CEFET/RJ

Data Preprocessing

Eduardo Ogasawara
eduardo.ogasawara@cefet-rj.br
https://eic.cefet-rj.br/~eogasawara

Data Quality: Why Preprocess the Data?

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.
[2] S. García, J. Luengo, and F. Herrera, 2014, Data Preprocessing in Data Mining. Springer.

Major Tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data reduction
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- Data transformation
  - Discretization
  - Normalization
  - Concept hierarchy generation

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.
[2] S. García, J. Luengo, and F. Herrera, 2014, Data Preprocessing in Data Mining. Springer.

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., Occupation = " " (missing data)
  - noisy: containing noise, errors, or outliers
    - e.g., Salary = "−10" (an error)
  - inconsistent: containing discrepancies in codes or names, e.g.,
    - Age = "42", Birthday = "03/07/2010"
    - Was rating "1, 2, 3", now rating "A, B, C"
    - discrepancy between duplicate records
  - Intentional (e.g., disguised missing data)
    - Jan. 1 as everyone's birthday?

[1] X. Chu, I.F. Ilyas, S. Krishnan, and J. Wang, 2016, Data cleaning: Overview and emerging challenges, In: Proceedings of the ACM SIGMOD International Conference on Management of Data, p. 2201–2206

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

[1] R.J.A. Little, 1988, A test of missing completely at random for multivariate data with missing values, Journal of the American Statistical Association, v. 83, n. 404, p. 1198–1202.
[2] B. Efron, 1994, Missing data, imputation, and the bootstrap, Journal of the American Statistical Association, v. 89, n. 426, p. 463–475.

How to Handle Missing Data?

- Ignore the tuple:
  - usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

[1] R.R. Andridge and R.J.A. Little, 2010, A review of hot deck imputation for survey non-response, International Statistical Review, v. 78, n. 1, p. 40–64.
[2] M.G. Kenward and J. Carpenter, 2007, Multiple imputation: Current perspectives, Statistical Methods in Medical Research, v. 16, n. 3, p. 199–218.

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which require data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

[1] X. Zhu and X. Wu, 2004, Class Noise vs. Attribute Noise: A Quantitative Study, Artificial Intelligence Review, v. 22, n. 3 (Nov.), p. 177–210.

How to Handle Noisy Data?

- Smoothing / Discretization
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Removing outliers
  - Box-plot based
  - Clustering: detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

[1] H. Liu, F. Hussain, C.L. Tan, and M. Dash, 2002, Discretization: An enabling technique, Data Mining and Knowledge Discovery, v. 6, n. 4, p. 393–423.
[2] M. Krzywinski and N. Altman, 2015, Points of Significance: Multiple linear regression, Nature Methods, v. 12, n. 12, p. 1103–1104.

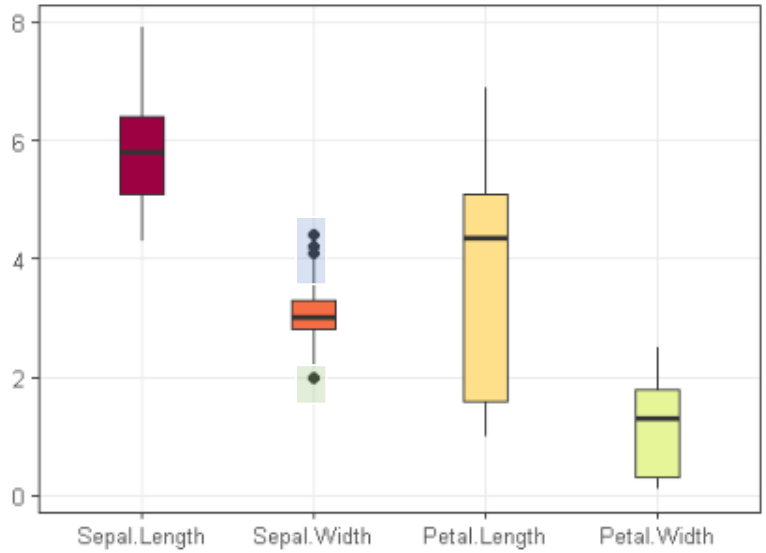
Data Cleaning as a Process

- Data discrepancy detection
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
- Data migration and integration
  - Data migration tools: enable transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface

[1] X. Chu, I.F. Ilyas, S. Krishnan, and J. Wang, 2016, Data cleaning: Overview and emerging challenges, In: Proceedings of the ACM SIGMOD International Conference on Management of Data, p. 2201–2206
[2] X. Wang and C. Wang, 2020, Time Series Data Cleaning: A Survey, IEEE Access, v. 8, p. 1866–1881.

Outlier removal based on boxplot

- Interval for regular data [Q_1-1.5·IQR, Q_3+1.5·IQR]
  - More conservative interval [Q_1-3·IQR, Q_3+3·IQR]



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
16	5.7	4.4	1.5	0.4	setosa
33	5.2	4.1	1.5	0.1	setosa
34	5.5	4.2	1.4	0.2	setosa
61	5.0	2.0	3.5	1.0	versicolor

[1] R. McGill, J.W. Tukey, and W.A. Larsen, 1978, Variations of box plots, American Statistician, v. 32, n. 1, p. 12–16.

Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real-world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

[1] B. Golshan, A. Halevy, G. Mihaila, and W.-C. Tan, 2017, Data integration: After the teenage years, In: Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, p. 101–106

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - Object identification: The same attribute or object may have different names in different databases
  - Derivable data: One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by correlation analysis and covariance analysis
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

[1] S. Bhattacherjee, A. Narang, and V.K. Garg, 2011, High throughput data redundancy removal algorithm with scalable performance, In: HiPEAC’11 - Proceedings of the 6th International Conference on High Performance and Embedded Architectures and Compilers, p. 87–96

Correlation Analysis (Nominal Data)

- χ^2 (chi-square) test
- $\chi^2 = \sum \frac{(observed - expected)^2}{expected}$
- The larger the χ^2 value, the more likely the variables are related
- The cells that contribute the most to the χ^2 value are those whose actual count is very different from the expected count
- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

[1] R.J. Larsen e M.L. Marx, 2017, An Introduction to Mathematical Statistics and Its Applications. Pearson Education.

