2110773 Data Mining Chapter2: Data Preprocessing

- GARBAGE IN → GARBAGE OUT
- IMPORTANT & TIME-CONSUMING TASK IN KDD
- PRACTICE IS EVERYTHING

- รศ. ดร. ญาใจ ลิ่มปิยะกรณ์

2

2110773-2 2/67

Types of Dataset

Record	Relational records Document data: text documents Transaction
Graph and network	World Wide Web Social or information networks Molecular Structures
Others	Image Video data: sequence of images Temporal/ Time-series Spatial data: maps

Data Object

3

2110773-2 2/67

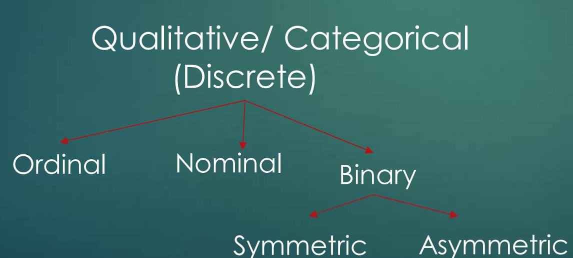
- Data sets are made up of data objects.
- A **data object** represents an entity. For examples:
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples , examples, instances, data points, objects, tuples*.
- Data objects are described by **attributes**.
- Database rows -> data objects; columns ->attributes.
- Attribute (or **dimension, feature, variable**): a data field, representing a characteristic or feature of a data object, e.g., *customer _ID, name, address, phone*

Attribute Data Types

4

2110773-2 2/67

1. Qualitative/ Quantitative
2. Categorical/ Numeric
3. Discrete/ Continuous
   - Discrete: Has only a finite or countably infinite set of values. Sometimes, represented as integer variables
   - Continuous: Has real numbers (floating-point) as attribute values. Practically, real values can only be measured.



Attribute Types

- **Nominal:** categories, states, or "names of things". Categories cannot be compared
- **Binary:** Nominal attribute with only 2 states (0 and 1)
  - *Symmetric binary*: both outcomes equally important
  - *Asymmetric binary*: outcomes not equally important. Convention: assign 1 to most important outcome (e.g., covid19 positive)
- **Ordinal:** Values have a meaningful order (ranking) but magnitude between successive values is not known. Categories with an implied order

- Quantity (integer or real-valued)
- **Interval**
  - Measured on a scale of **equal-sized units**
  - Values have order
  - No true zero-point
- **Ratio**
  - Inherent **zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).

Data Type Examples

Data Type	Examples
Nominal	color, bloodType, zipCode, ID#, occupation, political party
Ordinal	medal, satisfaction, grade, frequency, academic ranking
Binary- symmetric	gender
Binary- asymmetric	labTest
Interval	celcius, farenheit, pH,
Ratio	kelvin, exam score, weight, height, pulse, monetary quantities

Interval Data: No true zero, differences (subtraction) are interpretable.
Data can be added/ subtracted at interval scale but nonsense be multiplied/ divided.
Ex. If a day's temperature in celcius/ farenheit is twice than the other day,
we cannot say that one day is twice as hot as another day.

Ratio Data: True zero exists. Zero means none of that variable value, e.g. zero kelvin means no heat.
The ratio of two measurements has a meaningful interpretation.

7

2110773-2 2/67

%	Adverb of Frequency	Example
100%	Always เสมอ	I always study after class
90%	Usually เป็นประจำ/ปกติ	I usually walk to work
80%	Normally / Generally โดยทั่วไป	I normally get good marks
70%	Often / Frequently บ่อย	I often read in bed at night
50%	Sometimes บางครั้ง	I sometimes sing in the shower
30%	Occasionally บางโอกาส	I occasionally go to bed late
10%	Seldom ไม่ค่อยจะ	I seldom put salt on my food
5%	Hardly ever / Rarely นานๆที	I hardly ever get angry
0%	Never ไม่เคย	Vegetarians never eat meat

8

2110773-2 2/67

Scales of Measurement

Data	Nominal	Ordinal	Interval	Ratio
Labeled				
Order				
Measurable Difference				
True Zero Starting Point				

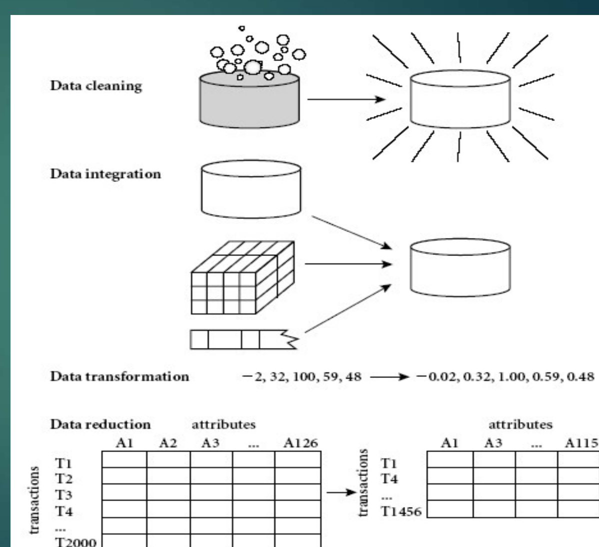
** A scale is an ordered set of values, continuous or discrete, or a set of categories to which an attribute is mapped.

Survey

1. How old are you? ___________years
2. Are you: Male Female
3. How much do you spend on groceries each week? _______Baht
4. How many cups of coffee do you buy in a week? ________
5. Which type of coffee do you like most?
   Latte Espresso Cappuccino Americano
6. How likely are you to buy more than a cup of coffee per day?
   Very Likely Likely Not Likely Very Unlikely

Data Preprocessing

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction



Data Cleaning

11

2110773-2 2/67

- Fill in missing data
- Smooth noisy data- random error or variance in a measured variable
- Identify or remove outliers
- Resolve inconsistencies
  - Same name means differently (BL= blue/ black)
  - Different names appear the same (Bill vs. Williams)
  - Inappropriate values (Male-Pregnant; born Feb 29, 2562; age=41 birthday=28/08/2010)
  - Due to inconsistent Unit of Measure

Missing Data

12

2110773-2 2/67

- Various reasons:
  - truly missed/ impossible to always have a value
  - Intentional (disguised missing data)
  - not measured due to no equipment or not able to measure in the past
  - Inconvenient, expensive
- Some methods
  - Leave as is, however, some algo can't deal w/ missing values and the program may refuse to continue or lead to inaccurate results
  - Remove the instance with missing value (e.g. in case of huge dataset or missing class label)
  - A global constant, e.g. 999,999 (valid values are much smaller) or -1 (valid values are non-negative). Watch out for zeros as some features can use this as the boolean representation! or "unknown" can be treated as a new class ?!
  - Imputing :
    - Attribute mean/median (Numerical variables); mode (Categorical variables)
    - **S**ubstitute w/ valid values of a certain feature e.g. fill in the seasonal averages of temperature for a certain location for missing temperature values given a date
    - Model-based/ inference-based: Regression, Decision Tree, k-nearest neighbor, Bayesian …)

