



DATA ANALYSIS AND VISUALIZATION

INSTRUCTOR: UMME AMMARAH





DATA WRANGLING/ DATA PREPROCESSING

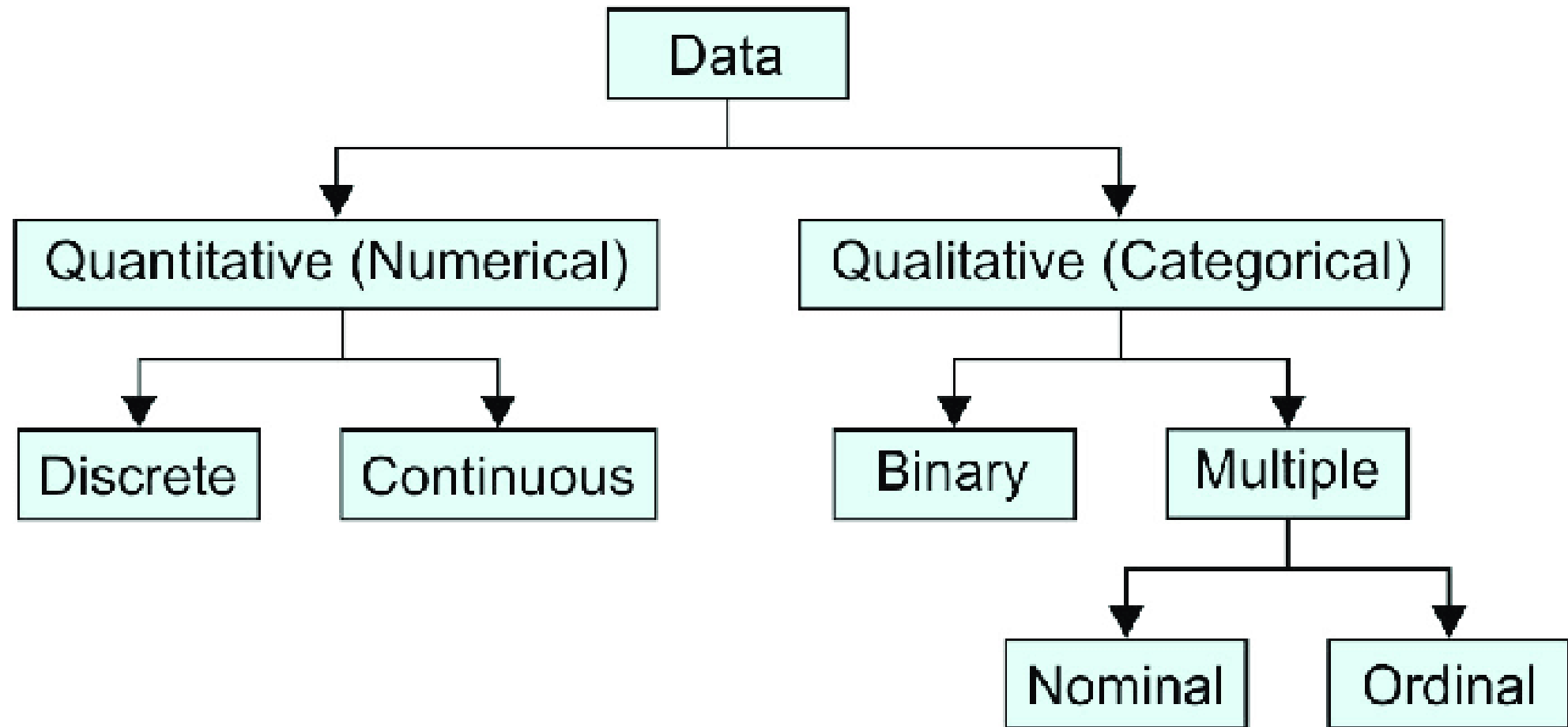


WHAT IS DATA WRANGLING

It is the process of converting raw data into a usable form. It may also be called:

- Data preprocessing
- Data preparation
- Data munging
- Data remediation

TYPES OF DATA



QUALITY OF DATA

Quality of data is measured on the basis of following factors:

- Accuracy
- Completeness
- Consistency
- Timeliness
- Believability
- Interpretability

WHY DATA PREPROCESSING?

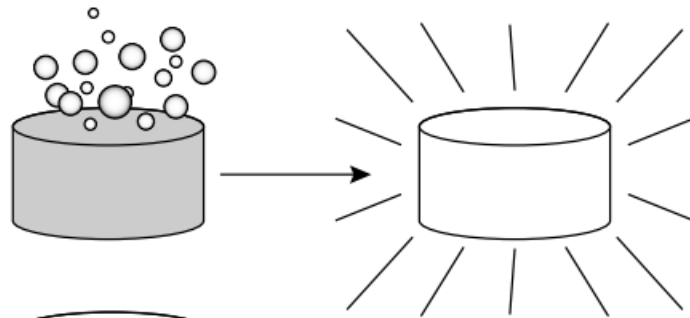
- Data in the real world is usually dirty
 - Incomplete
 - Noisy
 - Inconsistent

No quality data, no quality results!