Chi-Square Calculation: An Example

- χ^2(chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)
  - $\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$
- It shows that like science fiction and play chess are correlated in the group

$$90 = 1500 \cdot \frac{300}{1500} \cdot \frac{450}{1500}$$

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

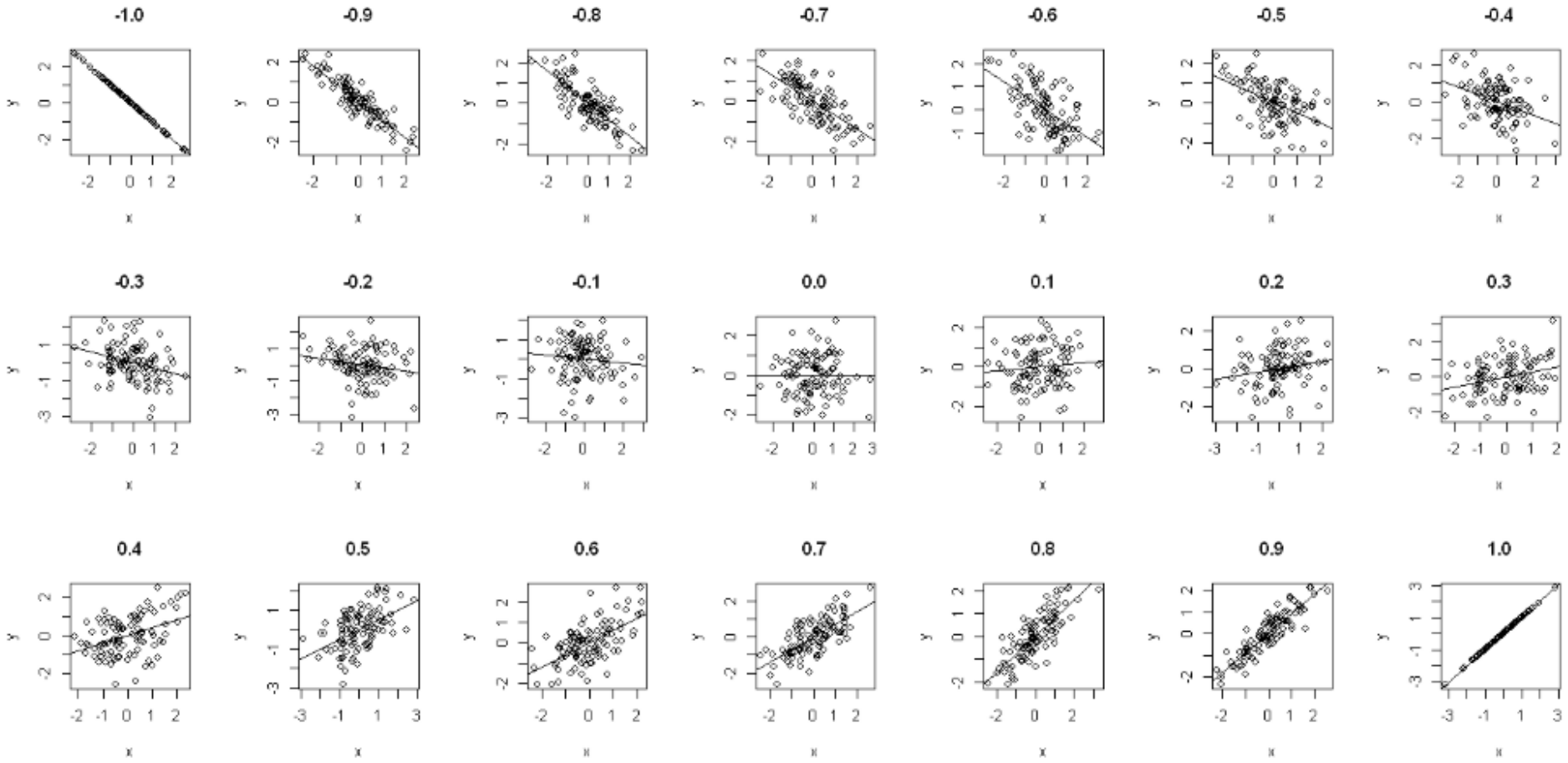
[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

Correlation Analysis (Numeric Data)

- Correlation coefficient (Pearson's product moment coefficient)
  - $r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A \sigma_B}$
  - where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\sum_{i=1}^{n}(a_i b_i)$ is the sum of the AB cross-product
- If $r_{A,B} > 0$, A and B are positively correlated
  - (A's values increase as B's). The higher, the stronger correlation
- $r_{A,B} = 0$: independent;
- $r_{A,B} < 0$: negatively correlated

[1] R.J. Larsen and M.L. Marx, 2017, An Introduction to Mathematical Statistics and Its Applications. Pearson Education.

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1

Data Reduction Strategies

- Data reduction: Obtain a reduced representation of the data set that is much smaller in volume but produces the same (or almost the same) analytical results
- Why data reduction?
  - A database/data warehouse may store terabytes of data
  - Complex data analysis may take a very long time to run on the complete data set
- Data reduction strategies
  - Dimensionality reduction, e.g., remove unimportant attributes
    - Feature subset selection, feature creation
    - Principal Components Analysis (PCA)
    - Regression Models
  - Numerosity reduction (some simply call it: Data Reduction)
    - Sampling
    - Aggregation
    - Data compression

[1] M.H. ur Rehman, C.S. Liew, A. Abbas, P.P. Jayaraman, T.Y. Wah, e S.U. Khan, 2016, Big Data Reduction Methods: A Survey, Data Science and Engineering, v. 1, n. 4, p. 265–284.

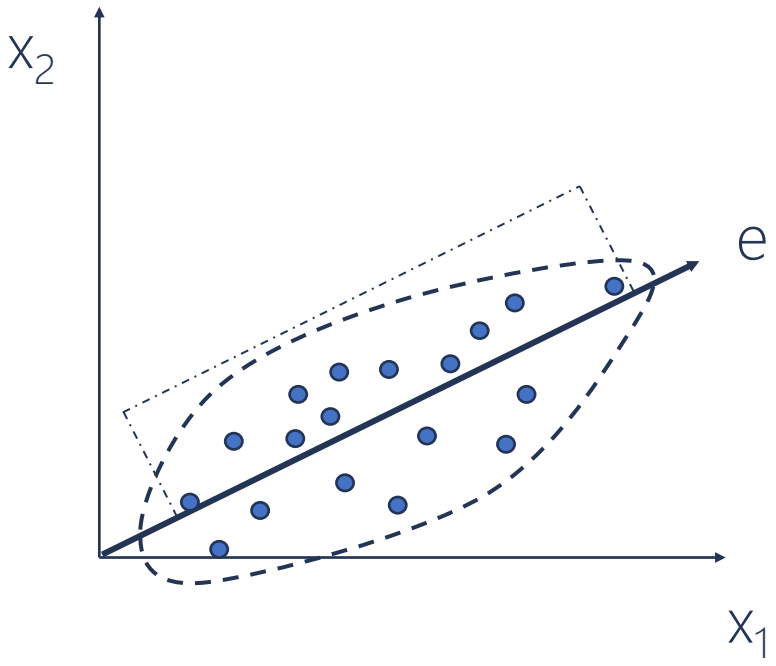
Dimensionality Reduction

- Curse of dimensionality
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points becomes less meaningful
    - critical to clustering, outlier analysis
  - The possible combinations of subspaces will grow exponentially
- Dimensionality reduction
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization
- Dimensionality reduction techniques
  - Principal Component Analysis
  - Supervised and nonlinear techniques (e.g., feature selection)
  - Feature creation
  - Regression Models