Noisy Data

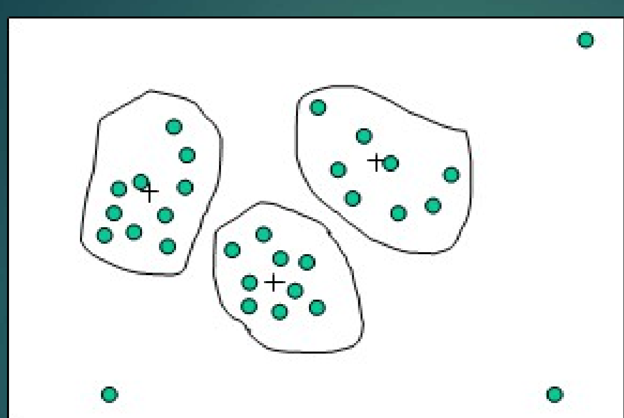
13

2110773-2 2/67

- Random error or variance in a measured variable
  - Regression- smooth by fitting the data into regression functions
- Outliers are noisy data or data points inconsistent with the majority of data, e.g. one's age = 200 year, height=3 metre, widely deviated points
  - Detect and remove outliers- Clustering
  - Truncate outliers- Bell curve, Box plots

14

2110773-2 2/67



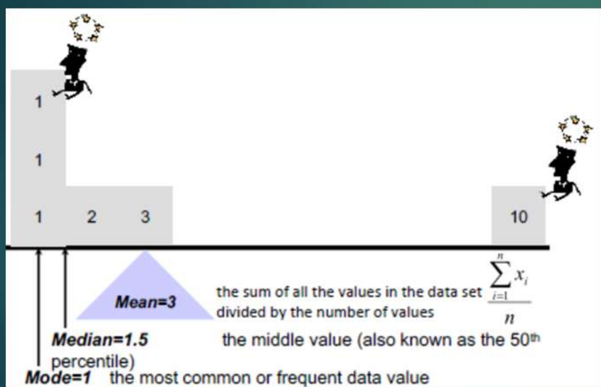
Clustering

Data Distribution

15

2110773-2 2/67

1. Central Tendency/ Center



Measure of Central Tendency (Representative value): Mean, Median, Mode

2. Spread/ Dispersion

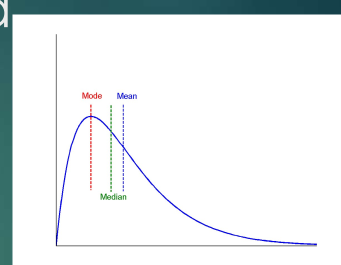
Measure	Definition
Range	the difference between the maximum and minimum data values
Interquartile Range	the difference between the 25th and 75th percentiles
Variance	a measure of dispersion of the data around the mean
Standard Deviation	a measure of dispersion expressed in the same units of measurement as your data (the square root of the variance)

Symmetric vs. Skewed Data

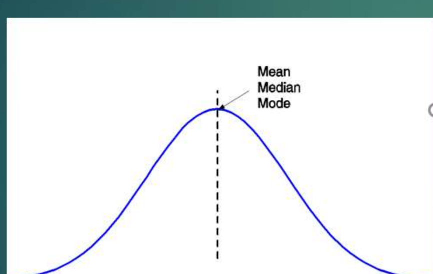
16

2110773-2 2/67

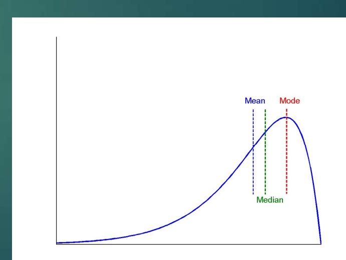
- Median, mean and mode of symmetric, positively and negatively skewed data



positively skewed



symmetric



negatively skewed

17

2110773-2 2/67

Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

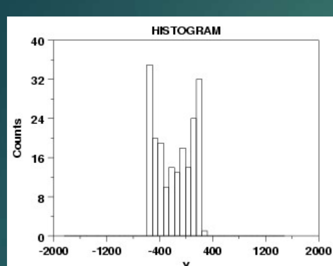
When to use Mean, Median, Mode

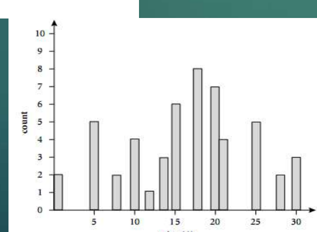
Graphical Displays of Distribution

18

2110773-2 2/67

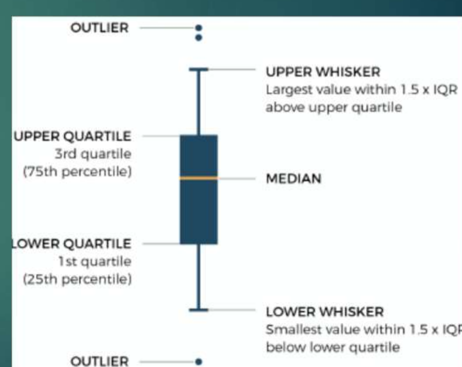
- Histogram Graph display of frequencies, shown as bars with numeric values on X axis





