Data Preprocessing

Cam Tu Nguyen

阮锦绣

Software Institute, Nanjing University
nguyenct@lamda.nju.edu.cn
ncamtu@gmail.com

Outline

- Data Preprocessing: Overview
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization

Outline

- Data Preprocessing: Overview
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization

Data Preprocessing: Overview

- Why Process Data?
  - To obtain Data Quality
    - Accuracy
    - Completeness
    - Consistency
    - Timeliness
    - Believability
    - Interpretability

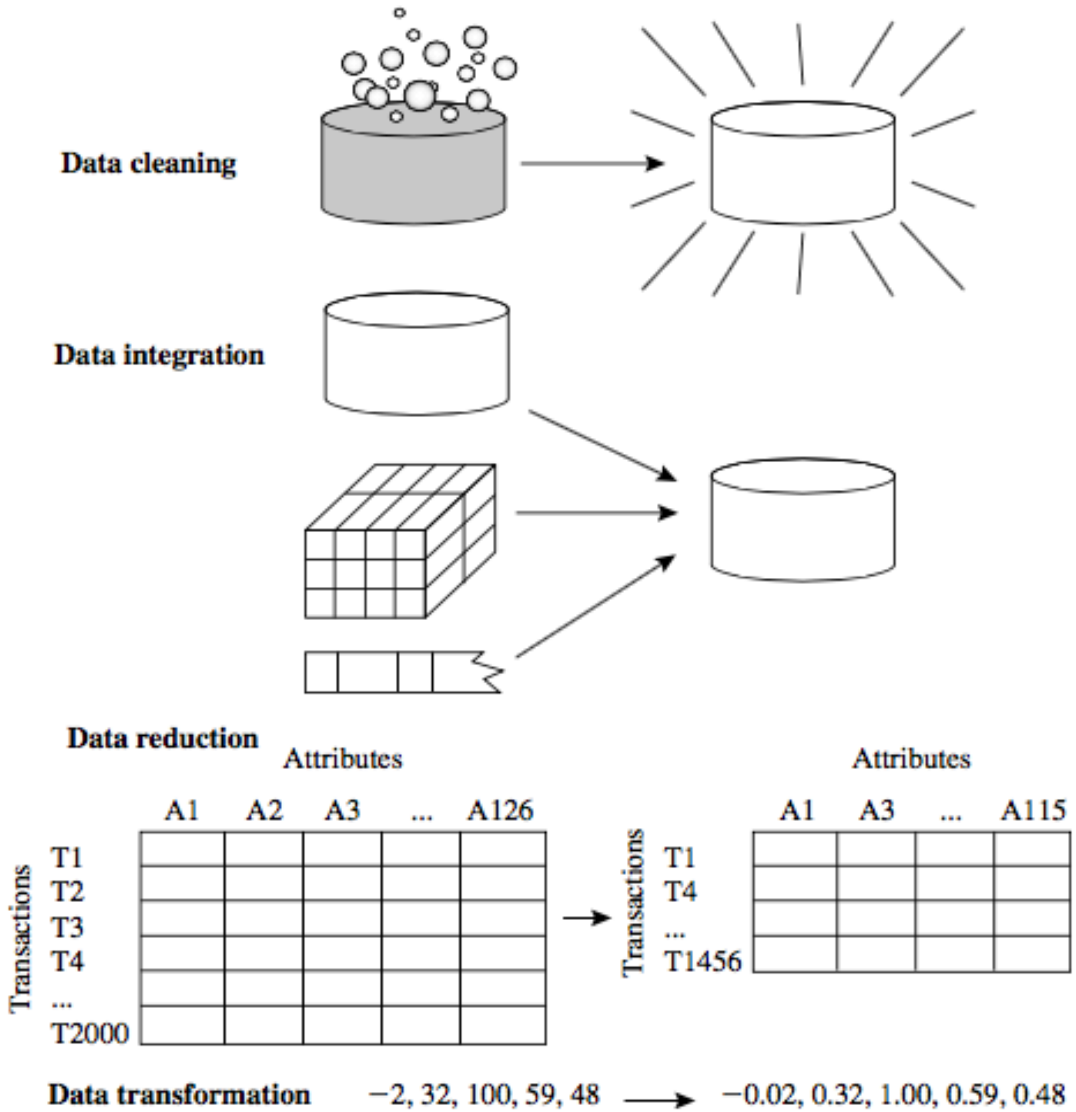
Data Preprocessing: Overview

- Inaccurate, incomplete and Inconsistent data are common in real world databases and data warehouses.

- Timeliness also affects data quality
  - Users do not update data in timely fashion

- Believability reflects how much the data are trusted by users
- Interpretability reflects how easy the data are understood.

Major Tasks in Data Preprocessing



Major Tasks in Data Preprocessing

- Data Cleaning
  - Filling in missing values
  - Smoothing noisy data
  - Identifying or removing outliers
  - Resolving inconsistencies
- Data Integration
  - Inconsistencies and redundancies may occur when integrating data from multiple sources.
- Data Reduction
  - Dimensionality Reduction
  - Numerosity Reduction
- Data Transformation
  - Normalization
  - Discretization
  - Concept Hierarchy Generation

Outline

- Data Preprocessing: Overview
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization

Data Cleaning

- Handle Missing Values
  1. Ignore the tuple
  2. Fill in the missing value manually
  3. Use a global constant to fill in the missing value
     - Unknown or $-\infty$
  4. Use a measure of central tendency for the attribute (e.g. the mean or median) to fill in the missing values
  5. Use the attribute mean or median for all samples belong to the same class of the given tuple
  6. Use the most probable value to fill in the missing value
     - Use regression, inference-based tools using Bayesian formalisim or decision trees.

Data Cleaning

- Handle Noisy Data
  - Noise is a random error or variance in a measured variable.
  - Smoothing techniques to remove numeric noises
    - Binning
      - Smoothing by bin means
      - Smoothing by bin medians
      - Smoothing by bin boundaries
    - Regression
    - Outlier Analysis
      - Clusters

  - Many smoothing methods are also used for data discretization (a form of data transformation) and data reduction.