[1] M. Dash, H. Liu, e J. Yao, 1997, Dimensionality reduction of unsupervised data, In: Proceedings of the International Conference on Tools with Artificial Intelligence, p. 532–539

Principal Component Analysis (PCA)

- Find a projection that captures the largest amount of variation in data
- The original data are projected onto a much smaller space, resulting in dimensionality reduction. We find the eigenvectors of the covariance matrix, and these eigenvectors define the new space



[1] H. Abdi and L.J. Williams, 2010, Principal component analysis, Wiley Interdisciplinary Reviews: Computational Statistics, v. 2, n. 4, p. 433–459.

Principal Component Analysis (Steps)

- Given n data vectors from n-dimensions, find k ≤ n orthogonal vectors (principal components) that can be best used to represent data
  - Normalize input data: Each attribute falls within the same range
  - Compute k orthonormal (unit) vectors, i.e., principal components
  - Each input data (vector) is a linear combination of the k principal component vectors
  - The principal components are sorted in order of decreasing "significance" or strength
  - Since the components are sorted, the size of the data can be reduced by eliminating the weak components, i.e., those with low variance (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data)
- Works only for numeric data only

[1] H. Abdi and L.J. Williams, 2010, Principal component analysis, Wiley Interdisciplinary Reviews: Computational Statistics, v. 2, n. 4, p. 433–459.

Projection with PCA

Data

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5.0	3.6	1.4	0.2
5.4	3.9	1.7	0.4

X

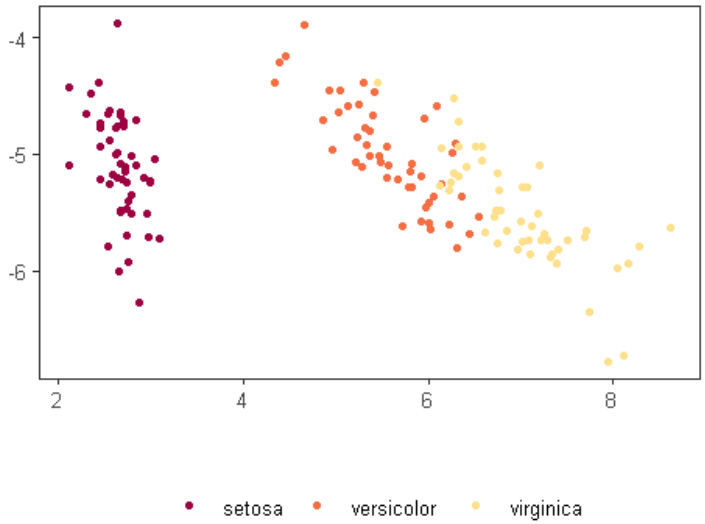
PCA

	PC1	PC2
Sepal.Length	0.5210659	-0.37741762
Sepal.Width	-0.2693474	-0.92329566
Petal.Length	0.5804131	-0.02449161
Petal.Width	0.5648565	-0.06694199

=

P

PC1	PC2	Species
2.640270	-5.204041	setosa
2.670730	-4.666910	setosa
2.454606	-4.773636	setosa
2.545517	-4.648463	setosa
2.561228	-5.258629	setosa
2.975946	-5.707321	setosa



[1] A. Tharwat, T. Gaber, A. Ibrahim, and A.E. Hassanien, 2017, Linear discriminant analysis: A detailed tutorial, AI Communications, v. 30, n. 2, p. 169–190.
https://www.r-bloggers.com/2014/01/computing-and-visualizing-lda-in-r/

Attribute Subset Selection

- Another way to reduce dimensionality of data
- Redundant attributes
  - Duplicate much or all the information contained in one or more other attributes
  - E.g., purchase price of a product and the amount of sales tax paid
- Irrelevant attributes
  - Contain no information that is useful for the data mining task at hand
  - E.g., students' ID is often irrelevant to the task of predicting students' GPA

[1] L. Carlos Molina, L. Belanche, and À. Nebot, 2002, Feature selection algorithms: A survey and experimental evaluation, In: Proceedings - IEEE International Conference on Data Mining, ICDM, p. 306–313
[2] M.A. Hall and G. Holmes, 2003, Benchmarking Attribute Selection Techniques for Discrete Class Data Mining, IEEE Transactions on Knowledge and Data Engineering, v. 15, n. 6, p. 1437–1447.

