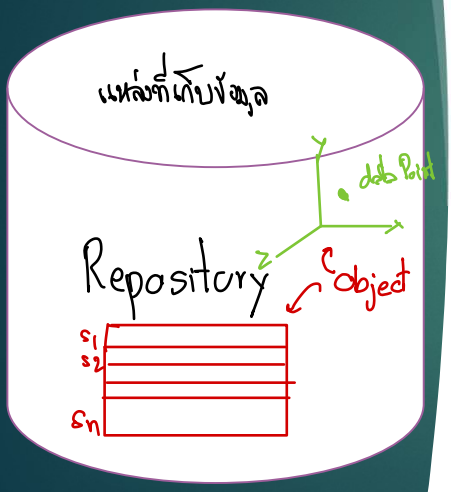
2110773 Data Mining Chapter2: Data Preprocessing

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

- รศ. ดร. ญาใจ ลิ่มปิยะกรณ์

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

- 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*

supervised — classification → classifier (categorical) / regression → regressor (numeric) → output model (predictive)

unsup — Associate / Clustering → output model / descriptive model

Attribute Data Types

1. *Qualitative/ Quantitative*
2. *Categorical/ Numeric*
3. *Discrete/ Continuous*

datum = เอกพจน์ data

- 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.

Subjective เอามาคำนวณไม่ได้ — objective คำนวณได้

Qualitative/ Categorical (Discrete)

- Ordinal
- Nominal
- Binary (แบ่งได้ 2 ประเภท)
  - Symmetric
  - Asymmetric

Quantitative/ Numeric (Continuous)

- Interval
- Ratio

data don't lie.

1 เท่า: 6°C, 41°F — ไม่ 2 เท่า: 0°C, 50°F

0°C ไม่ได้แปลว่าร้อนหรือไม่ร้อน ไม่มี true zero

C, F → Interval เลข 0 ไม่มี 0 แท้จริง

ค. หมวดหมู่ที่นับไม่ได้ → Quantify (IQ, อุณหภูมิ) → ตัวเลข (uom : unit of measure) C°, F° โดนไม่ต้องเติม ในเรื่องตัวบุคคล

2 เท่า (80 kg / 40 kg 176 lb / 88 lb) 2 เท่าเหมือนกัน ถึงเป็น Ratio

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°).

Kevin ล็อบอก no heat หรือๆ

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

ค่าที่เป็นเชิงสัญลักษณ์ เอามาคำนวณไม่ได้

ฟัน Gold, Silver ไม่รู้ "magnitude" (ขนาด) ว่าห่างกันเท่าไหร่

มีลำดับ

color = {red, green, ...}

bloodType = {A, B, O, AB}

gender = {M, F}

labTest = {+, -} +, - ความหมายต่างกัน

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.

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

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

Scales of Measurement

Data	Nominal	Ordinal	Interval	Ratio
Labeled การแทน	✓	✓	✓	✓
Order ลำดับ	✗	✓	✓	✓
Measurable Difference การคำนวณ	✗	✗	✓	✓
True Zero Starting Point	✗	✗	✗	✓

เพราะแทนด้วยตัวอักษร

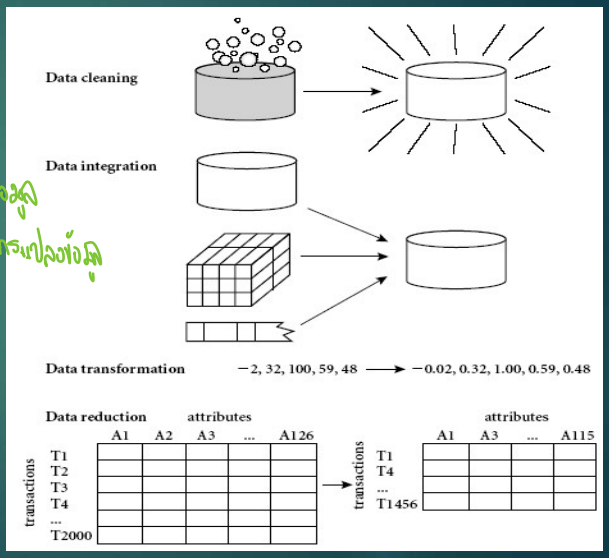
ตอนเราเก็บข้อมูลมาก ๆ เยอะมาก ๆ ก็ใช้สถิติมาแทนด้วย เช่น สถิติบอกว่าเป็น numeric