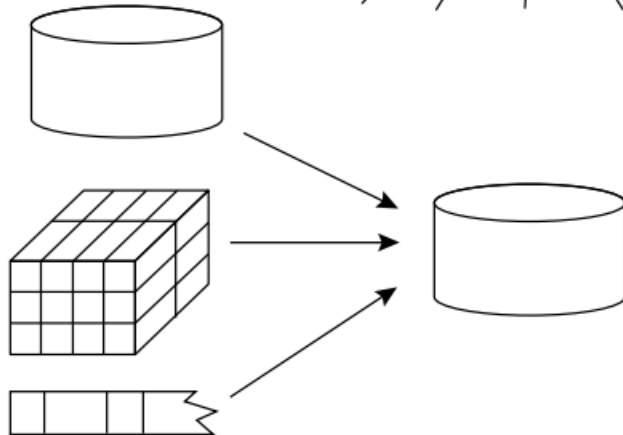
TASKS IN DATA PREPROCESSING

- Data cleaning
- Data integration
- Data reduction
- Data transformation
- Data discretization (for numerical data)

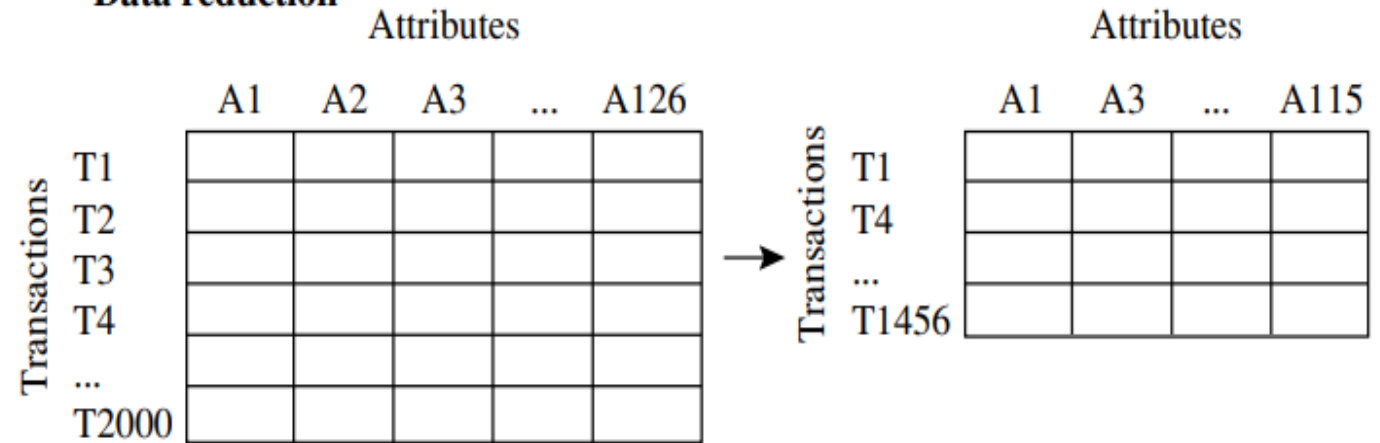
Data cleaning



Data integration



Data reduction



Data transformation

$-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$



DATA CLEANING



DATA CLEANING

- Fill in missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data

MISSING VALUES

- Data is not always available
- 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

HOW TO HANDLE MISSING DATA?

1. Ignore the tuple
2. Fill in the missing value manually
3. Use a global constant to fill in the missing value
4. Use a measure of central tendency for the attribute (e.g., the mean or median)
5. Use the attribute mean or median for all samples belonging to the same class.
6. Use the most probable value, inference-based such as regression, Bayesian formula, decision tree

NOISY DATA

- Random error in a measured variable is noise.
- Incorrect attribute values may due to
 - Faulty data collection instruments
 - Data entry problems
 - Data transmission problems
- Other data problems which requires data cleaning
 - duplicate records, incomplete data, inconsistent data

OUTLIERS

- Data points inconsistent with the majority of data
- Different outliers
 - Valid: CEO's salary,
 - Noisy: One's age = 200, widely deviated points
- Removal methods
 - Clustering
 - Curve-fitting
 - Hypothesis-testing with a given model

HOW TO HANDLE NOISY DATA

- Binning method:
 - first sort data and partition into bins
- Clustering
 - detect and remove outliers
- Regression
 - smooth by fitting the data into regression functions

BINNING

- Binning methods smooth a sorted data value by consulting its “neighborhood,” that is, the values around it
- Sorted values are distributed into a number of “buckets,” or bins