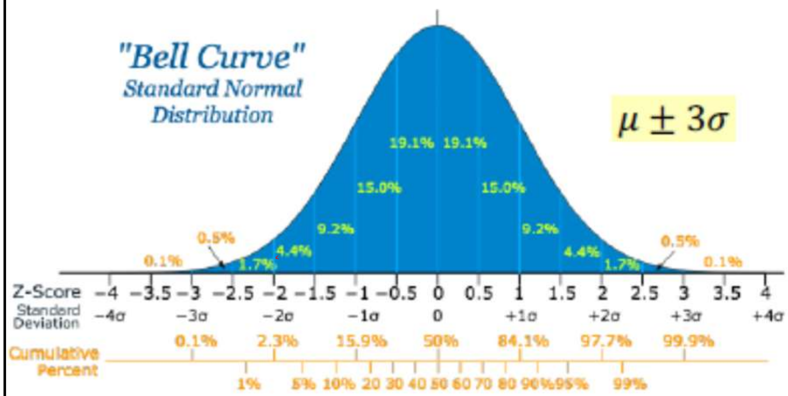
Singleton Histogram

- Box plots



19

Truncate Outliers: Bell Curve



$\mu \pm 3\sigma$

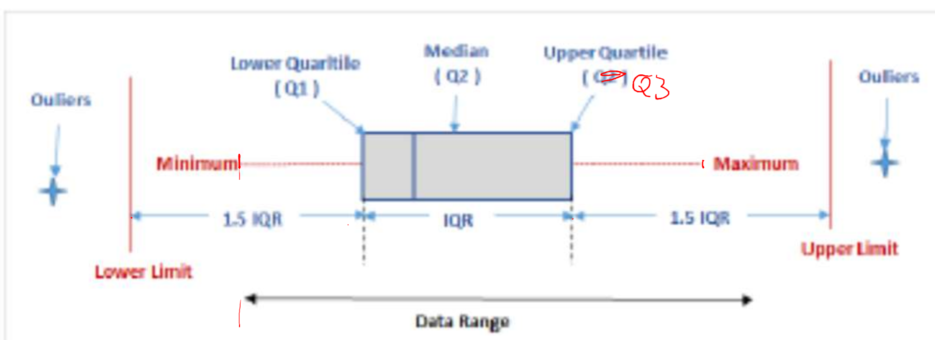
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^{\,2} - \mu^2$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^{\,2} - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$$

Variance and standard deviation (sample: s, population: σ)

Standard deviation is the square root of variance

20



Truncate Outliers: Box Plots

Interquartile Range

21

2110773-2 2/67

- **IQR is a measure of spread indicating where the bulk of the values lie.**
  - **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
  - **Inter-quartile range**: IQR = $Q_3 - Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier**: usually, a value higher/lower than 1.5 x IQR

IQR Calculation

22

2110773-2 2/67

Odd set of numbers

- Step 1: **Put the numbers in order.**
  1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.
- Step 2: **Find the median.**
  1, 2, 5, 6, 7, **9**, 12, 15, 18, 19, 27.
- Step 3: **Place parentheses around the numbers above and below the median.**
  Not necessary **statistically**, but it makes Q1 and Q3 easier to spot.
  (1, 2, 5, 6, 7), 9, (12, 15, 18, 19, 27).
- Step 4: **Find Q1 and Q3**
  Think of Q1 as a median in the lower half of the data and think of Q3 as a median for the upper half of data.
  (1, 2, **5**, 6, 7), **9**, (12, 15, **18**, 19, 27). Q1 = 5 and Q3 = 18.
- Step 5: **Subtract Q1 from Q3 to find the interquartile range.**
  18 – 5 = 13.

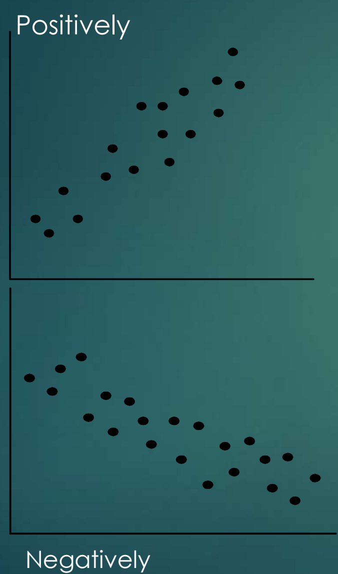
Even set of numbers

- Step 1: **Put the numbers in order.**
  3, 5, 7, 8, 9, 11, 15, 16, 20, 21.
- Step 2: **Make a mark in the center of the data:**
  3, 5, 7, 8, 9, | 11, 15, 16, 20, 21.
- Step 3: **Place parentheses around the numbers above and below the mark you made in Step 2–it makes Q1 and Q3 easier to spot.**
  (3, 5, 7, 8, 9), | (11, 15, 16, 20, 21).
- Step 4: **Find Q1 and Q3**
  Q1 is the median (the middle) of the lower half of the data, and Q3 is the median (the middle) of the upper half of the data.
  (3, 5, **7**, 8, 9), | (11, 15, **16**, 20, 21). Q1 = 7 and Q3 = 16.
- Step 5: **Subtract Q1 from Q3.**
  16 – 7 = 9.