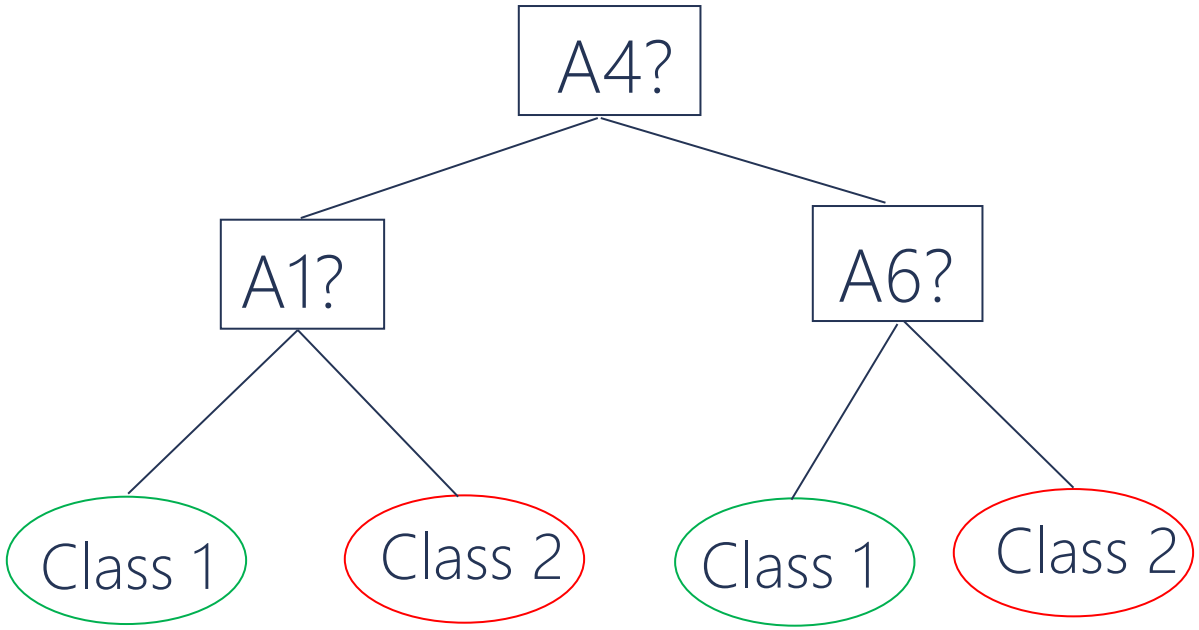
Heuristic Search in Attribute Selection

- There are 2^d possible attribute combinations of d attributes
- Typical heuristic attribute selection methods:
  - Best single attribute under the attribute independence assumption: choose by significance tests
  - Best step-wise feature selection:
    - The best single attribute is picked first
    - Then next best attribute condition to the first, ..
  - Step-wise attribute elimination:
    - Repeatedly eliminate the worst attribute

[1] I. Kononenko and S.J. Hong, 1997, Attribute selection for modelling, Future Generation Computer Systems, v. 13, n. 2–3, p. 181–195.
[2] M. Dash and H. Liu, 1997, Feature selection for classification, Intelligent Data Analysis, v. 1, n. 3, p. 131–156.

Example of Decision Tree induction for feature selection

Initial attribute set:
{A1, A2, A3, A4, A5, A6}



⇨ Reduced attribute set: {A1, A4, A6}

[1] J.R. Quinlan, 1996, Bagging, boosting, and C4.5, In: Proceedings of the National Conference on Artificial Intelligence, p. 725–730

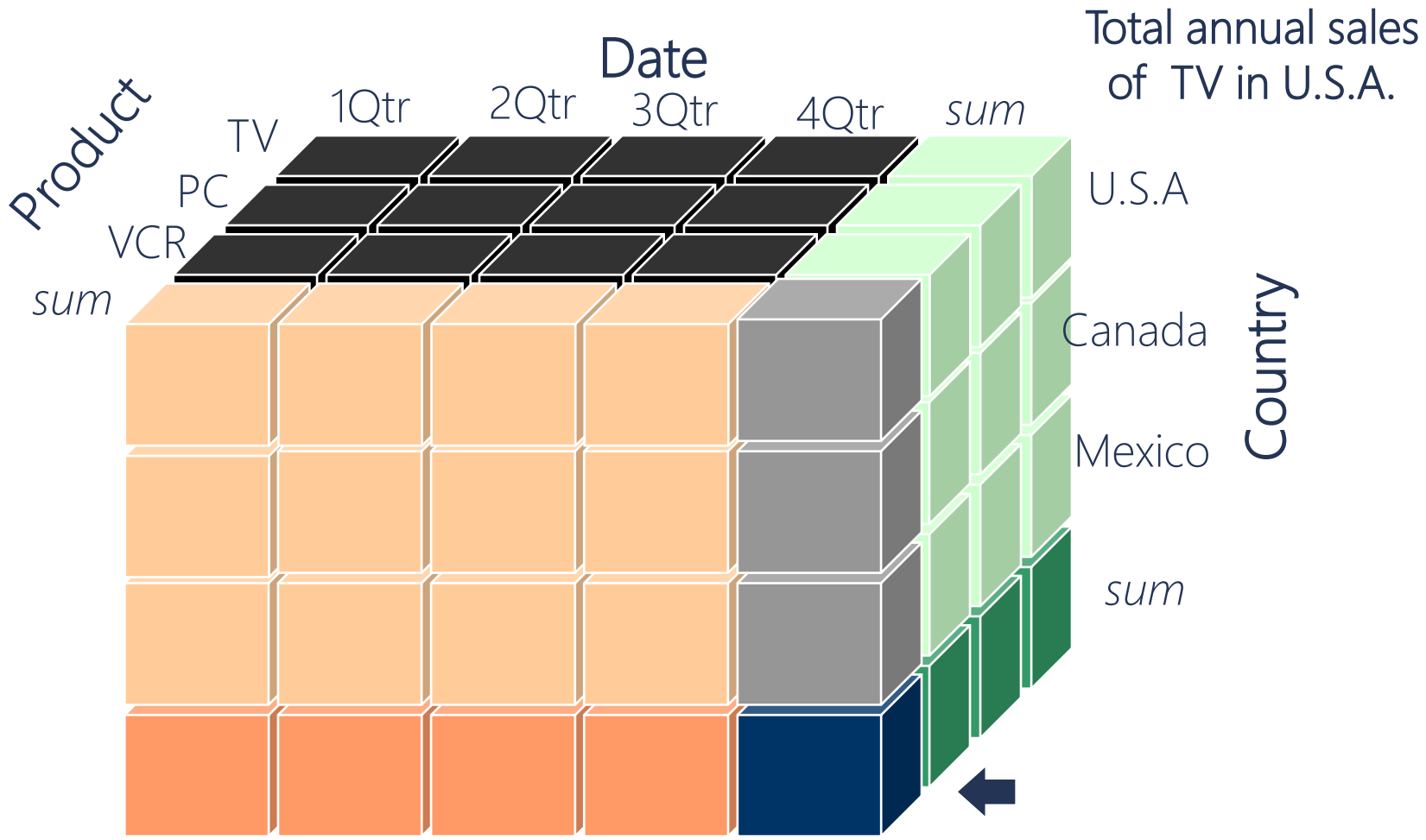
Feature Generation

- Create new attributes (features) that can capture the important information in a data set more effectively than the original ones
- Three general methodologies
  - Attribute extraction
    - Domain-specific
  - Mapping data to new space (see data reduction)
    - E.g., Fourier transformation, wavelet transformation
  - Attribute construction
    - Data discretization

[1] S. Markovitch and D. Rosenstein, 2002, Feature generation using general constructor functions, Machine Learning, v. 49, n. 1, p. 59–98.
[2] I. Daubechies, 1990, The Wavelet Transform, Time-Frequency Localization and Signal Analysis, IEEE Transactions on Information Theory, v. 36, n. 5, p. 961–1005.

Data Aggregation

- Aggregation: Summarization, data cube construction



[1] V. Harinarayan, A. Rajaraman, and J.D. Ullman, 1996, Implementing Data Cubes Efficiently, SIGMOD Record (ACM Special Interest Group on Management of Data), v. 25, n. 2, p. 205–216.

Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values
  - each old value can be identified with one of the new values
- Methods
  - Attribute/feature construction
    - New attributes constructed from the given ones
    - Complex aggregation
  - Normalization: Scaled to fall within a smaller, specified range
  - Discretization / Smoothing
  - Concept hierarchy climbing
  - Categorical Mapping

[1] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, and F. Herrera, 2017, A survey on data preprocessing for data stream mining: Current status and future directions, Neurocomputing, v. 239, p. 39–57.

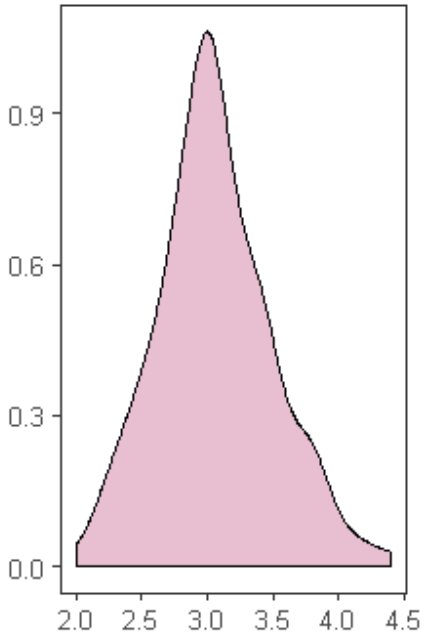
Normalization

- Min-max normalization: to [nminA, nmaxA]
  - $nv = \frac{v - min_A}{max_A - min_A}(nmax_A - nmin_A) + nmin_A$
- Z-score normalization (μ: mean, σ: standard deviation):
  - $nv = \frac{v - \mu_A}{\sigma_A}$
- Normalization by decimal scaling
  - $nv = \frac{v}{10^j}$, where j is the smallest integer such that max(|nv|) < 1
- Let income range ($12,000,$98,000) with μ = 54,000, σ = 16,000, then $73,600
  - is mapped to $\frac{73600-12000}{98000-12000}(1-0)+0 = 0.716$ using min-max (0-1)
  - is mapped to $\frac{73600-54000}{16000} = 1.225$ using z-score
  - Is mapped to $\frac{v}{10^6} = 0.736$ using decimal scaling

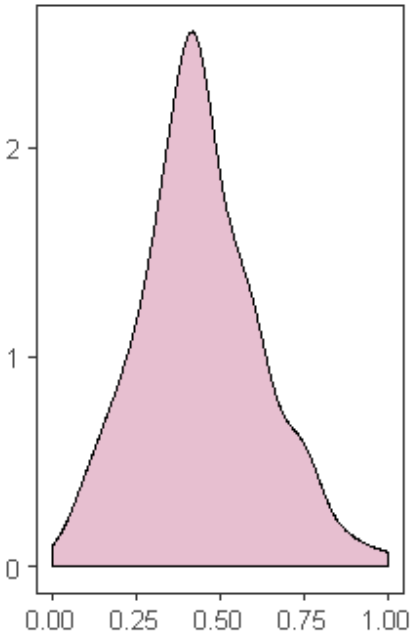
[1] E. Ogasawara, L. Murta, G. Zimbrão, and M. Mattoso, 2009, Neural networks cartridges for data mining on time series, In: Proceedings of the International Joint Conference on Neural Networks, p. 2302–2309

Normalization

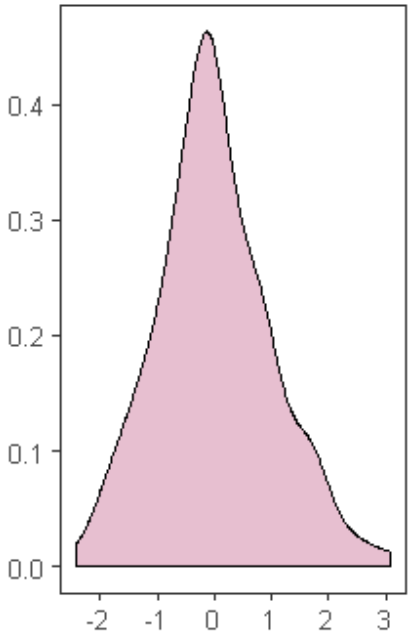
Data



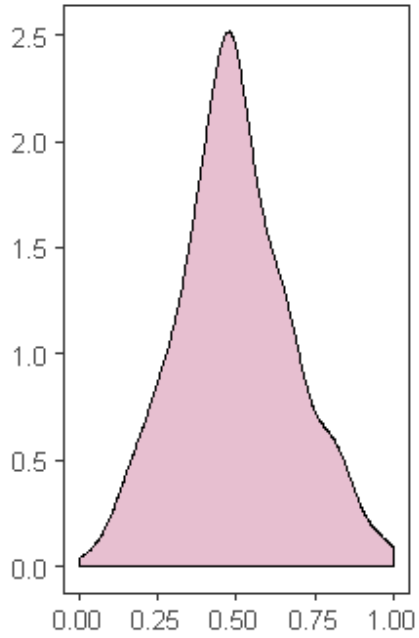
Min-max [0-1]



Z-score/N(0,1)



N(0.5, $\sqrt{\frac{0.5}{2.698}}$)



Discretization & Smoothing

- Discretization is the process of transferring continuous functions, models, variables, and equations into discrete counterparts
- Smoothing is a technique that creates an approximating function that attempts to capture important patterns in the data while leaving out noise or other fine-scale structures/rapid phenomena
- An important part of the discretization/smoothing is to set up bins for proceeding the approximation

[1] R. Kerber, 1992, Chimerge: discretization of numeric attributes, In: Proceedings Tenth National Conference on Artificial Intelligence, p. 123–128
[2] H. Liu, F. Hussain, C.L. Tan, and M. Dash, 2002, Discretization: An enabling technique, Data Mining and Knowledge Discovery, v. 6, n. 4, p. 393–423.

Binning methods for data smoothing

- Equal-width (distance) partitioning
  - Divides the range into N intervals of equal size: uniform grid
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B –A)/N
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well
- Equal-depth (frequency) partitioning
  - Divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

[1] M. Krzywinski, 2016, Points of View: Binning high-resolution data, Nature Methods, v. 13, n. 6, p. 463.