BINNING METHODS FOR DATA SMOOTHING

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

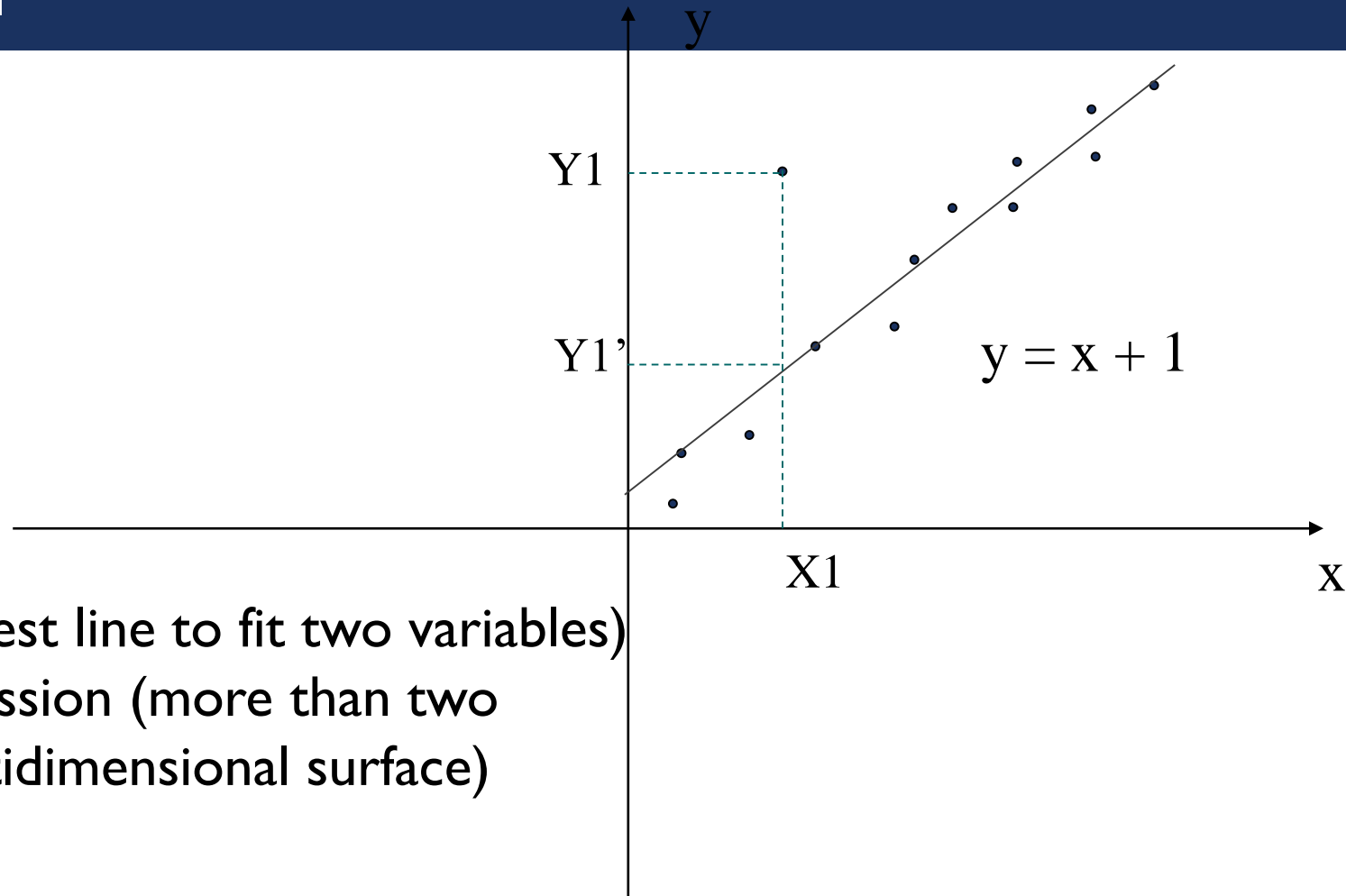
Smoothing by bin boundaries:

Bin 1: 4, 4, 15

Bin 2: 21, 21, 24

Bin 3: 25, 25, 34

REGRESSION



- Linear regression (best line to fit two variables)
- Multiple linear regression (more than two variables, fit to a multidimensional surface)

REGRESSION IMPUTATION

- *Correlated feature*
- *Linear regression*

$$y = a + bx$$

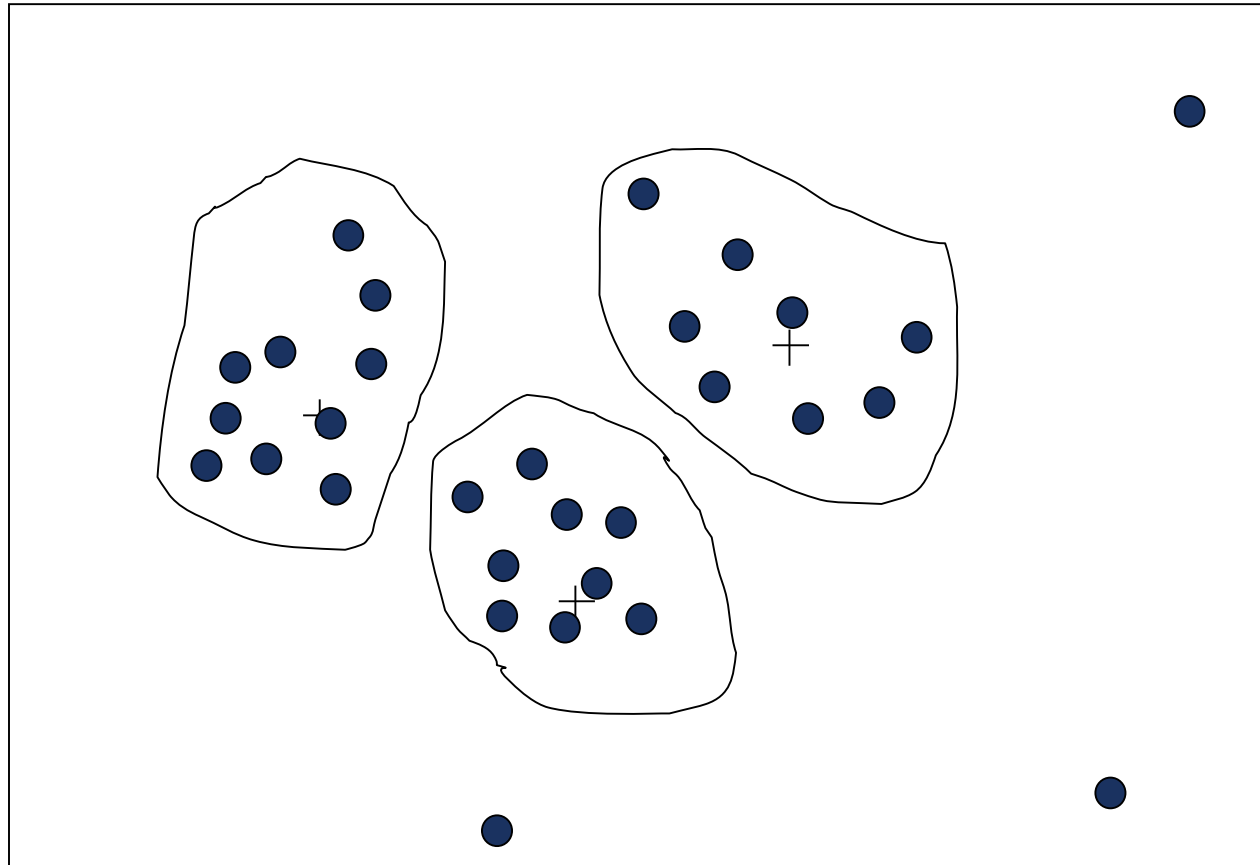
- Slope b and intercept a

Y	X3
13	5
26	0
26	8
18	10
NA	1
NA	5

$$a = \bar{y} - b\bar{x}$$
$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma_{xy}}{\sigma_x^2}$$