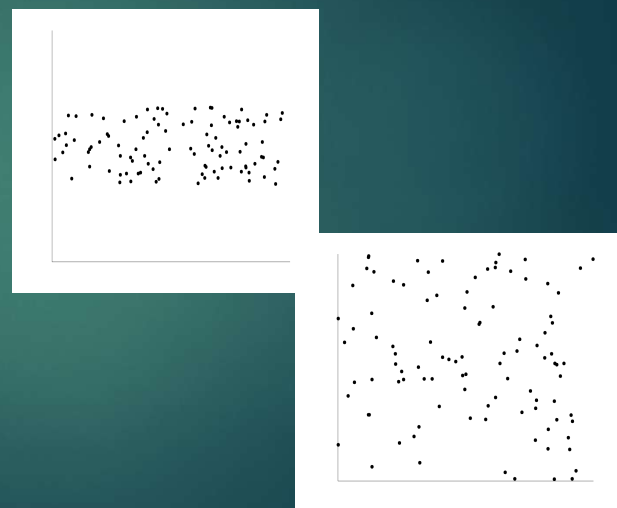
Correlated Data

23

2110773-2 2/67



Uncorrelated Data



Regression

24

2110773-2 2/67

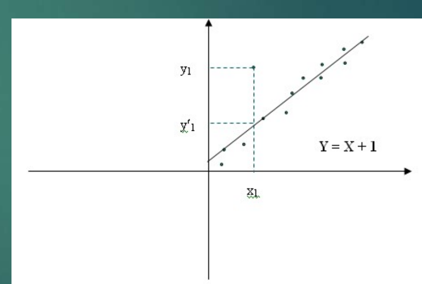
- Linear Regression

  $Y = \alpha + \beta X$

- Multiple Linear Regression

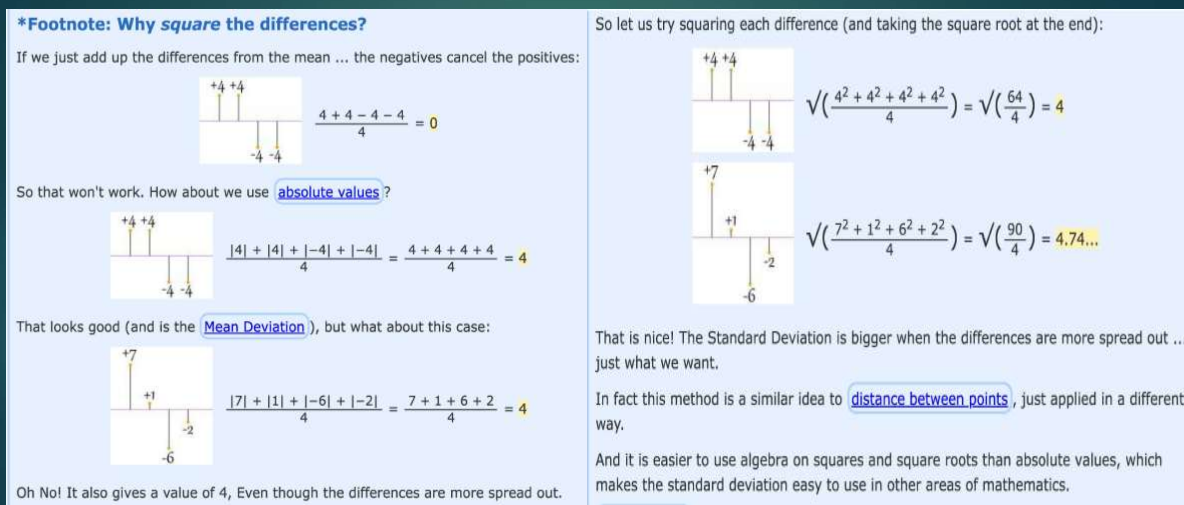
  $Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_m X_m$

- Smooth out noise
- Fill in missing value



25

2110773-2 2/67



***Footnote: Why *square* the differences?**

If we just add up the differences from the mean ... the negatives cancel the positives:

+4 +4 -4 -4

$$\frac{4+4-4-4}{4} = 0$$

So that won't work. How about we use absolute values?

+4 +4 -4 -4

$$\frac{|4| + |4| + |-4| + |-4|}{4} = \frac{4+4+4+4}{4} = 4$$

That looks good (and is the Mean Deviation), but what about this case:

+7 +1 -2 -6

$$\frac{|7| + |1| + |-6| + |-2|}{4} = \frac{7+1+6+2}{4} = 4$$

Oh No! It also gives a value of 4, Even though the differences are more spread out.

So let us try squaring each difference (and taking the square root at the end):

+4 +4 -4 -4

$$\sqrt{\left(\frac{4^2+4^2+4^2+4^2}{4}\right)} = \sqrt{\left(\frac{64}{4}\right)} = 4$$

+7 +1 -2 -6

$$\sqrt{\left(\frac{7^2+1^2+6^2+2^2}{4}\right)} = \sqrt{\left(\frac{90}{4}\right)} = 4.74...$$

That is nice! The Standard Deviation is bigger when the differences are more spread out ... just what we want.

In fact this method is a similar idea to distance between points, just applied in a different way.

And it is easier to use algebra on squares and square roots than absolute values, which makes the standard deviation easy to use in other areas of mathematics.

Data Integration

26

2110773-2 2/67

- Integration of multiple databases
- Handle data inconsistencies, majorly due to
  - Unit of Measure differences
  - Value differences
- Manage data redundancies
  - Correlation analysis

Data Transformation$_1$

27

* Many models implemented in Sklearn might perform poorly if the numeric features do not more or less follow a standard Gaussian (normal) distribution. Except for tree-based models, the objective function of Sklearn algorithms assumes the features follow a **normal distribution**.
* **Standardization** or **Scaling** numeric features is required for distance-based algorithms e.g. SVM, kNN to achieve better results
* Scaling and Normalization are very similar and confusing, sometimes used interchangeably
* what's the difference?

Data Transformation$_2$

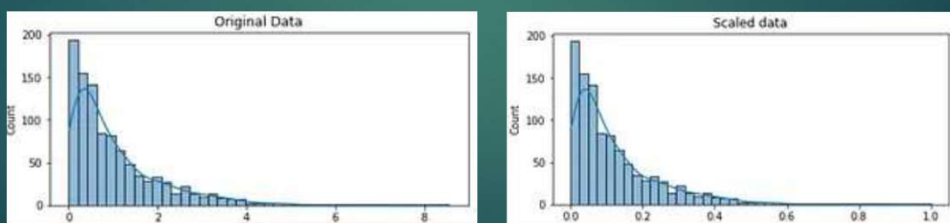
28

- Standardization (Scaling) / Normalization
  - **numeric variables** are transformed in both cases
    - Min-max Scaling using min and max values of distribution→ MinMaxScaler()
    - Z-score using variance and mean → StandardScaler()
    - Sigmoidal
    - Log Transforms →PowerTransformer()
- Data Type Conversion

https://towardsdatascience.com/how-to-differentiate-between-scaling-normalization-and-log-transformations-69873d365a94

Scaling

- Scaling is a method to standardize the range of independent variables or features of data.
- Change **range** of data to same scale, e.g. 0-1, 0-100
- Applied in distance-based algorithms, e.g. SVM, kNN
- Same importance for a change of "1" in any numeric features
- By scaling, variables are compared on equal footing



From Kaggle source

Scaling: case study

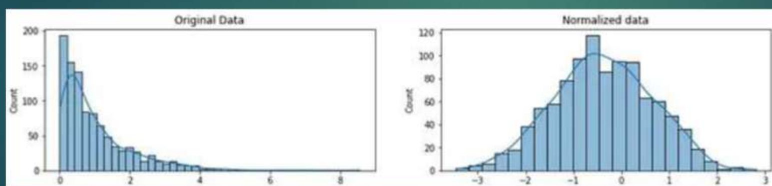
- Purpose: Change the values of numeric columns to a common scale
- Example: *age*(x1) ranges 0-100; *income*(x2) ranges 0-1,000,000
- Observing *income* will influence the result more due to its larger value
- Example of two deep neural network models w/ and w/o data scaling, accuracy = 88.93%, 48.80% respectively

Elevation	Aspect	Slope	Horizontal_D	Vertical_Dist	Horizontal_D	Hillshade_9a	Hillshade_N	Hillshade_3p	Horizontal_Distance_To_Fire_Points
2596	51	3	258	0	510	221	232	148	6279
2590	56	2	212	-6	390	220	235	151	6225
2804	139	9	268	65	3180	234	238	135	6121
2785	155	18	242	118	3090	238	238	122	6211
2595	45	2	153	-1	391	220	234	150	6172
2579	132	6	300	-15	67	230	237	140	6031
2606	45	7	270	5	633	222	225	138	6256
2605	49	4	234	7	573	222	230	144	6228
2617	45	9	240	56	666	223	221	133	6244
2612	59	10	247	11	636	228	219	124	6230
2612	201	4	180	51	735	218	243	161	6222
2886	151	11	371	26	5253	234	240	136	4051
2742	134	22	150	69	3215	248	224	92	6091
2609	214	7	150	46	771	213	247	170	6211
2503	157	4	67	4	674	224	240	151	5600
2495	51	7	42	2	752	224	225	137	5576
2610	259	1	120	-1	607	216	239	161	6096
2517	72	7	85	6	595	228	227	133	5607
2504	0	4	95	5	691	214	232	156	5572