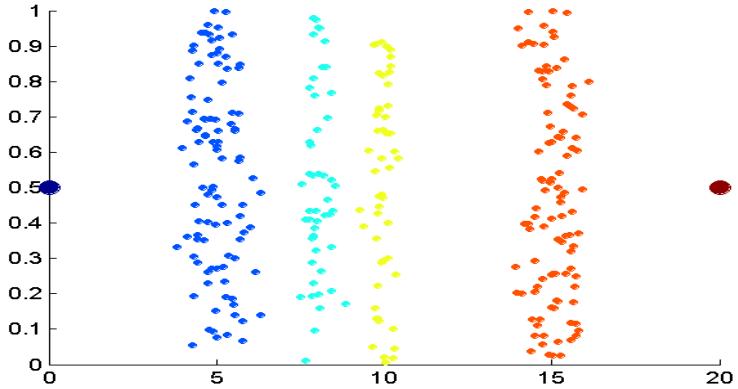
Binning methods for data smoothing

- Sorted data for price (in dollars):
  - 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Comparison of binning methods
  - Partition of equal-length (3 bins): (34-4)/3
    - Bin 1 [4-13[: 4, 8, 9
    - Bin 2 [14-23[: 15, 21, 21
    - Bin 3 [23-34]: 24, 25, 26, 28, 29, 34
  - Partition into equal-frequency (equi-depth) (3 bins):
    - Bin 1: 4, 8, 9, 15
    - Bin 2: 21, 21, 24, 25
    - Bin 3: 26, 28, 29, 34
- Smoothing using equal-frequency
  - by bin means:
    - Bin 1: 9, 9, 9, 9
    - Bin 2: 23, 23, 23, 23
    - Bin 3: 29, 29, 29, 29
  - by bin boundaries:
    - Bin 1: 4, 4, 4, 15
    - Bin 2: 21, 21, 25, 25
    - Bin 3: 26, 26, 26, 34

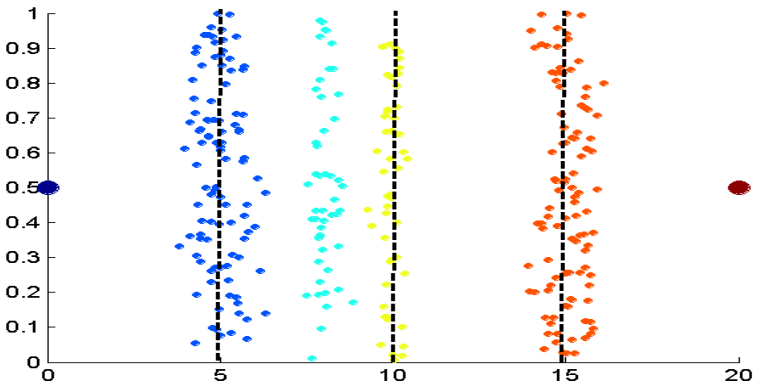
[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

Influence on binning during discretization techniques

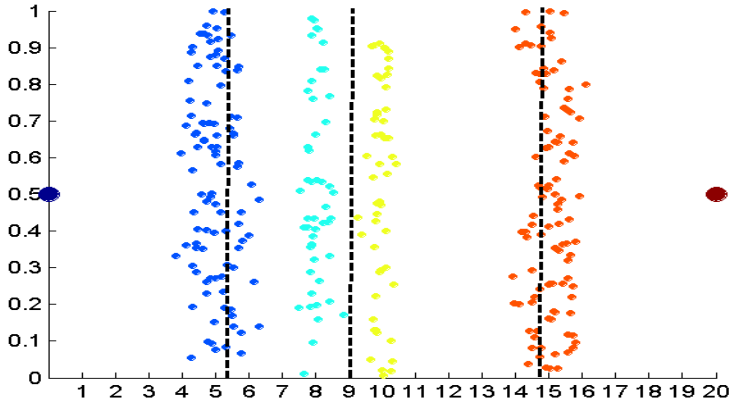
Smooth using 4 bins



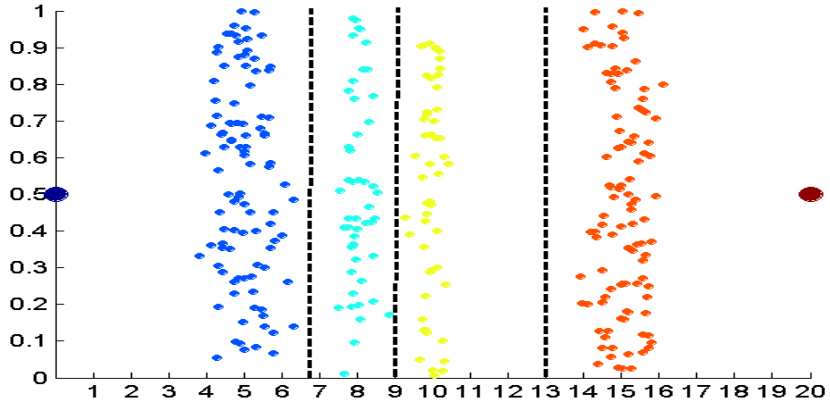
data



Equal interval width (binning)



Equal frequency (binning)



K-means clustering

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

Concept Hierarchy Generation

- Concept hierarchy organizes concepts hierarchically
  - It is usually associated with each dimension in a data warehouse
- Concept hierarchies facilitate drilling and rolling in data warehouses to view data in multiple granularity
- Concept hierarchies can be specified by domain experts
  - Domain semantics
- Concept hierarchy can be automatically formed
  - Numeric data: using discretization methods shown
  - Less semantics

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

Examples

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
  - city < state < country
  - week < month < year
- Specification of a hierarchy for a set of values by explicit data grouping
  - {city, state, country}

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

Categorical Mapping

- n binary derived inputs: one for each value of the original attribute
  - This 1-to-n mapping is commonly applied when n is relatively small
- As n grows, the number of inputs to the model increases and consequently the number of parameters to be estimated increases
  - Thus, this method is not applicable to high-cardinality attributes with hundreds or thousands of distinct values

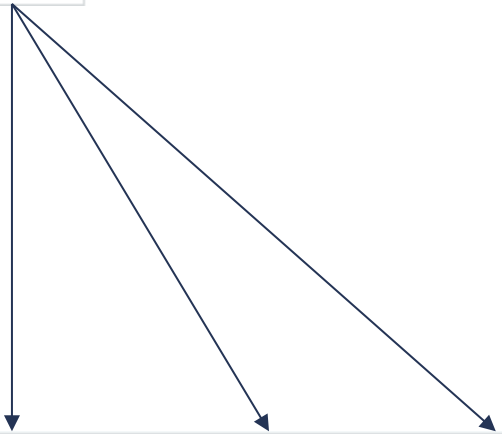
[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

Example Categorical Mapping

iris

Filter

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Speciessetosa	Speciesversicolor	Speciesvirginica
1	5.1	3.5	1.4	0.2	1	0	0
2	4.9	3.0	1.4	0.2	1	0	0
3	4.7	3.2	1.3	0.2	1	0	0
4	4.6	3.1	1.5	0.2	1	0	0
5	5.0	3.6	1.4	0.2	1	0	0