- a and b, together, are called the regression coefficients

CLUSTERING



IDENTIFYING OUTLIERS: Z-SCORE

z-score method for identifying outliers states that a data value is an outlier if it has a Z- score that is either less than or greater than 3.

$$Z\text{-score} = \frac{X - \text{mean}(X)}{SD(X)}$$

Example: Suppose we have the following values for salary (in thousands of dollars),
10,7,20,12,75,15,9,18,4,12,8,14

IDENTIFYING OUTLIERS: IQR(INTER QUARTILE RANGE)

Robust statistical methods for outlier detection
Less sensitive to the presence of the outliers themselves

The quartiles of a data set divide the data set into the following four parts, each containing 25% of the data:

The first quartile (Q_1) is the 25th percentile

The second quartile (Q_2) is the 50th percentile, that is, the median.

The third quartile (Q_3) is the 75th percentile

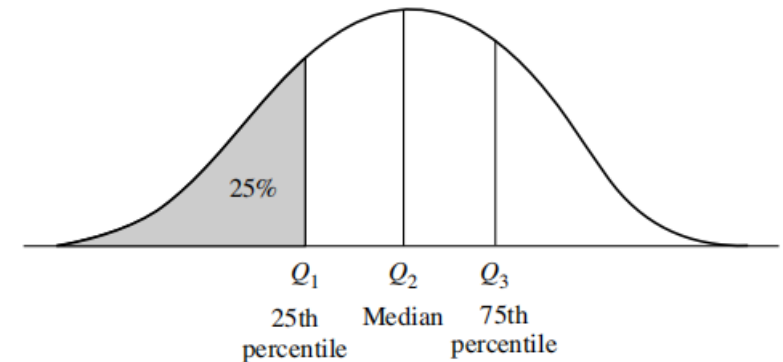
The IQR is a measure of variability, much more robust than the SD

$IQR = Q_3 - Q_1$

may be interpreted to represent the spread of the middle 50% of the data

A data value is an outlier if

- a. it is located $1.5(IQR)$ below the first quartile Q_1 , or
- b. it is located $1.5(IQR)$ above the third quartile Q_3 .





DATA INTEGRATION



DATA INTEGRATION

Merging of data from multiple sources

- Schema integration
 - integrate metadata from different sources
- Detecting and resolving data value conflicts

ENTITY IDENTIFICATION PROBLEM

How can equivalent real-world entities from multiple data sources be matched up?

- Schema Integration
- Object matching

HOW TO HANDLE IDENTIFICATION PROBLEM

- Using meta-data associated with attributes (name, meaning, data type, and value range)
 - It also supports data transformation, especially with varying data codes.
- Consider Data Structure when matching attributes
 - functional dependencies and referential constraints must align between sources.

REDUNDANCY

- Data redundancy occurs when the same piece of data is stored in two or more separate places
 - An attribute may be redundant if it can be “derived” from another attribute
 - Inconsistencies in attribute or dimension naming can also cause redundancies
- Some redundancies can be detected by correlation analysis.

CORRELATION ANALYSIS

- Given two attributes, such analysis can measure how strongly one attribute implies the other, based on the available data.
 - For nominal data:
 - χ^2 (chi-square) test
 - For numeric attributes:
 - correlation coefficient
 - covariance,

χ^2 CORRELATION TEST

- 1) Create a contingency table for any two attributes.
- 2) State hypothesis
- 3) Alpha value is selected (significance value)
- 4) Calculate expected values
- 5) If calculated value chi-square is less than or equal to the critical/tabular value, then the attributes are not correlated.

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

$$e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B = b_j)}{n},$$

	dog	cat	bird	total
men	207	282	241	730
women	234	242	232	708
total	441	524	473	1438

$$\frac{\text{row total} * \text{column total}}{\text{grand total}}$$

$$\frac{(\text{Observed_value} - \text{Calculated_value})^2}{\text{Calculated_value}}$$

	dog	cat	bird	total
men	223.87343533	266.00834492	240.11821975	730
women	217.12656467	257.99165508	232.88178025	708
total	441	524	473	1438

observed (o)	calculated (c)	(o-c)^2 / c
207	223.87343533	1.2717579435607573
282	266.00834492	0.9613722161954465
241	240.11821975	0.003238139990850831
234	217.12656467	1.3112758457617977
242	257.99165508	0.991245364156322
232	232.88178025	0.0033387601600580606
		4.542228269825232

EXAMPLE CONT.

- The degrees of freedom is defined as : $(\text{no. of rows} - 1) * (\text{no. of columns} - 1) = (2-1) * (3-1) = 2$

critical value of $\chi^2 \geq$ calculated value of χ^2

$$5.991 > 4.54$$

So we can infer that no correlation exist.

Library: `scipy`

Function: `chi2_contingency`

Critical values of the Chi-square distribution with d degrees of freedom

Probability of exceeding the critical value							
d	0.05	0.01	0.001	d	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

CORRELATION COEFFICIENT

- If r is greater than 0, then A and B are positively correlated, may indicate redundancy.

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

- If the resulting value is equal to 0, then A and B are independent.

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

- If the resulting value is less than 0, then A and B are negatively correlated.

COVARIANCE

- For assessing how much two attributes change together.

1) Calculate expected value for each attribute

2) Compute covariance

covariance of 0 imply independence

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

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

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

EXAMPLE

Stock Prices for *AllElectronics* and *HighTech*

Time point	<i>AllElectronics</i>	<i>HighTech</i>
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

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

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

$$\begin{aligned} \text{Cov}(\text{AllElectronics}, \text{HighTech}) &= \frac{6 \times 20 + 5 \times 10 + 4 \times 14 + 3 \times 5 + 2 \times 5}{5} - 4 \times 10.80 \\ &= 50.2 - 43.2 = 7. \end{aligned}$$

The positive covariance we can say that stock prices for both companies rise together.