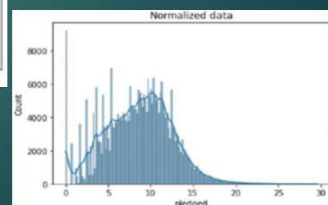
https://medium.com/@urvashilluniya/why-data-normalization-is-necessary-for-machine-learning-models-681b65a05029

Normalization

- Change **shape of distribution**
- Change the observations so that they can be described as **Normal** distribution, also known as **Gaussian** distribution
- Histogram and Boxplot can be used for identifying the underlying distribution of features



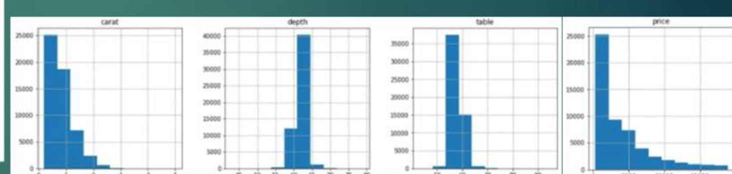
From Kaggle source

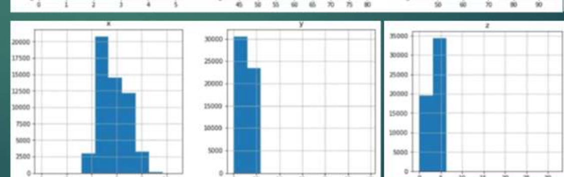


Box-Cox Transformation

Normalization- case study

	count	mean	std	min	25%	50%	75%	max
carat	53940.0	0.80	0.47	0.2	0.40	0.70	1.04	5.01
depth	53940.0	61.75	1.43	43.0	61.00	61.80	62.50	79.00
table	53940.0	57.46	2.23	43.0	56.00	57.00	59.00	95.00
price	53940.0	3932.80	3989.44	326.0	950.00	2401.00	5324.25	18823.00
x	53940.0	5.73	1.12	0.0	4.71	5.70	6.54	10.74
y	53940.0	5.73	1.14	0.0	4.72	5.71	6.54	58.90
z	53940.0	3.54	0.71	0.0	2.91	3.53	4.04	31.80

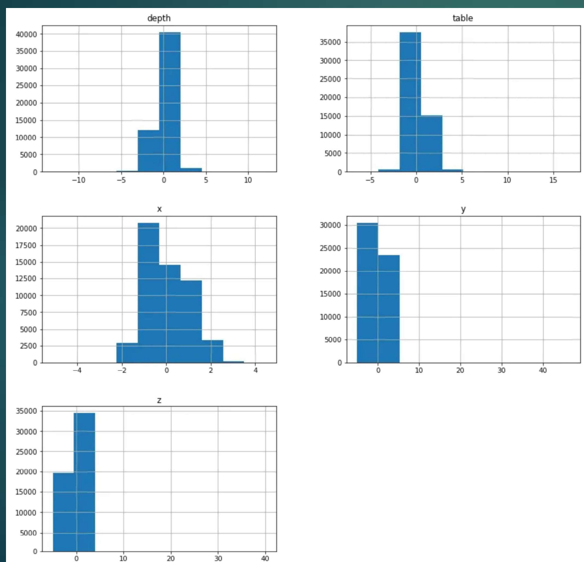




Normalization w/ StandardScaler()

33

2110773-2 2/67




```
>>> diamonds[to_scale].var()
depth    1.000019
table    1.000019
x        1.000019
y        1.000019
z        1.000019
dtype: float64

>>> diamonds[to_scale].mean().round(3)
depth   -0.0
table    0.0
x        0.0
y       -0.0
z       -0.0
dtype: float64
```

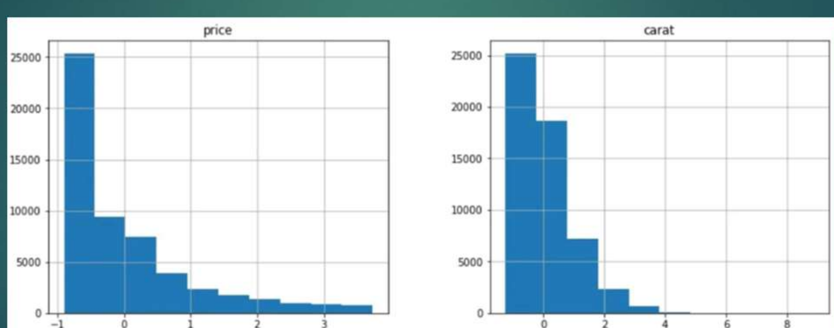
Depth and *x* now genuinely look like a Gaussian distribution. However, the features table, y, and z are still squished into the corner of their plots, suggesting the presence of outliers

Skewed Distribution w/ StandardScaler()

34

2110773-2 2/67

- When a feature does not follow a linear distribution, it would be unwise to use the mean and the standard deviation to scale it.
- To implement non-linear transformations, Sklearn offers a PowerTransformer() using logarithmic functions to support Box-Cox and Yeo-Johnson transform.



Log Transform

35

2110773-2 2/67

Normalization

- Base 2 — the base 2 logarithm of 8 is 3, because $2^3 = 8$
- Base 10 — the base 10 logarithm of 100 is 2, because $10^2 = 100$
- Natural Log — the base of the natural log is the mathematical constant "e" or Euler's number which is equal to 2.718282. So, the natural log of 7.389 is 2, because $e^2 = 7.389$.