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

Sampling

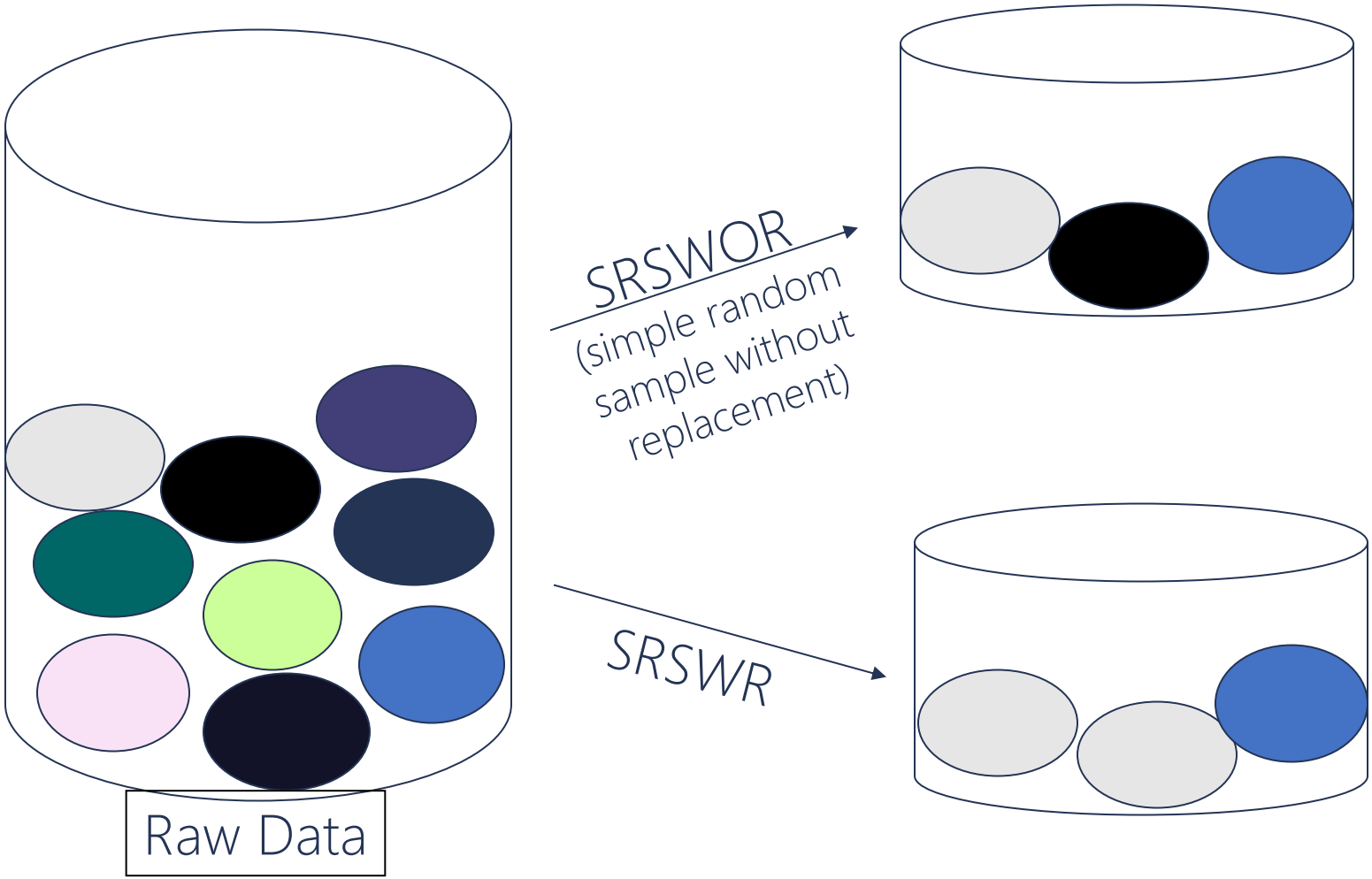
- Sampling: obtaining a small sample to represent the entire dataset
- Enable a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Key principle: Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
  - Develop adaptive sampling methods, e.g., stratified sampling
- Note: Sampling may not reduce database IOs (page at a time)

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

Types of Sampling

- Simple random sampling
  - There is an equal probability of selecting any item
- Sampling without replacement
  - Once an object is selected, it is removed from the population
- Sampling with replacement
  - A selected object is not removed from the population
- Stratified sampling
  - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
  - Used in conjunction with skewed data