DATA REDUCTION



DATA REDUCTION

Data reduction techniques can be applied to obtain a reduced representation of the data set.

- Dimensionality reduction
- Numerosity reduction
- Data compression

DIMENSIONALITY REDUCTION

It is the process of reducing the number of attributes under consideration.

- Wavelet Transforms
- Principal Components Analysis
- Attribute Subset Selection

DISCRETE WAVELET TRANSFORM (DWT)

Store only a small fraction of the strongest of the wavelet coefficients

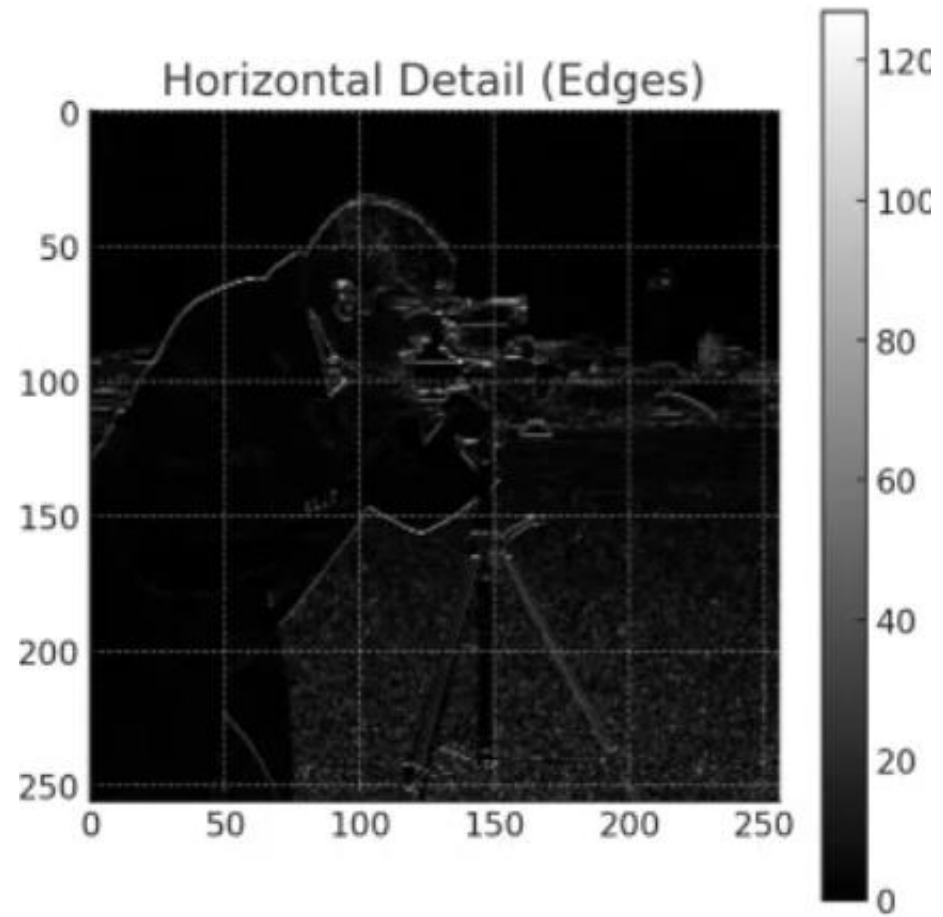
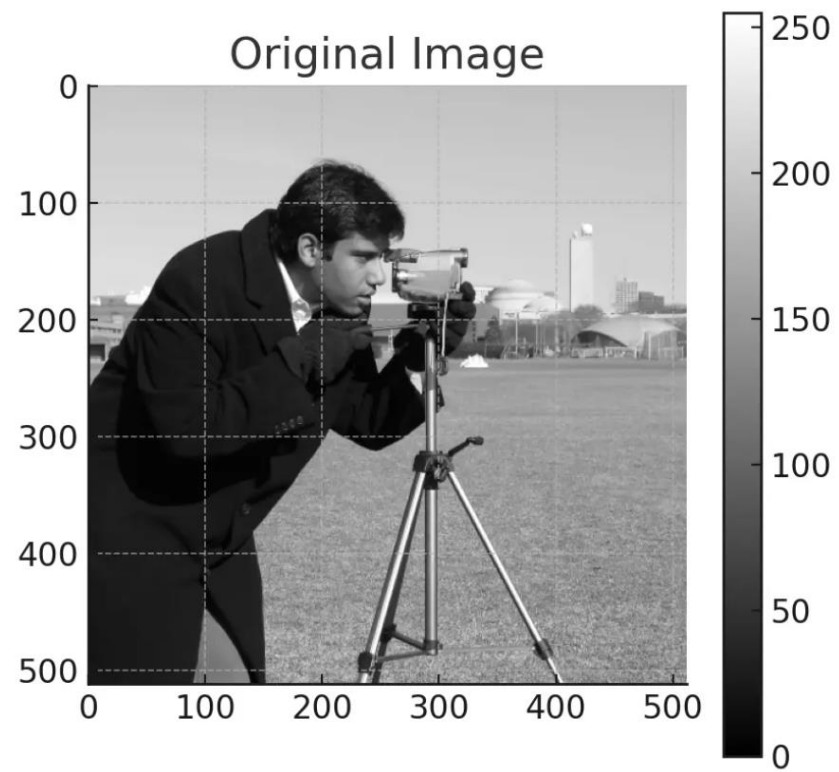
- Length, L , must be an integer power of 2 (padding with 0's, when necessary)
- Each transform has 2 functions: smoothing, difference
- Applies to pairs of data, resulting in two set of data of length $L/2$
- Applies two functions recursively, until reaches the length 2