Data Cleaning

- Handle Noisy Data

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

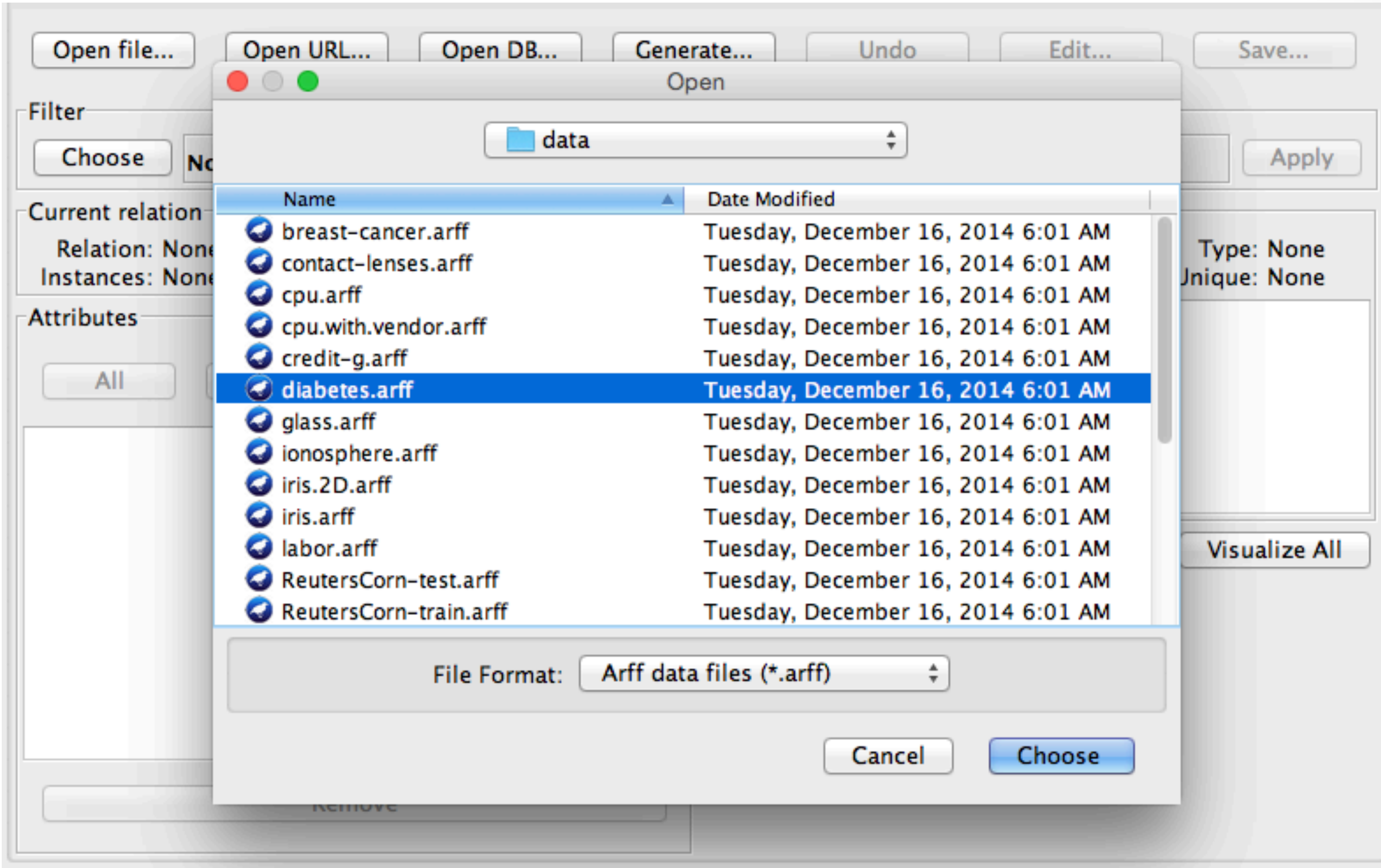
Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

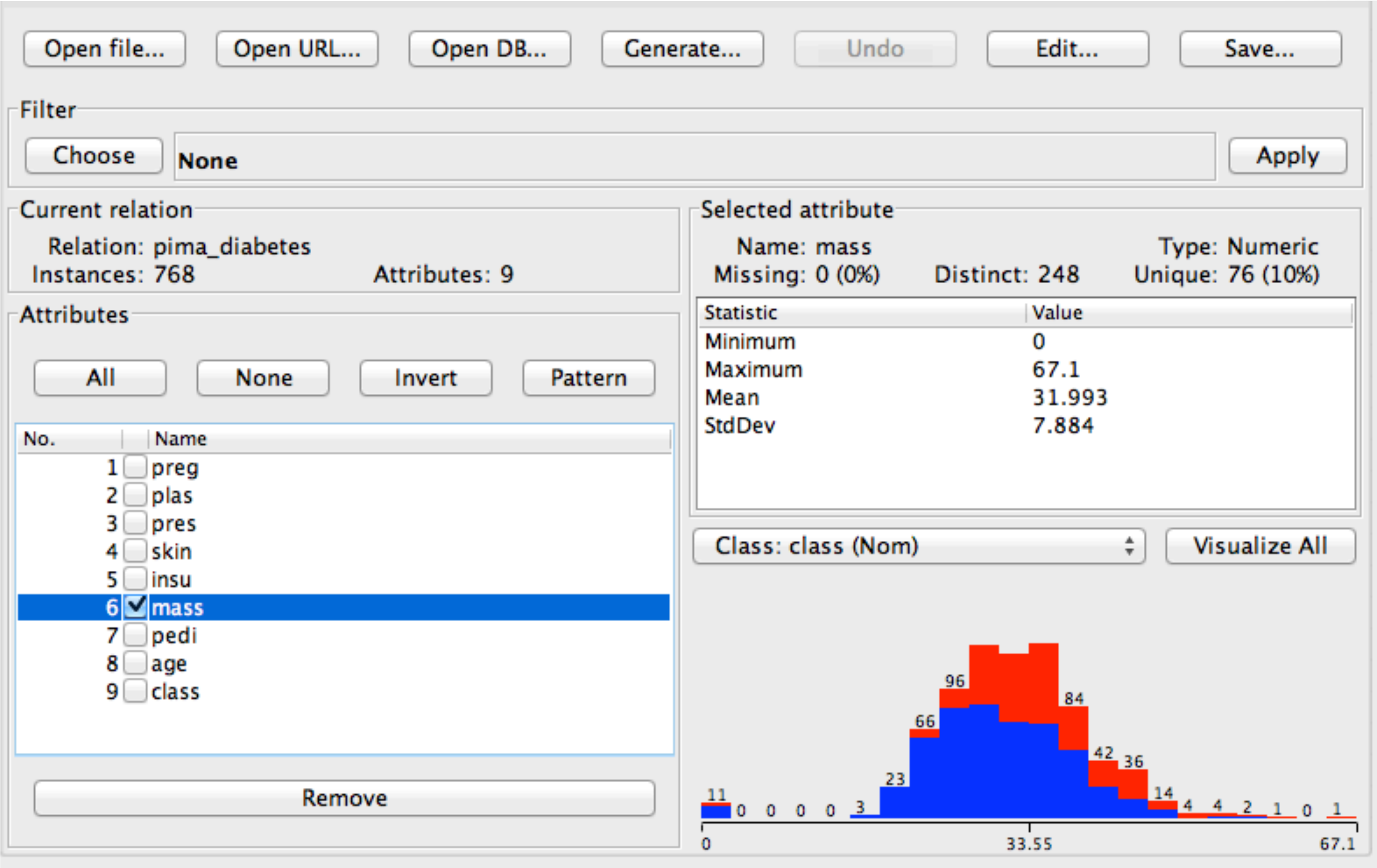
Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