Natural log transformation function of NumPy

```
import numpy as np

x = [1, 2, 3, 4, 5]
y = np.log(x)
y
```

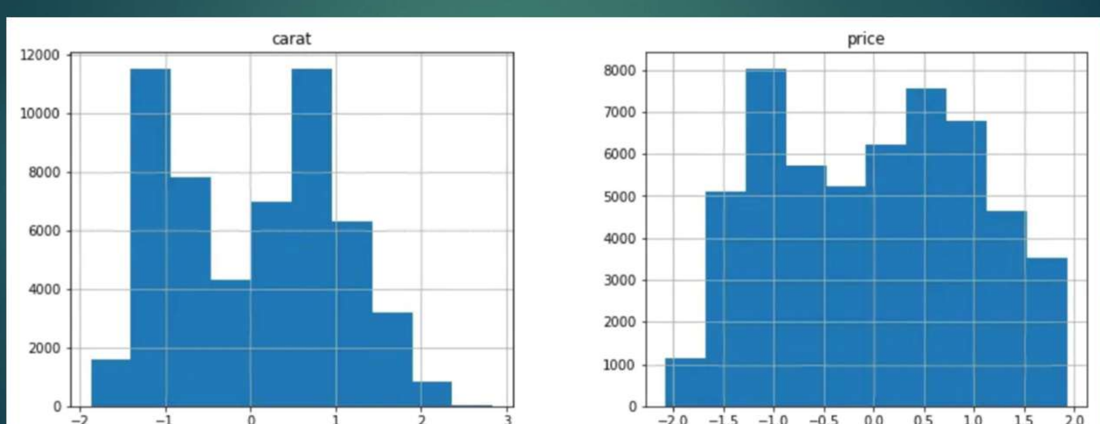
	Income	Age	Department	log_income
0	15000	25	HR	9.615805
1	1800	18	Legal	7.495542
2	120000	42	Marketing	11.695247
3	10000	51	Management	9.210340

Skewed Distribution w/ PowerTransformer()

36

2110773-2 2/67

- The new features look much better than the old skewed ones.



Min-Max Scaling

เป็นการแปลงข้อมูลเชิงเส้นจากช่วงที่เป็นไปได้เดิมของค่าอินพุต ให้เป็นช่วงข้อมูลใหม่ที่กำหนด ปกติคือช่วง [0-1]

กำหนดให้ v คือค่าคุณลักษณะเดิม; v' คือค่าคุณลักษณะใหม่

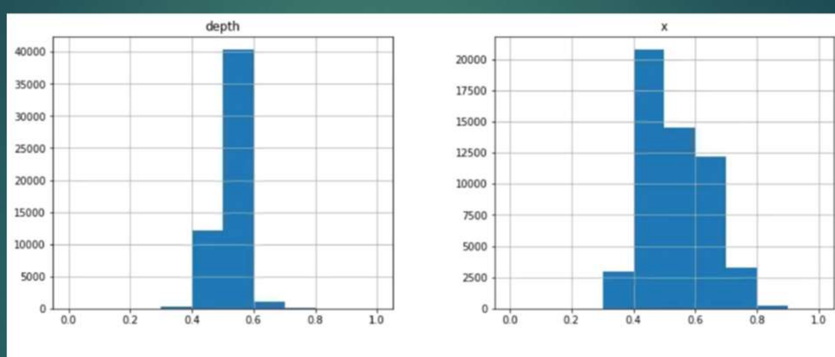
minA, maxA คือ ค่าต่ำสุดและสูงสุดเดิมของคุณลักษณะ A

new_minA, new_maxA คือ ค่าต่ำสุดและสูงสุดใหม่ของคุณลักษณะ A จะได้ว่า

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \quad \text{(สูตรที่ 1)}$$

Normalization w/ MinMaxScaler()

MinMaxScaler does not work well with features with outliers.



Even though it forces features to follow a normal distribution, the features won't have unit variance and a mean of 0:

StandardScaler/ PowerTransformer/ MinMaxScaler

- Scale data using StandardScaler(), a transformer used when we want a feature to follow a normal distribution with mean 0 and unit variance. Used most often with distributions without too many outliers.
- Log transform data using PowerTransformer(), a transformer used when we want a heavily skewed feature to be transformed into a normal distribution as close as possible.
- Normalize data using MinMaxScaler(), a transformer used when we want the feature values to lie within specific min and max values. It doesn't work well with many outliers and is prone to unexpected behaviors if values go out of the given range in the test set. It is a less popular alternative to scaling.

Data Scaling: Sigmoidal

- แปลงค่าอินพุตให้อยู่ในช่วง -1 ถึง 1 โดยใช้ฟังก์ชันซิกมอยด์ ซึ่งไม่ใช่ฟังก์ชันเชิงเส้น ข้อดีของวิธีนี้ คือ จะยังคงมีการเก็บรักษาค่าแปลกแยกไว้ การคำนวณหาค่าข้อมูลใหม่ y'

$$y' = \frac{1-e^{-\alpha}}{1+e^{-\alpha}}$$

โดยที่ $$\alpha = \frac{y - mean}{stddev}$$

Sigmoidal Normalization Example

Alpha	Y	Sig Y'			
0.24	45	0.12		Average	27.90
0.10	35	0.05			
0.58	70	0.28		Std-dev	72.08
0.57	69	0.28			
-0.08	22	-0.04			
-0.25	10	-0.12			
-0.32	5	-0.16			
-0.18	15	-0.09			
0.03	30	0.01			
3.78	300	0.96			
-0.10	21	-0.05			
-0.23	11	-0.12			
-0.19	14	-0.10			
-0.01	27	-0.01			
-1.08	-50	-0.49			
-1.43	-75	-0.61			
-0.36	2	-0.18			
-0.36	2	-0.18			
-0.36	2	-0.18			

Data Type Conversion: Label encoding

41

2110773-2 2/67

- CATEGORICAL → NUMERIC
- IN CASE THE ALGORITHM NEEDS NUMERICAL VALUES
- THE METHOD CAN BE PROBLEMATIC AS THE LEARNER MAY CONCLUDE THAT THERE IS AN ORDER. FOR EXAMPLE, AFRICA AND NORTH AMERICA DIFFER BY 4.

Label	Encoded Label
Africa	1
Asia	2
Europe	3
South America	4
North America	5
Other	6

Data Type Conversion: One hot encoding

42

2110773-2 2/67

- The encoding produces a sparse matrix (grid of numbers) w/ lots of zeroes (false values) and occasional ones (true values).

	is_africa	is_asia	is_europe	is_sam	is_nam
Africa	1	0	0	0	0
Asia	0	1	0	0	0
Europe	0	0	1	0	0
South America	0	0	0	1	0
North America	0	0	0	0	1
Other	0	0	0	0	0

Binning 1

43

2110773-2 2/67

- Data type conversion from numeric → categorical
- First **sort** data and **partition** into bins
- Label each bin w/ a symbol or value
- Given attribute values (for one attribute e.g., age):
  - 0, 4, 12, 16, 16, 18, 24, 26, 28
- Equi-width binning – for bin width of e.g., 10:
  - Bin 1: 0, 4 [-,10) bin
  - Bin 2: 12, 16, 16, 18 [10,20) bin
  - Bin 3: 24, 26, 28 [20,+) bin