Survey

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

Data Preprocessing

- Data Cleaning
- Data Integration รวมหลายข้อมูล
- Data Transformation การแปลงข้อมูล
- Data Reduction ลดข้อมูล



Data Cleaning

- 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

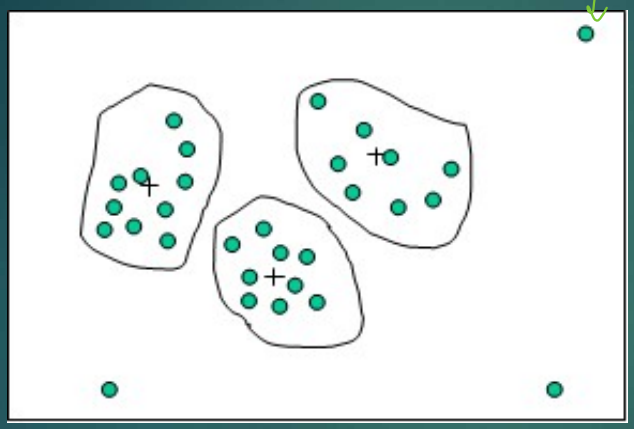
Missing Data

- 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)
    - Substitute 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

- 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

Normal Curve



อธิบายการจัดกลุ่มของข้อมูล

Clustering

Data Distribution

การกระจายของข้อมูล

15

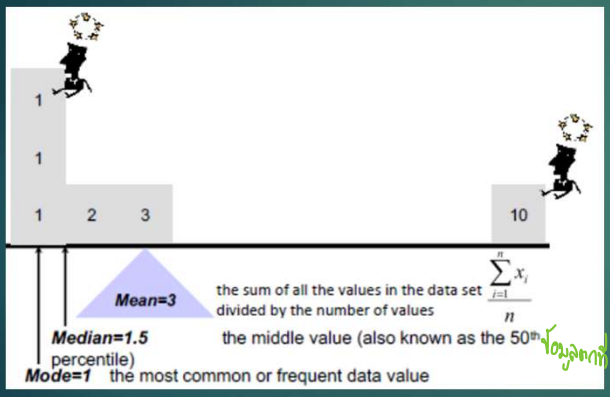
2110773-2 2/66

ข้อมูลอยู่กลางๆ ค่ากลาง เป็น bell curve

1. Central Tendency/ Center

ดูเข้าหา mean

2. Spread/ Dispersion การจัดกระจาย, กระเจิง



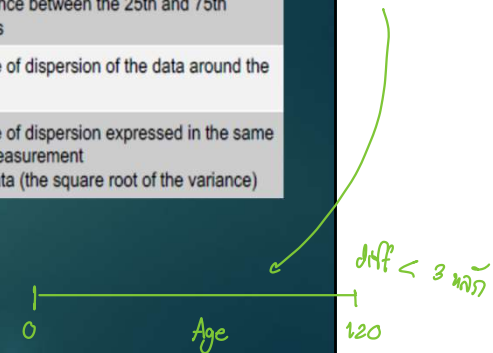
ข้อมูลตรงนี้ 50%

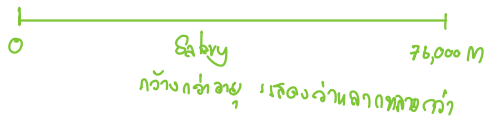
Measure	Definition
Range	the difference between the maximum and minimum data values
Interquartile Range	the difference between the 25th and 75th percentiles
Variance	a measure of dispersion of the data around the mean
Standard Deviation	a measure of dispersion expressed in the same units of measurement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