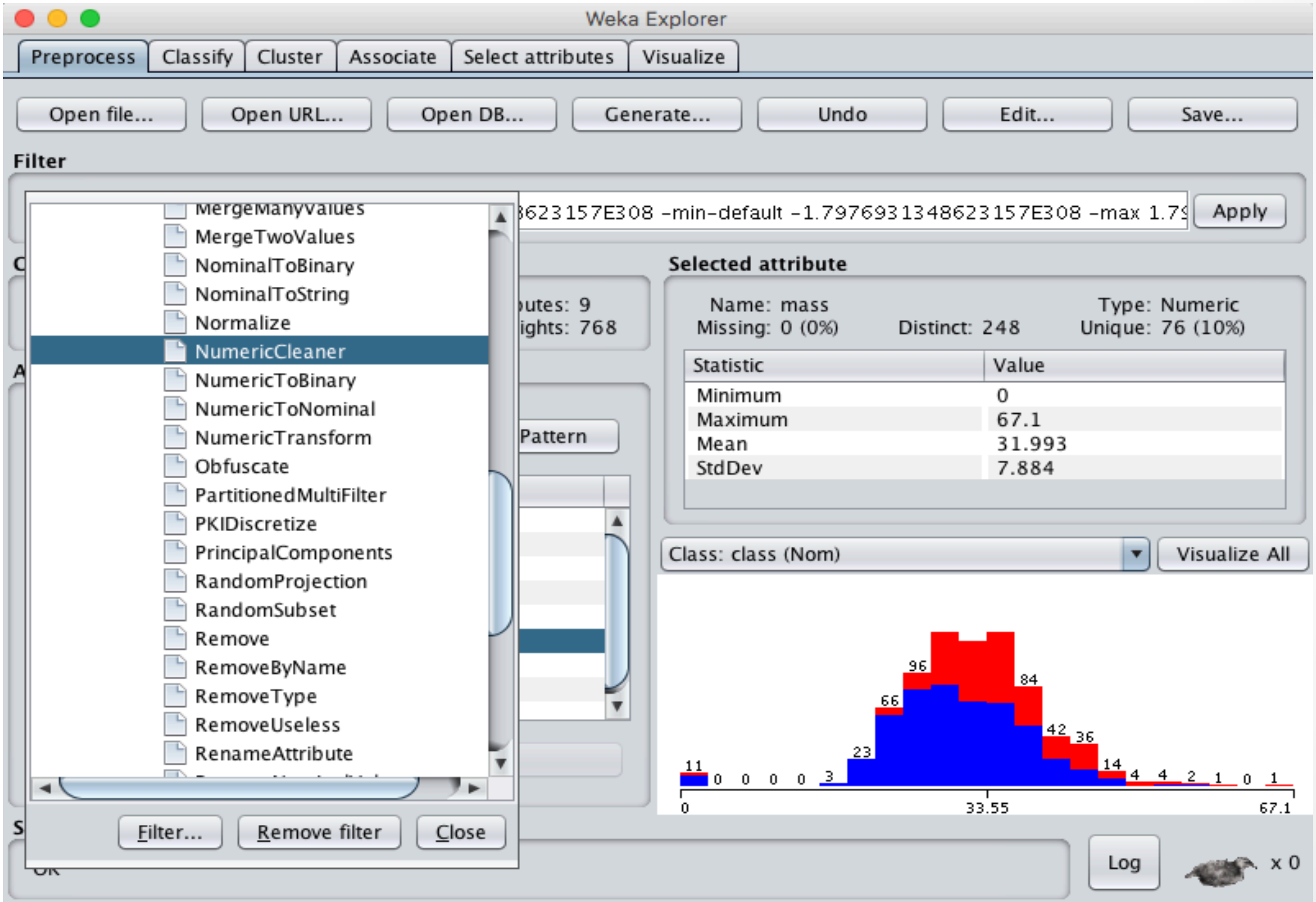
Handle Missing Values with Weka



Handle Missing Values with Weka

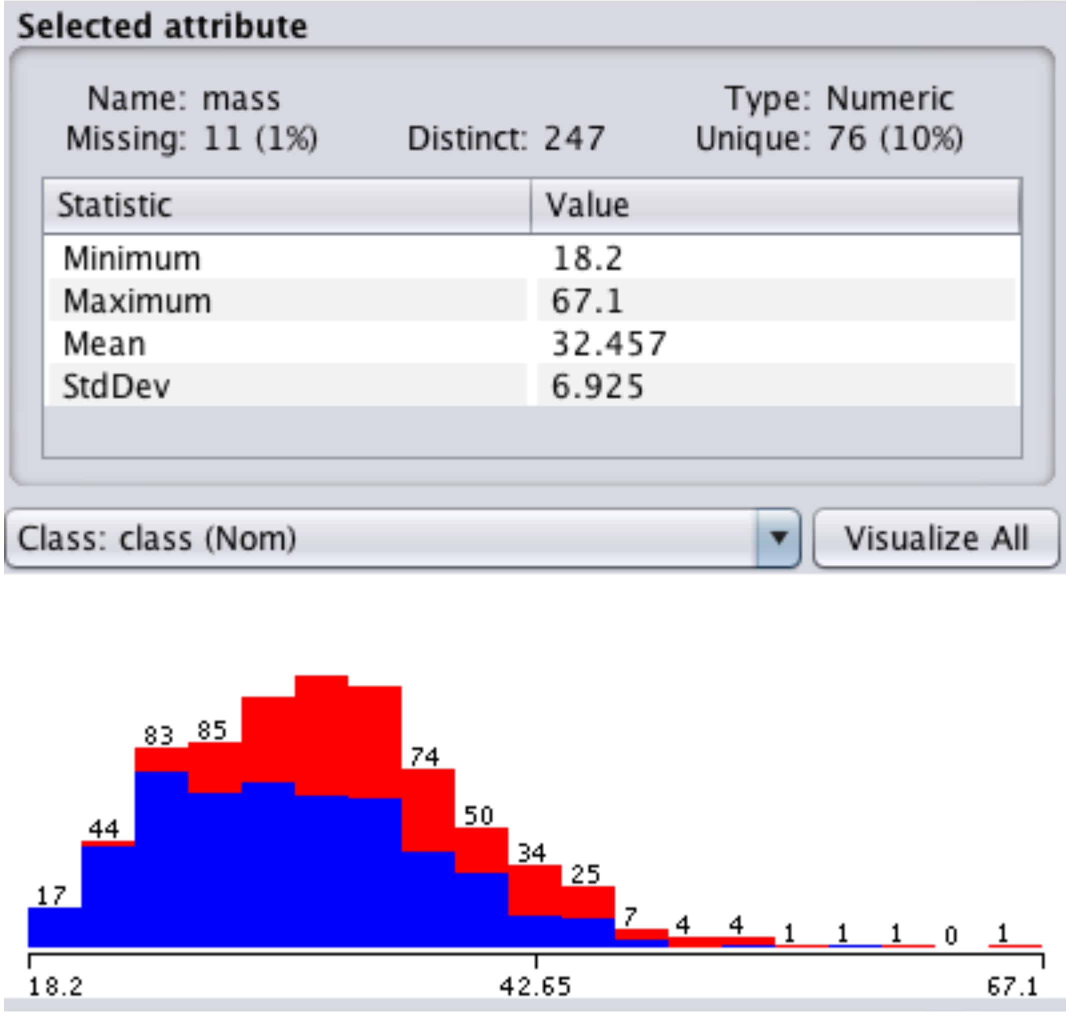


Handle Missing Values with Weka

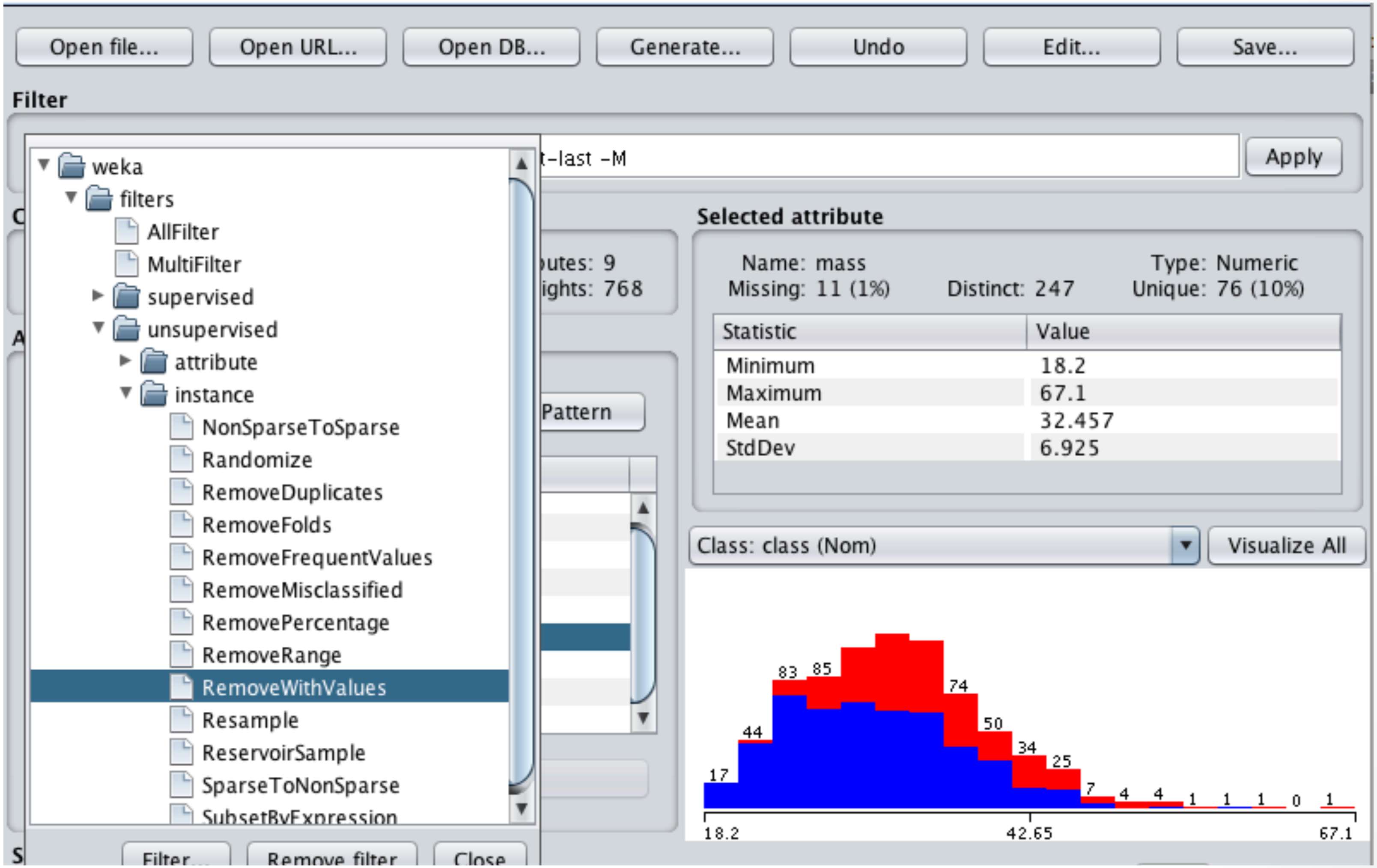


Handle Missing Values with Weka

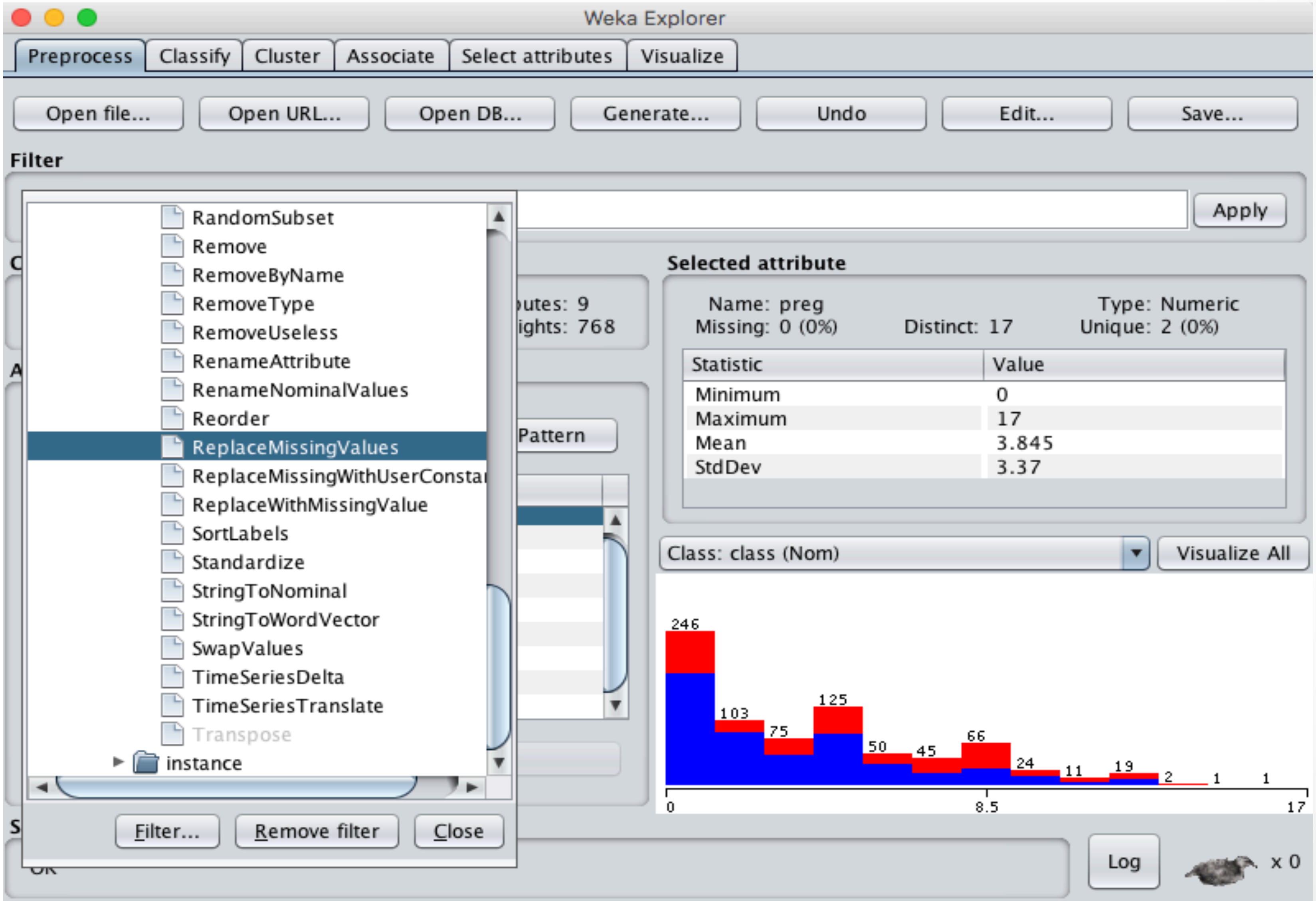
Replace mass values near 0 by NaN, we obtain



Remove missing values



Replace Missing Values



Outline

- Data Preprocessing: Overview
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization

Entity Identification Problem

- How to know customer_id in one database and cust_number in another refer to the same attribute?

- Metadata can be used to help avoid errors in schema integration
  - Metadata such as name, meaning, data type, range of values permitted for the attributes, and null rules to handle blank, zero, or null values.