  ** – to denote negative infinity, + for positive infinity
- Alternative Equi-width: Width = (Max – Min) / #intervals

Binning 2

44

2110773-2 2/67

- Equi-depth/ Equi-freuency:
  ใช้ความถี่ในการแบ่งข้อมูลออกเป็น N ช่วง
- Equi-frequency binning – for bin density of e.g., 3:
  - Bin 1: 0, 4, 12 [-,14) bin
  - Bin 2: 16, 16, 18 [14,21) bin
  - Bin 3: 24, 26, 28 [21,+] bin

- Given a list of product prices
  4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- แบ่งข้อมูลโดยวิธีแบ่งเป็นความลึกที่เท่ากัน
  Bin1: 4, 8, 9, 15; Bin2: 21, 21, 24, 25; Bin3: 26, 28, 29, 34
- ปรับเรียบโดยใช้ค่า bin means (ค่าเฉลี่ยของแต่ละบิน) :
  Bin 1: 9, 9, 9, 9; Bin 2: 23, 23, 23, 23; Bin 3: 29, 29, 29, 29
- ปรับเรียบโดยใช้ค่า bin boundaries (ค่าขอบของแต่ละบินที่ใกล้มากกว่า)
  Bin1: 4, 4, 4, 15; Bin2: 21, 21, 25, 25; Bin3: 26, 26, 26, 34

Note:

- Binning inevitably leads to loss of information, however, it reduces the chance of overfitting.
- Certainly, there will be improvements in speed and reduction of memory or storage requirements and redundancy.

Data Reduction

- Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Complex data analysis may take a very long time to run on the complete/ huge data set.

- Data Reduction Strategies
  - Data Aggregation
  - Dimensionality Reduction/ Feature selection
  - Numerosity Reduction
  - Discretization and Concept Hierarchy Generation

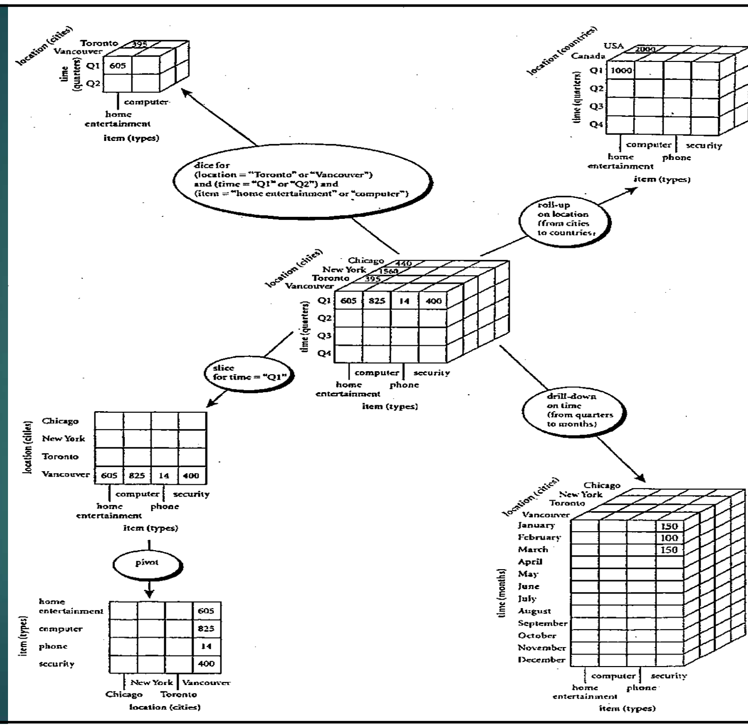
Data Aggregation

2110773-2 2/67

- การลดข้อมูลโดยใช้ค่าผลรวม
- Data Cube in Data Warehouse
- มิติข้อมูล คือ มุมมอง (perspective) ซึ่งองค์กรสนใจ ต้องการเก็บบันทึกข้อมูลไว้ เช่น เวลา สถานที่ ประเภท
- แทนที่จะเก็บข้อมูลดิบของรายการขายทั้งหมดที่เกิดขึ้น องค์กรจะลดปริมาณข้อมูลโดยจัดเก็บ***ข้อมูลรวมของยอดขาย***สำหรับแต่ละมิติที่น่าสนใจในโครงสร้างการจัดเก็บแบบลูกบาศก์ข้อมูล (data cube)

OLAP

Online Analytical Processing



47

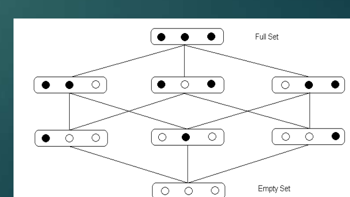
2110773-2 2/67

Dimensionality Reduction/ Feature Selection

48

2110773-2 2/67

- select *m* from *n* features, *m*≤ *n,* saving in search space
- suggestion:
  - ✓ remove *key/ ID* attribute
  - ✓ remove attributes with (too many) unique values
  - ✓ remove attributes with missing values > 50%
  - ✓ remove *irrelevant*, *redundant* features. Be cautious of removing relevant features as it is harmful.



Principal Component Analysis (PCA)

- Datasets typically contain a large number of features, but such high-dimensional feature spaces are not always helpful.
- In general, all the features are *not* equally important.
- Dimensionality reduction algorithms aim to reduce the dimension of the feature space to a fraction of the original number of dimensions.
- Principal Component Analysis (PCA) is linear dimensionality reduction technique.
- PCA is one of the most popular dimensionality reduction algorithms that takes advantage of existing correlations between the input variables in the dataset and combines those correlated variables into a new smaller set of uncorrelated variables called ***principal components***.
- PCA requires feature scaling if there is a significant difference in the scale between the features of the dataset.

Numerosity Reduction

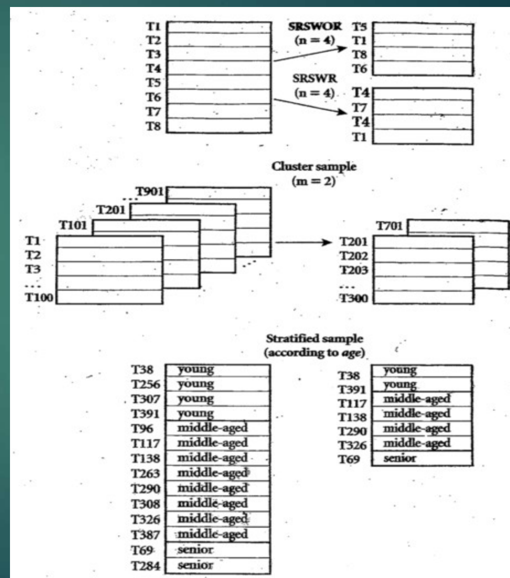
- Replace original data by smaller form of data representation
- ใช้เครื่องมือ เช่น แผนภาพฮิสโตแกรม (Histogram) หรือวิธีการจัดกลุ่ม (Clustering) ช่วยแสดงการกระจายของข้อมูล และใช้ค่าตัวแทนกลุ่มแทนค่าข้อมูลจริง
- หรืออาจใช้วิธีทางสถิติ เช่น การสุ่มตัวอย่าง (Sampling/ Instance selection) แทนการใช้ประชากรทั้งหมด