ตัวช่วยกว้าง แสดงว่าข้อมูลกระจายเยอะ

√variance

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





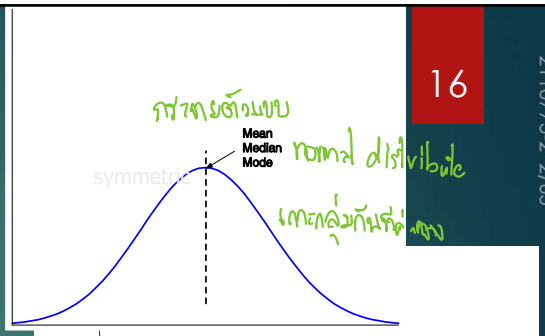
กว้างกว่าอายุ แสดงว่าหลากหลายกว่า

Symmetric vs. Skewed Data

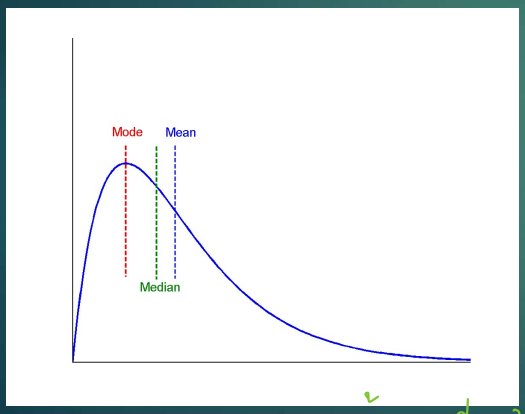
16

2110773-2 2/65

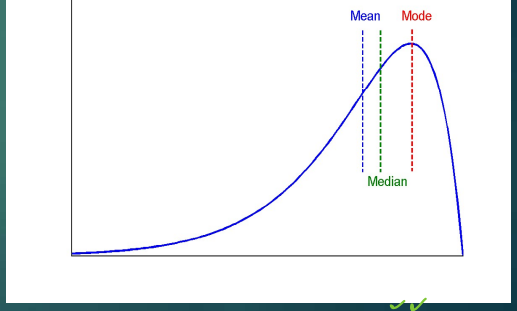
- Median, mean and mode of symmetric, positively and negatively skewed data



การกระจายตัวแบบ normal distribute เกาะกลุ่มกันที่ตรงกลาง



positively skewed เบ้ขวา (ดูที่หาง)



negatively skewed เบ้ซ้าย

เพราะข้อมูลเกาะทางซ้ายมาก

จึงต้องดึงไปทางขวาเพื่อเข้าสู่ normal distribution

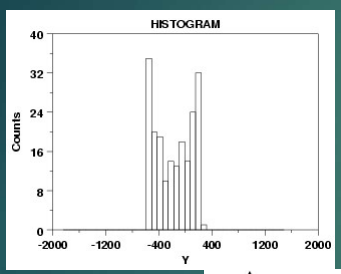
ตัวที่ใหญ่กว่าลู่สู่ต่ำกว่า

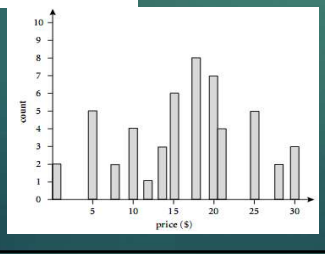
Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed) ถ้ามีการบิด	Median

