperform all other techniques on all problems



C2. No Free Lunch (NFL) Theorem

A wide variety of techniques exist for modeling

- traditional statistical approaches, such as generalized linear models (GLMs) and linear discriminant analysis, to modern machine learning (ML) techniques, such as artificial neural networks (ANNs), support-vector machines (SVMs), Bayes (or belief) networks, and decision trees

Arising in Statistical Machine Learning the NFL Theorem states that ...

- no technique will outperform all other techniques on all problems
- any modeling effort should consider a range of techniques



C3. Data Preparation

Critical modeling aspect is *data representation*

- transforms numbers/labels into a representation suitable for input to a model
- is the primary point where domain expertise can come into play

How to represent or encode an input variable in a model?

- number of calls per day a customer makes to an ISP



C3. Data Preparation

Critical modeling aspect is *data representation*

- transforms numbers/labels into a representation suitable for input to a model
- is the primary point where domain expertise can come into play

How to represent or encode an input variable in a model?

- number of calls per day a customer makes to an ISP
 - a **scalar** (e.g., customer makes 5 calls, then the no. is 5).
 - use **bins** (e.g., 0-1, 2-5, 6-12, 12+) expressed as a binary vector
 - if customer made 5 or 20 calls, this is [0 1 0 0] or [0 0 0 1], respectively
 - many possible representations, which **influences model accuracy**



C3. Data Preparation

Critical modeling aspect is *data representation*

- transforms numbers/labels into a representation suitable for input to a model
- is the primary point where domain expertise can come into play

How to represent or **encode** an input variable in a model?

- number of calls per day a customer makes to an ISP
 - a **scalar** (e.g., customer makes 5 calls, then the no. is 5).
 - use **bins** (e.g., 0-1, 2-5, 6-12, 12+) expressed as a binary vector
 - if customer made 5 or 20 calls, this is [0 1 0 0] or [0 0 0 1], respectively
 - many possible representations, which **influences model accuracy**

Poor data preparation likely leads to inferior models



C4. Model Selection

- There is a continuum of models from simple to complex
 - simple models capture the primary trends in a data set, but their simplicity may prevent them from capturing subtle patterns.
 - complex models capture subtle patterns, but they memorize training data quirks, instead of capturing patterns that will be useful for prediction.



C4. Model Selection

- There is a continuum of models from simple to complex
 - simple models **capture** the primary **trends** in a data set, but their simplicity may prevent them from capturing subtle patterns.
 - complex models capture **subtle patterns**, but they **memorize** training data quirks, **instead of capturing patterns** that will be useful for prediction.
- A key issue is *model selection*, picking the appropriate level of complexity for a model given data.
- Many methods exist for model selection
 - (e.g., cross validation, Bayesian averaging, ensembles, regularization)
- **Rigorous model selection is important, otherwise the models will be far less accurate than they could be**



C5. Model Evaluation: Test Data

Once a model is built, the question arises: “Is it accurate?”

Unfortunately, modelers can inflate model accuracy estimates

Failing to use an independent test set.

- A set of training examples are used to construct the model.
- It is meaningless to assess the accuracy of the model using the training set, because one could always build a model that has extremely high accuracy on the training set.
- To obtain a fair estimate of performance, the **model must be evaluated on examples that were not contained** in the **training set**.
- To obtain independent training and test sets, the available **data must be split into two nonoverlapping subsets**, with the test set reserved only for evaluation.



C6. Model Evaluation: Stationarity

Assuming stationarity of the test environment.

- For many difficult problems, a model built based on historical data will become a poorer and poorer predictor as time goes on, because the environment is *nonstationary* – the rules and behaviors of individuals change over time.



C7. Model Evaluation: Filtering

Filtering data to bias results.

- In a large data set, one segment of the population may be easier to predict than another
 - For example, customers with low incomes are likely to be more cost sensitive, and hence, might reliably churn (i.e., customer turnover) when their bills rise above a certain amount.



C7. Model Evaluation: Filtering

Filtering data to bias results.

- In a large data set, one segment of the population may be easier to predict than another
 - For example, customers with low incomes are likely to be more cost sensitive, and hence, might reliably churn (i.e., customer turnover) when their bills rise above a certain amount.
- If a model is trained and tested just on this segment of the population, it will be more accurate than a model that must handle the entire population.
- Selective filtering turns a hard problem into an easier problem.
 - For example, focusing on customers in the first three months of service, and assuming that performance on the broader customer base is comparable.



C8. Acquiring Data

- Although a great deal is said about the availability of data
- The truth is it isn't always easy to gain access to it for varying reasons
 - corporate confidentiality agreements (re: 3rd party data)
 - user privacy rights
 - conflicts of interest, such as in national security (in the defense domain)
 - global competitiveness
 - available for purchase only

The KONECT (Koblenz Network Collection) is a project to collect large network datasets of all types in order to perform research in network science and related fields, collected by the Institute of Web Science and Technologies at the University of Koblenz-Landau. KONECT contains over a hundred network datasets of various types, including directed, undirected, bipartite, weighted, unweighted, signed and rating networks. The networks of KONECT are collected from many diverse areas such as social networks, hyperlink networks, authorship networks, physical networks, interaction networks and communication networks. The KONECT project has developed network analysis tools which are used to compute network statistics, to draw plots and to implement various link prediction algorithms. The result of these analyses are presented on these pages. Whenever we are allowed to do so, we provide a download of the networks.

Stanford Large Network Dataset Collection

- Social networks : online social networks, edges represent interactions between people
- Networks with ground-truth communities : ground-truth network communities in social and information networks
- Communication networks : email communication networks with edges representing communication
- Citation networks : nodes represent papers, edges represent citations
- Collaboration networks : nodes represent scientists, edges represent collaborations (co-authoring a paper)
- Web graphs : nodes represent webpages and edges are hyperlinks
- Amazon networks : nodes represent products and edges link commonly co-purchased products
- Internet networks : nodes represent computers and edges communication
- Road networks : nodes represent intersections and edges roads connecting the intersections
- Autonomous systems : graphs of the Internet
- Signed networks : networks with positive and negative edges (friend/foe, trust/distrust)
- Location-based online social networks : Social networks with geographic check-ins
- Wikipedia networks and metadata : Talk, editing and voting data from Wikipedia
- Twitter and Memetracker : Memetracker phrases, links and 467 million Tweets
- Online communities : Data from online communities such as Reddit and Flickr
- Online reviews : Data from online review systems such as BeerAdvocate and Amazon
- Information cascades : ...



TU-Berlin-DIMA / myriad-toolkit

Myriad Data Generator Toolkit



C9. Constraints & Discrepancies

Corporate/Operational Constraints

- managing time pressures/project schedule
- differences of opinion in the interpretation of results
- mitigation strategy
 - show consistency across many methods

Technological

- trustworthiness of the computing platform and the analytics
- incorrect usage of the tools
- mitigation strategy
 - “calibrate the instrument,” conduct experiments on known problems