Instance Selection

Use portion rather the whole huge dataset

Sampling methods

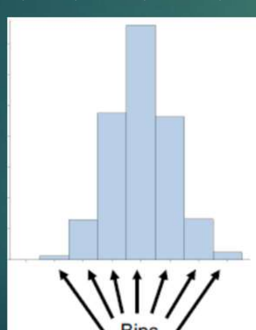
- Simple Random Sample Without Replacement (SRSWOR)
- Simple Random Sample With Replacement (SRSWR)
- Cluster Sample
- Stratified sampling treating each stratum as a population
  - Proportional allocation
  - Equal sample sizes



Discretization & Concept Hierarchy

Discretization

- Binning methods
  - Equi-width/ Equal-width
  - Equi-depth/ Equal-frequency



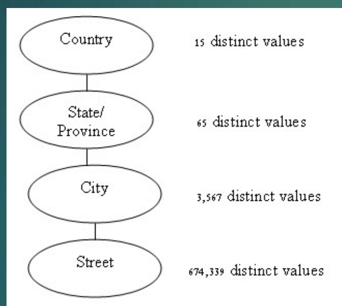
Concept Hierarchy

- การลดข้อมูลประเภท Categorical หรือข้อมูลที่ไม่ต่อเนื่องด้วยการสร้างเป็นลำดับชั้นมโนทัศน์ (concept hierarchy) โดยนิยามลำดับการเทียบ (mapping) กลุ่มมโนทัศน์ (concept) ระดับล่างไปสู่มโนทัศน์ในระดับสูงขึ้น
- Higher concept มีความเป็นทั่วไป (general) มากกว่า Lower concept

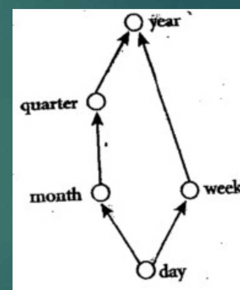
53

2110773-2 2/67

Schema hierarchy ส่วนมากเป็นความสัมพันธ์ระหว่างคุณลักษณะในฐานข้อมูล ซึ่งอาจเป็นความสัมพันธ์แบบ



Total Order



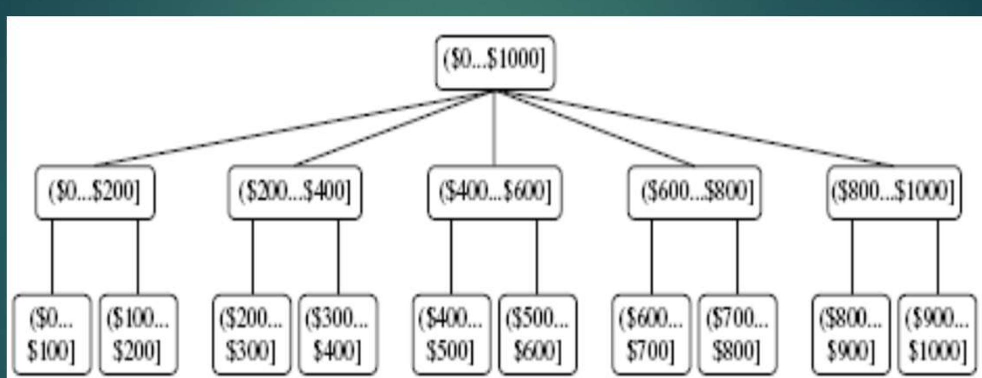
lattice

Partial Order

54

2110773-2 2/67

Set-grouping hierarchy การแบ่งค่าคุณลักษณะออกเป็นช่วงๆ และการแบ่งช่วงค่าสามารถทำต่อเป็นลำดับชั้น



Operation-derived Hierarchy

55

2110773-2 2/67

- การกำหนดลำดับชั้นมโนทัศน์ (concept hierarchy) จะขึ้นอยู่กับการใช้งานหรือการปฏิบัติงานของผู้ใช้/ ผู้เชี่ยวชาญ ตัวอย่างเช่น email address หรือ URL ของหน้าเว็บต่างๆ

Rule-based Hierarchy

56

2110773-2 2/67

- การกำหนดลำดับชั้นมโนทัศน์ อ้างอิงจากกฎชุดหนึ่ง ตัวอย่างเช่น
  กำหนดให้ P1 = retail price of X; P2 = actual cost of X
- lowProfitMargin(X) ← price(X, P1) and cost(X, P2) and (P1-P2) < $50
- mediumProfitMargin(X) ← price(X, P1) and cost(X, P2) and ((P1-P2) >= $50 and (P1-P2) <= $250)
- highProfitMargin(X) ← price(X, P1) and cost(X, P2) and (P1-P2) >$250

Data Preparation

57

2110773-2 2/67

1) Examining the Data Set

2) Feature Selection/ Narrowing down columns manually

- Remove Id or key
- Remove Irrelevant variables Feature understanding is extremely important! <u>Require Domain Expert</u>
- Remove Calculated fields
- Remove flat values

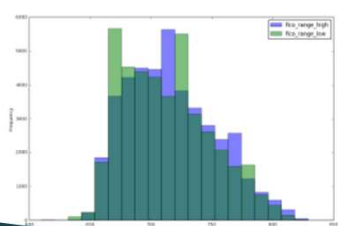
3) Preparing features

- Drop unqualified features
  - Variables with missing values > 50%
  - Too many unique values
- Handling missing values
- Investigate categorical features
  - Recode, consolidation (grouping)
  - Convert ordinal to numeric
  - Convert categorical to numeric
- Check all numeric variables
  - Truncate outliers
  - Feature Transformation

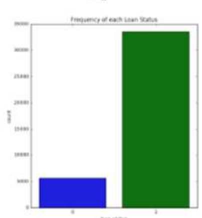
58

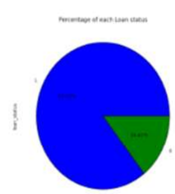
2110773-2 2/67

- Numerical variables
  - Out of ranges
  - Distribution: histogram



- Categorical variables
  - Miscodes
  - Distribution: frequency table, bar chart
- Target variable
  - Understand proportion of each class: bar chart, pie chart





Examining the Dataset