EXAMPLES

$S = [2, 2, 0, 2, 3, 5, 4, 4]$ to $S' = [23/4, -11/4, 1/2, 0, 0, -1, -1, 0]$

$$\begin{array}{c|c} \begin{array}{l} [2, 1, 4, 4] \\ [1\frac{1}{2}, 4] \\ [2\frac{3}{4}] \end{array} & \begin{array}{l} [0, -1, -1, 0] \\ [\frac{1}{2}, 0] \\ [-1\frac{1}{4}] \end{array} \end{array}$$

EXAMPLE



PRINCIPAL COMPONENT ANALYSIS

Given N data vectors from n -dimensions, find $k \leq n$ orthogonal vectors (principal components) that can be best used to represent data

- Normalize input data
- Compute k orthonormal (unit) vectors, i.e., principal components
- 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

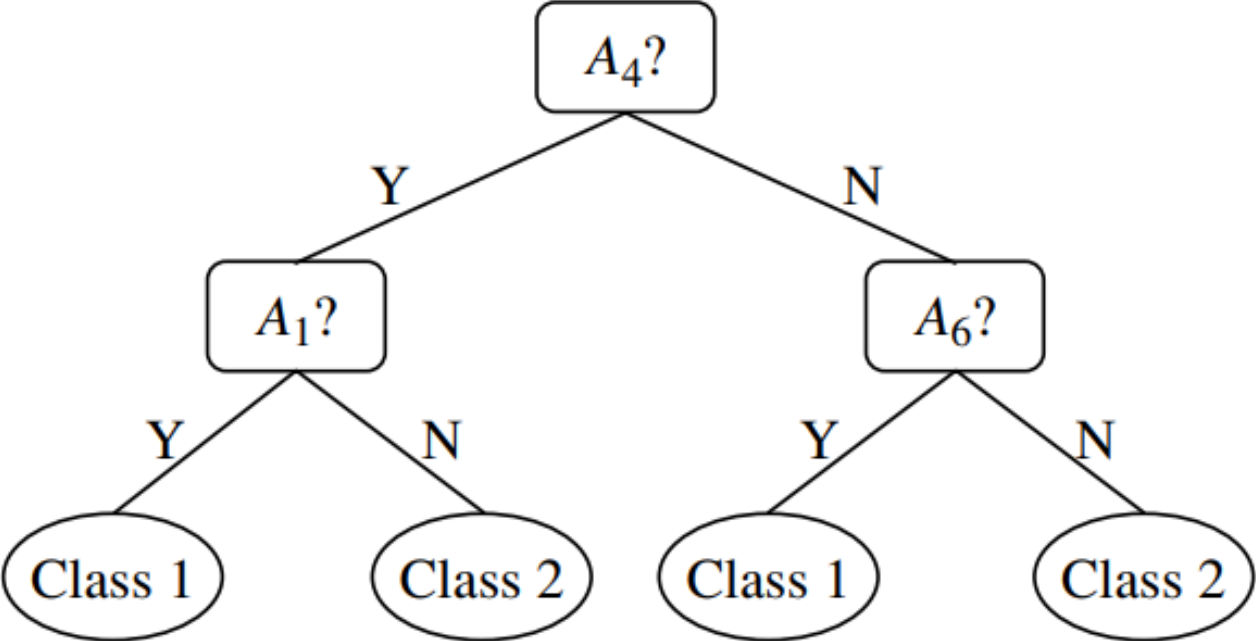
Works for numeric data only

ATTRIBUTE SUBSET SELECTION

Reduces the data set size by removing irrelevant or redundant attributes (or dimensions).

Attribute selection methods:

- Best step-wise feature selection: The best single-attribute is picked first
- Step-wise attribute elimination: Repeatedly eliminate the worst attribute
- Best combined attribute selection and elimination
- Decision Tree Induction

Forward selection	Backward elimination	Decision tree induction
<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p> <p>Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>	<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p> <p>$\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>	<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p>  <pre> graph TD A4["A4?"] -- Y --> A1["A1?"] A4 -- N --> A6["A6?"] A1 -- Y --> C1_1(("Class 1")) A1 -- N --> C2_1(("Class 2")) A6 -- Y --> C1_2(("Class 1")) A6 -- N --> C2_2(("Class 2")) </pre> <p>\Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>

NUMEROSITY REDUCTION

It replace the original data volume by alternative, smaller forms of data representation.