Redundancy and Correlation Analysis

- Recognize redundancies by correlation analysis
  - Problem: Given two attributes, measure how strongly one attribute implies the other, based on available data.
  - For nominal data, we use χ^2-test (chi-square test)
  - For numeric attributes:
    - Correlation coefficient
    - Covariance

Chi-square test for Nominal data

Test for Hypothesis that A and B are independent

		A				
		a_1	a_2	….	a_c	Sum
B	b_1	n_{11}				$n_{1.}$
	b_2	n_{21}				$n_{2.}$
	…			n_{ij}		
	b_r					
	Sum	$n_{.1}$	$n_{.2}$		$n_{.c}$	$n_{..}$

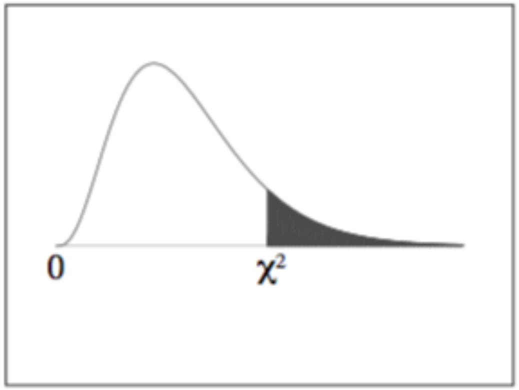
$$e_{ij} = \frac{count(B = b_i) \times count(A = a_j)}{n} = \frac{n_{i.} \times n_{.j}}{n_{..}}$$

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

Check for significance level with (r-1)(c-1) degrees of freedom

Chi-square test for Nominal data

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750

Chi-square test for Nominal data

- Example:
  - A group of 1500 people were surveyed. The **gender** and **preferred_reading** are noted.

	male	*female*	*Total*
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Note: Are *gender* and *preferred_reading* correlated?

Correlation Coefficient for Numeric Attributes

- Evaluate the correlation between two attributes A and B, we can use correlation coefficient (also known and Pearson's product moment coefficient):

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_ib_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B},$$

a_i, b_i are the values of A, B in the i-th data object (instance, tuple)

$\bar{A}, \bar{B}$ are the means of A, B

σ_A, σ_B are the standard deviations of A, B

Correlation Coefficient for Numeric Attributes

- Correlation Coefficient

$$-1 \leq r_{A,B} \leq +1$$

  - If the correlation coefficient is larger than 0, then A and B are positively correlated.
  - If the correlation coefficient is smaller than 0; then A and B are negatively correlated.
  - If the correlation coefficient is 0, then A and B are independent.
  - The larger the absolute value, the stronger the relationship between A, B.

- Note that, correlation DOES NOT imply causality.

Covariance of Numeric Data

- Correlation and Covariance are two similar measures to assess how much two attributes change together.

Expected values of A, B

$$E(A) = \bar{A} = \frac{\sum_{i=1}^{n} a_i}{n} \qquad E(B) = \bar{B} = \frac{\sum_{i=1}^{n} b_i}{n}.$$

Covariance

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}.$$

$$= E(A \cdot B) - \bar{A}\bar{B}.$$

Correlation

$$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B},$$

Covariance of Numeric Data

- Example

Stock Prices for *AllElectronics* and *HighTech*

Time point	*AllElectronics*	*HighTech*
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

If the stocks are affected by the same industry trends, will the prices of two company raise or fall together?

Covariance of Numeric Data

- Solution

$$E(AllElectronics) = \frac{6+5+4+3+2}{5} = \frac{20}{5} = \$4$$

$$E(HighTech) = \frac{20+10+14+5+5}{5} = \frac{54}{5} = \$10.80.$$

$$Cov(AllElectroncis, HighTech) = \frac{6\times 20 + 5\times 10 + 4\times 14 + 3\times 5 + 2\times 5}{5} - 4\times 10.80$$
$$= 50.2 - 43.2 = 7.$$

Outline

- Data Preprocessing: Overview
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization

Data Reduction: Overview

- Dimensionality Reduction
  - Wavelet Transform
  - Principle components analysis
  - Attribute Subset Selection
- Numerosity Reduction
  - Parametric methods
    - Statistical models instead of actual data
  - Nonparametric methods
    - Sampling
    - Clustering
    - Histograms
  - Data Compression

Dimension Reduction

- Discrete Wavelet Transform
  - Linear Signal Processing that, when applied to a data vector X, transforms it to a numerically different vector X', of **wavelet coefficients.**
  - The two vectors are of the same length.
  - Dimension Reduction is obtained by setting small coefficients to zeros, thus, obtaining sparse vector.
  - Examples are Haar Wavelet transform or Daubechies D4 Transform.

Dimension Reduction

- Haar Wavelet Transform
  - The forward transform
    - Given: a sequence of N elements $s_0, s_1, \ldots, s_{N-1}$
    - Calculate the averages and differences of consecutive elements
      - There are N/2 averages
      - N/2 (different) coefficients
      - The averages become the input for the next recursive step.
      - The recursion stops when it has only one average and one coefficient.

$$a_i = \frac{s_i + s_{i+1}}{2} \qquad c_i = \frac{s_i - s_{i+1}}{2}$$

    - We replace the original sequence of N elements with an average (the last round average) and a set of coefficients whose size is an increasing power of two.
  - The reverse transform

$$s_i = a_i + c_i$$
$$s_{i+1} = a_i - c_i$$

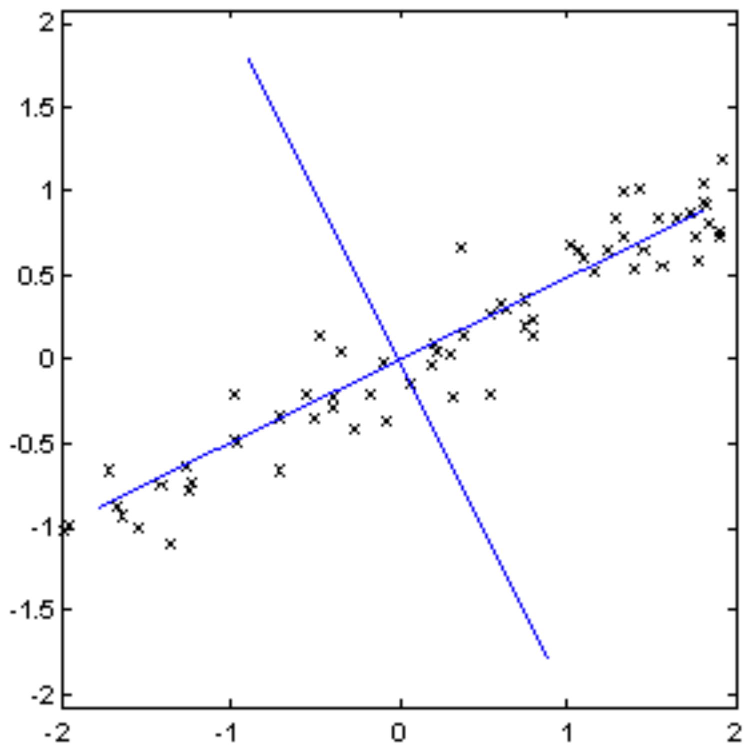
Dimension Reduction

- Haar forward transform via matrix multiply

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} \Leftarrow \begin{bmatrix} a_0 \\ c_0 \\ a_1 \\ c_1 \\ a_2 \\ c_2 \\ a_3 \\ c_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & -\frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \bullet \begin{bmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \\ s_7 \end{bmatrix}$$