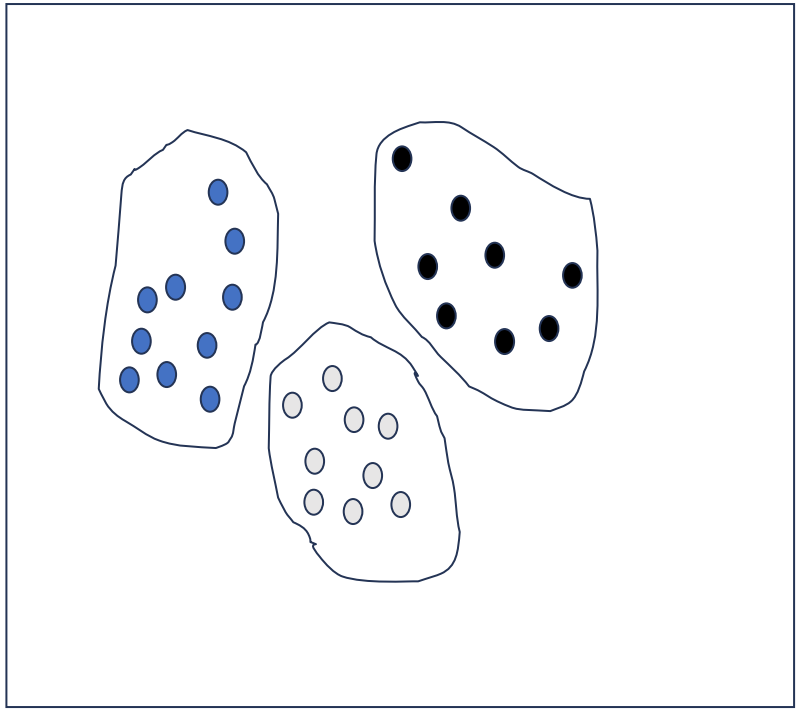
Sampling: With or without Replacement



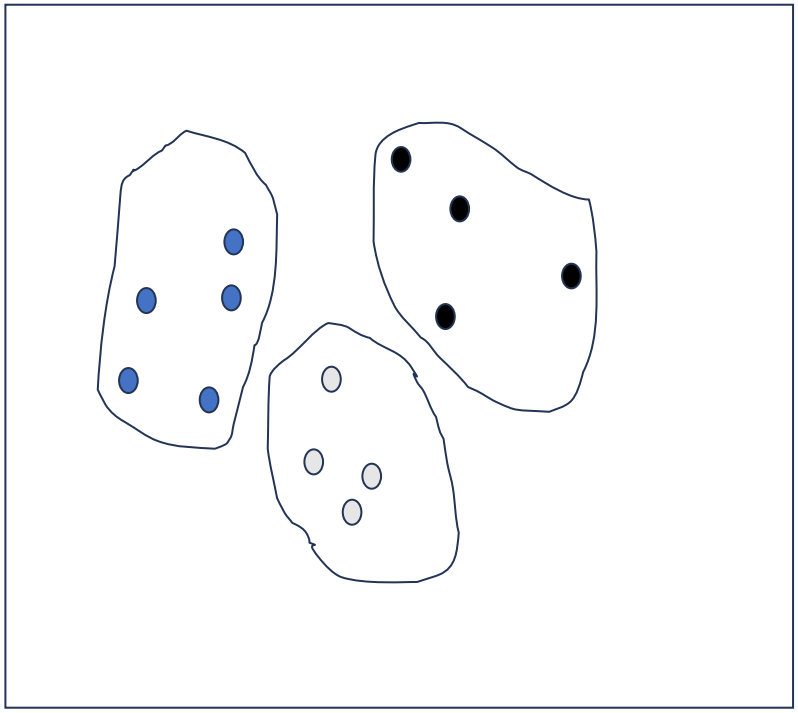
[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample



[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

Sampling - Examples

2/3

	setosa	versicolor	virginica
dataset	50	50	50
random sample	42	41	37
stratified sample	40	40	40

1/3

	setosa	versicolor	virginica
random sample	8	11	11
stratified sample	10	10	10

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
34	5.5	4.2	1.4	0.2	setosa
107	4.9	2.5	4.5	1.7	virginica
76	6.6	3.0	4.4	1.4	versicolor
22	5.1	3.7	1.5	0.4	setosa
116	6.4	3.2	5.3	2.3	virginica
113	6.8	3.0	5.5	2.1	virginica

[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

Balancing dataset

- Class-imbalance problem:
  - Rare positive example but numerous negative ones, e.g., medical diagnosis, fraud, oil-spill, fault
- Traditional methods assume a balanced distribution of classes and equal error costs
  - not suitable for class-imbalanced data
- Typical preprocessing methods for imbalance data in 2-class classification
  - Oversampling: re-sampling of data from positive class
  - Under-sampling: randomly eliminate tuples from negative class

[1] Z.-H. Zhou and X.-Y. Liu, 2006, Training cost-sensitive neural networks with methods addressing the class imbalance problem, IEEE Transactions on Knowledge and Data Engineering, v. 18, n. 1, p. 63–77.
[2] A. Tharwat and W. Schenck, 2020, Balancing Exploration and Exploitation: A novel active learner for imbalanced data, Knowledge-Based Systems, v. 210

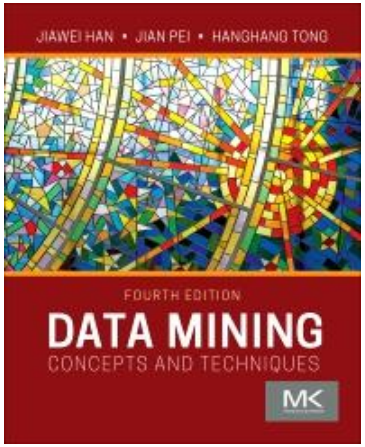
Oversampling/Under sampling

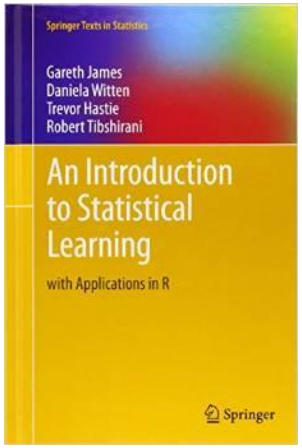
- Consider that the iris dataset had:
  - 20 observations of setosa
  - 50 observations of versicolor
  - 11 observations of virginica
- How does oversampling and subsampling address it?

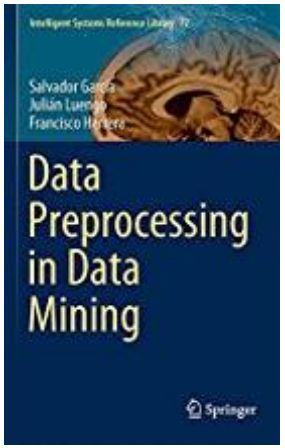
	setosa	versicolor	virginica
unbalanced	20	50	11
oversampling	50	50	50
subsampling	11	11	11

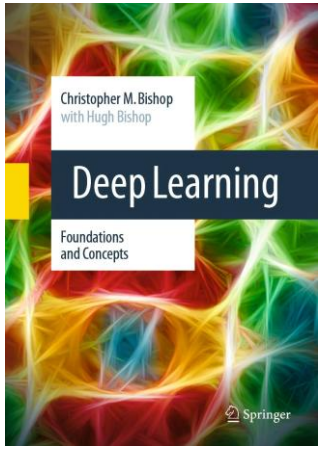
[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

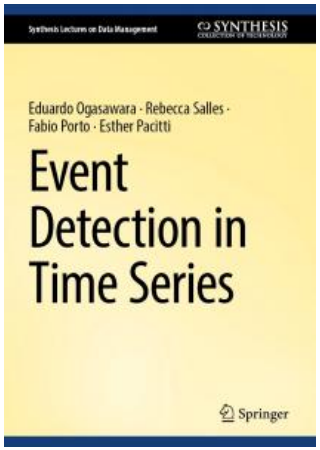
Main References











[1] J. Han, J. Pei, and H. Tong, Data Mining: Concepts and Techniques, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.

[2] G. M. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: With Applications in R. Springer Nature, 2021.

[3] S. Garcia, J. Luengo, and F. Herrera, Data Preprocessing in Data Mining. Springer, 2014.

[4] C. M. Bishop and H. Bishop, Deep Learning: Foundations and Concepts. Springer Nature, 2023.

[5] E. Ogasawara, R. Salles, F. Porto, and E. Pacitti, Event Detection in Time Series, 1st ed. in Synthesis Lectures on Data Management. Cham: Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-75941-3.

Slides and videos at: https://eic.cefet-rj.br/~eogasawara/data-mining/