- Regression
- Histograms
- Clustering
- Sampling

NUMEROSITY REDUCTION

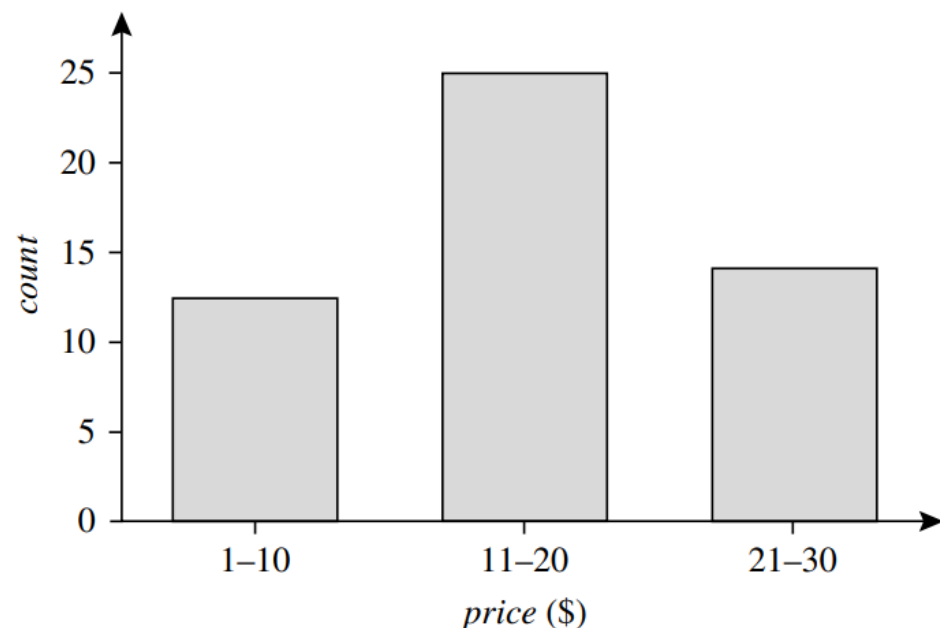
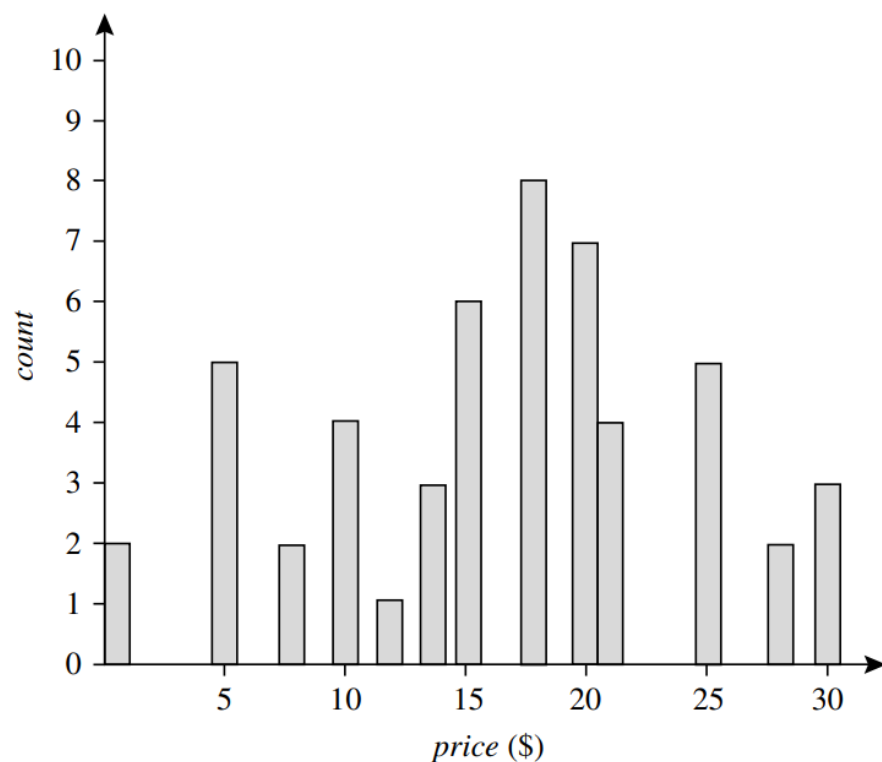
- Parametric methods(e.g., regression)
 - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
 - Ex.: Log-linear models—obtain value at a point in m -D space as the product on appropriate marginal subspaces
- Non-parametric methods
 - Do not assume models
 - Major families: histograms, clustering, sampling, ...

HISTOGRAMS

Histograms use binning to approximate data distributions and are a popular form of data reduction.

- Equal width bins
- Equal frequency bins

Histograms. The following data are a list of *AllElectronics* prices for commonly sold items (rounded to the nearest dollar). The numbers have been sorted: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.



DATA COMPRESSION

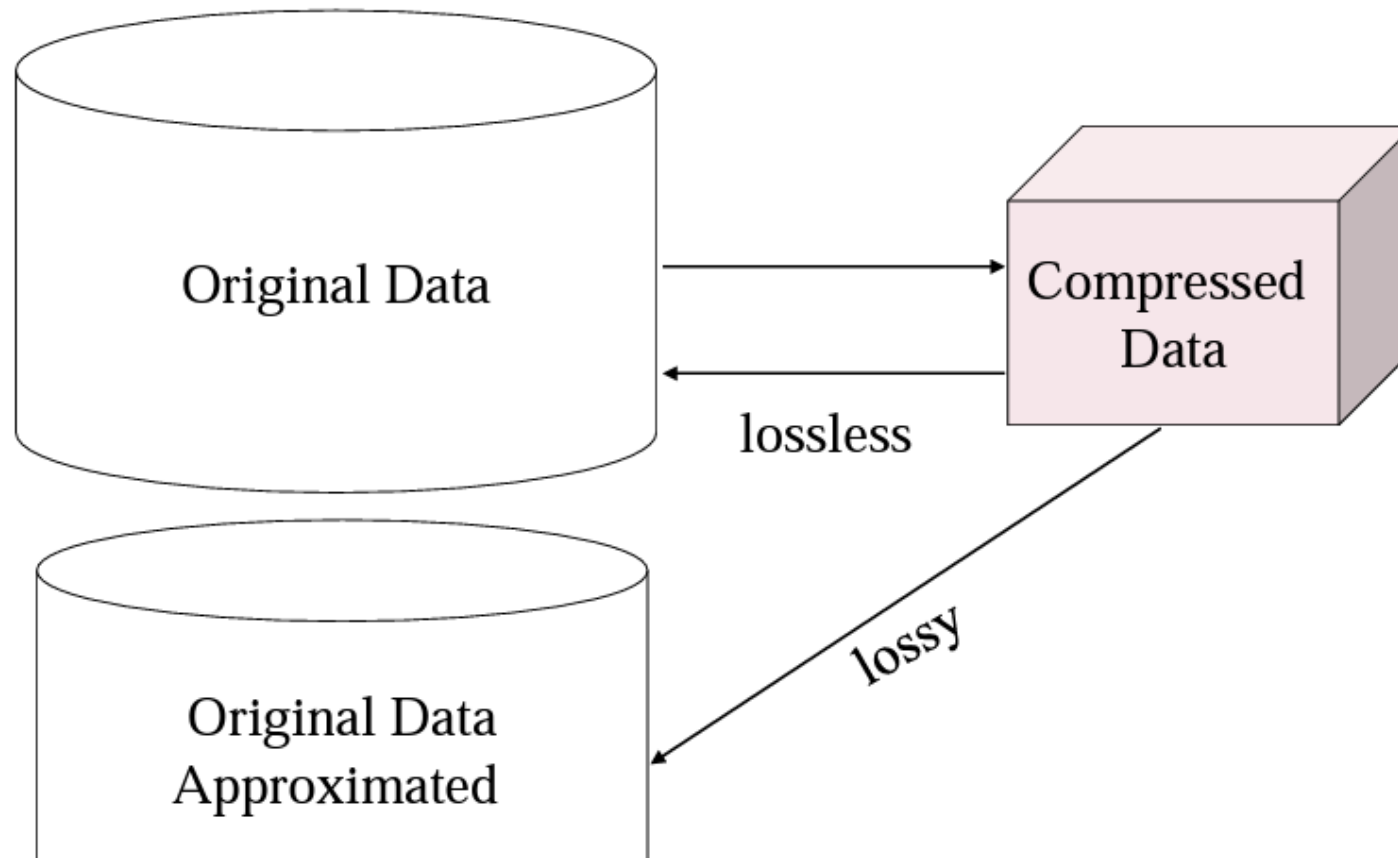
In data compression, transformations are applied so as to obtain a reduced or “compressed” representation of the original data.