Dimension Reduction

Principle Component Analysis



Dimension Reduction

Principle Component Analysis

- Principle Components
  - The first component corresponds to axis with largest variance
  - The second component corresponds to the axis with the second largest variance.
  - …

- PCA can be used for dimension reduction
  - Project the original attribute vector to the space that spans by at k principle components ($k <$ the original number of attributes p).

Dimension Reduction

Principle Component Analysis

Random Vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

Variance-Covariance Matrix

$$\text{var}(\mathbf{X}) = \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

Consider the linear combinations

$$\begin{aligned} Y_1 &= e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\ Y_2 &= e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\ &\vdots \\ Y_p &= e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p \end{aligned}$$

Dimension Reduction

Principle Component Analysis

- i-th Principle Component
  - Select $e_{i1}, e_{i2}, \ldots, e_{ip}$ that maximizes

$$\mathrm{var}(Y_i) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{ik} e_{il} \sigma_{kl} = \mathbf{e}_i' \Sigma \mathbf{e}_i$$

Subject to $\mathbf{e}_i'\mathbf{e}_i = \sum_{j=1}^{p} e_{ij}^2 = 1$

$$\mathrm{cov}(Y_1, Y_i) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{1k} e_{il} \sigma_{kl} = \mathbf{e}_1' \Sigma \mathbf{e}_i = 0,$$

$$\mathrm{cov}(Y_2, Y_i) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{2k} e_{il} \sigma_{kl} = \mathbf{e}_2' \Sigma \mathbf{e}_i = 0,$$

$$\vdots$$

$$\mathrm{cov}(Y_{i-1}, Y_i) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{i-1,k} e_{il} \sigma_{kl} = \mathbf{e}_{i-1}' \Sigma \mathbf{e}_i = 0$$

Dimension Reduction

Principle Component Analysis

- How do we find coefficients?
  - The solution involves the eigenvalues and eigenvectors o the variance-covariance matrix.
  - Let $\lambda_1, \ldots, \lambda_p$, these are ordered so that $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_p$
  - Let corresponding eigenvectors be $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$
  - It turns out that the elements for these eigenvectors will be coefficients of our principle components.
  - The variance for the i-th principle component is equal to the i-th eigenvalue.

$$\textbf{var}(Y_i) = \text{var}(e_{i1}X_1 + e_{i2}X_2 + \ldots e_{ip}X_p) = \lambda_i$$

  - Moreover, the principle components are uncorrelated with one another.

Dimension Reduction

Principle Component Analysis

- Spectral Decomposition Theorem
  - The variance-covariance matrix can be written as the sum over p eigenvalues, multiplied by the product of the corresponding eigenvector times its transpose.

$$
\begin{aligned}
\boldsymbol{\Sigma} &= \textstyle\sum_{i=1}^{p} \lambda_i \mathbf{e}_i \mathbf{e}_i' \\
&\cong \textstyle\sum_{i=1}^{k} \lambda_i \mathbf{e}_i \mathbf{e}_i'
\end{aligned}
$$

  - The second expression is a useful approximation if $\lambda_{k+1}, \lambda_{k+2}, \ldots, \lambda_p$ are small.