- Lossless
- Lossy

DATA COMPRESSION

- String compression
 - There are extensive theories and well-tuned algorithms
 - Typically lossless, but only limited manipulation is possible without expansion
- Audio/video compression
 - Typically lossy compression, with progressive refinement
 - Sometimes small fragments of signal can be reconstructed without reconstructing the whole

DATA COMPRESSION





DATA TRANSFORMATION



DATA TRANSFORMATION

In data transformation, the data are transformed or consolidated into forms appropriate may be more efficient, and the patterns found may be easier to understand.

- Smoothing
- Attribute construction
- Aggregation
- Normalization
- Discretization
- Concept hierarchy generation for nominal data

NORMALIZATION TECHNIQUES:

- min-max normalization

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

- z-score normalization

$$v' = \frac{v - \text{mean}_A}{\text{stand_dev}_A}$$

- normalization by decimal scaling

$$v' = \frac{v}{10^j} \quad \text{Where } j \text{ is the smallest integer such that } \text{Max} (|v'|) < 1$$

MIN-MAX NORMALIZATION

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

- Suppose that the minimum and maximum values for the attribute income are \$12,000 and \$98,000, respectively.
- We would like to map income to the range [0,1]
- By min-max normalization, a value of \$73,600 for income is transformed to
- $(73,600 - 12,000) / (98,000 - 12,000) * (1 - 0) + 0 = 0.716$.

EXAMPLE

marks	marks after Min-Max normalization
8	
10	
15	
20	

marks	marks after Min-Max normalization
8	0
10	0.16
15	0.58
20	1

Z-SCORE NORMALIZATION

- Example: Suppose that the mean and standard deviation of the values for the attribute income are \$54,000 and \$16,000, respectively.
- With z-score normalization, a value of \$73,600 for income is transformed to $73,600 - 54,000 / 16,000 = 1.225$.
- A variation of this z-score normalization replaces the standard deviation by the mean absolute deviation of A
- The mean absolute deviation of A, denoted s_A , is

$$s_A = \frac{1}{n}(|v_1 - \bar{A}| + |v_2 - \bar{A}| + \cdots + |v_n - \bar{A}|)$$

DECIMAL SCALING

- Suppose that the recorded values of A range from -986 to 917 .
- The maximum absolute value of A is 986 .
- To normalize by decimal scaling, we therefore divide each value by 1000 (i.e., $j = 3$) so that -986 normalizes to -0.986 and 917 normalizes to 0.917 .

NORMALIZATION

- Normalization can change the original data quite a bit, especially when using z-score normalization or decimal scaling
- It is also necessary to save the normalization parameters (e.g., the mean and standard deviation if using z-score normalization) so that future data can be normalized in a uniform manner.

DISCRETIZATION

- It is often necessary to transform a continuous attribute into a categorical attribute (discretization), and both continuous and discrete attributes may need to be transformed into one or more binary attributes (binarization)

WHY DISCRETIZATION IS USED?

- Reduce data size
- Transforming quantitative data to qualitative data
- Discretization can be performed recursively on an attribute

DISCRETIZATION BY BINNING: EXAMPLE

Age	10,11,13,14,17,19,30, 31, 32, 38, 40, 42,70 , 72, 73, 75
------------	--

Table: Before discretization

Attribute	Age	Age	Age
	10,11,13,14,17,19,	30, 31, 32, 38, 40, 42	70 , 72, 73, 75
After Discretization	Young	Middle	Old

DISCRETIZATION BY BINNING: EXAMPLE

Iris Dataset

Table 3.1. Discretized sepal length attribute

Bins	Domain	Counts
[4.3, 5.2]	Very Short (a_1)	$n_1 = 45$
(5.2, 6.1]	Short (a_2)	$n_2 = 50$
(6.1, 7.0]	Long (a_3)	$n_3 = 43$
(7.0, 7.9]	Very Long (a_4)	$n_4 = 12$

BINARIZATION

Table 2.6. Conversion of a categorical attribute to five asymmetric binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
<i>awful</i>	0	1	0	0	0	0
<i>poor</i>	1	0	1	0	0	0
<i>OK</i>	2	0	0	1	0	0
<i>good</i>	3	0	0	0	1	0
<i>great</i>	4	0	0	0	0	1