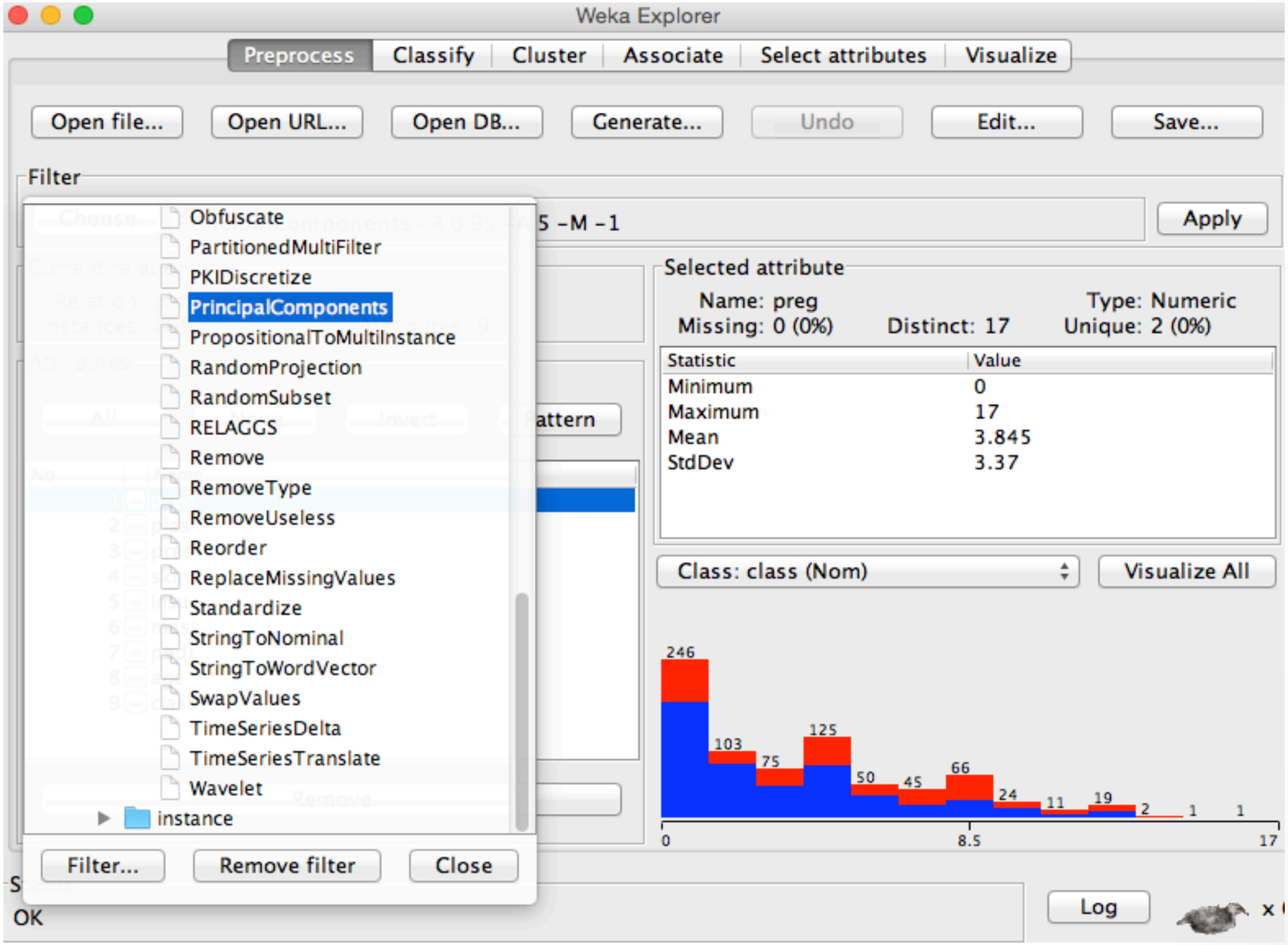
Dimension Reduction

Principle Component Analysis

- Dimension Reduction:
  - Using only k-components (eigenvectors) corresponding to k largest eigenvalues, we can transform X (p dimensions) to Y (with k dimensions) where k < p.

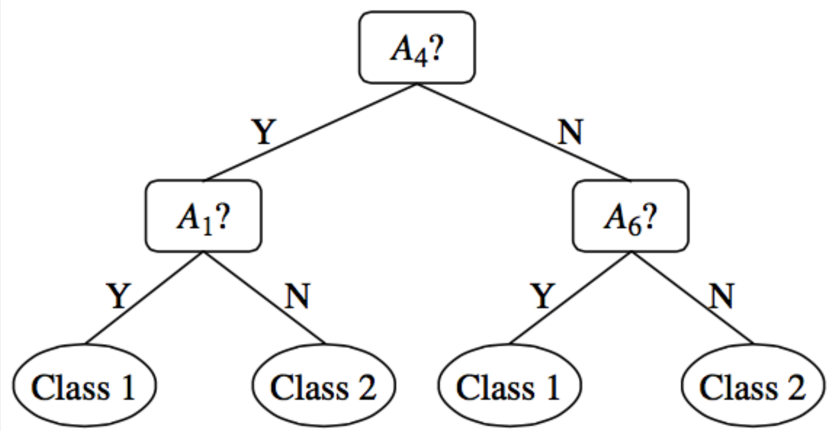
$$
\begin{aligned}
Y_1 &= e_{11}X_1 + e_{12}X_2 + ... + e_{1p}X_p \\
Y_2 &= e_{21}X_1 + e_{22}X_2 + ... + e_{2p}X_p \\
&\qquad\qquad\qquad\qquad\qquad\qquad\quad ... \\
Y_k &= e_{k1}X_1 + e_{k2}X_2 + ... + e_{kp}X_p
\end{aligned}
$$

Principle Component Analysis in Weka



Dimension Reduction

- Attribute Subset Selection

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => Reduced attribute set: $\{A_1, A_4, A_6\}$

Greedy (heuristic) methods for attribute subset selection.

Numerosity Reduction

- Sampling
  - Simple random sample without replacement.
  - Simple random sample with replacement
  - Cluster sample
  - Stratified sample

- It is possible (using the center limit theorem) to determine a sufficient sample size for estimating a given function with a specified degree of error.

Outline

- Data Preprocessing: Overview
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization

Data Transformation Strategies: Overview

- Smoothing
  - Remove noises from the data, techniques include binning, regression, and clustering
- Attribute construction
  - New attributes are constructed and added from the given set of attributes.
- Aggregation
  - Summary or aggregation operations are applied to the data.
- Normalization
  - Attributes are scaled so as to fall within ranges such as [-1,1] or [0,1]
- Discretization
  - Raw values of numeric attributes (e.g. age) are replaced by interval levels.
- Concept Hierarchy generation for nominal data:
    - E.g. street can be generalized to higher level concepts such as city or country.

Normalization

- Min-max normalizaton
  - Map a value, vi, of A to vi’ in the new range.

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A.$$

- Z-score normalization

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A},$$

- Normalization by decimal scaling: $v_i' = \frac{v_i}{10^j},$

Where j is the smallest integer such that $max(|v_i'|) < 1.$

Discretization

- Discretization by Binning
  - Binning method for data smoothing can also be used for discretization
- Discretization by Histogram Analysis
  - Equal-width histogram
  - Equal-frequency histogram
  - The histogram analysis algorithm can be applied recursively to obtain multilevel concept hierarchy.
- Discretization by Clustering, Decision Tree
- Discretization with ChiMerge
  - Find the best neighboring intervals and then merging them to form larger intervals, recursively.
  - It uses class label, the relative class frequency should be fairly consistent within an interval.
  - If two adjacent intervals have very similar distributions of classes then the intervals can be merged.

Summary

- Data quality is defined in terms of accuracy, completeness, consistency, timeliness, believability, and interpretability
- Data cleaning routines fill in missing values, smooth out noises while identifying outliers
- Data integration: reduce redundancies, inconsistencies.
- Data reduction
  - Dimensionality reduction
  - Numerosity reduction
- Data transformation
  - Normalization
  - Discretization.