When to use Mean, Median, Mode

Graphical Displays of Distribution

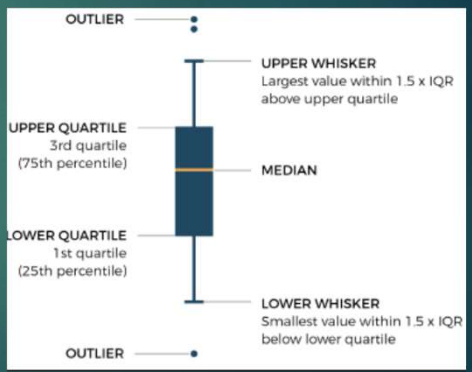
- Histogram Graph display of frequencies, shown as bars with numeric values on X axis



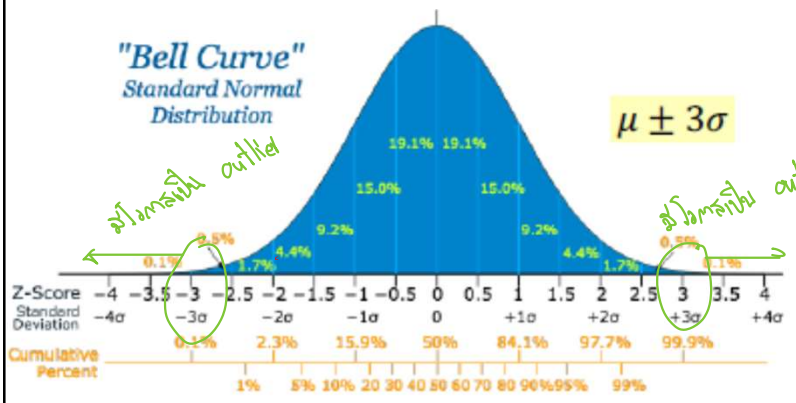


Singleton Histogram

- Box plots



19



Truncate Outliers: Bell Curve

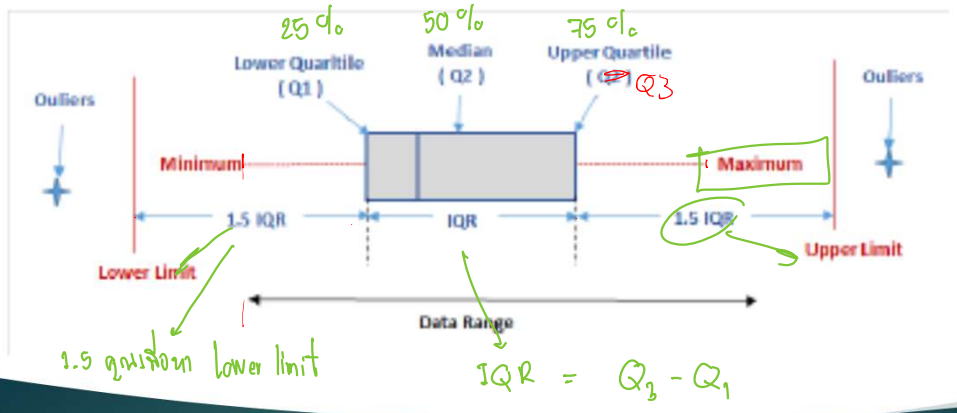
Variance and standard deviation (sample: s, population: σ)

Standard deviation is the square root of variance

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

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

20



Truncate Outliers: Box Plots

2110773-2 2/66

Interquartile Range : IQR

- **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

Odd set of numbers

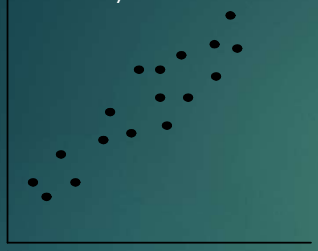
- Step 1: **Put the numbers in order.**
  1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.
- Step 2: **Find the median.**
  1, 2, 5, 6, 7, **9**, 12, 15, 18, 19, 27.
- Step 3: **Place parentheses around the numbers above and below the median.** Not necessary **statistically**, but it makes Q1 and Q3 easier to spot.
  (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.

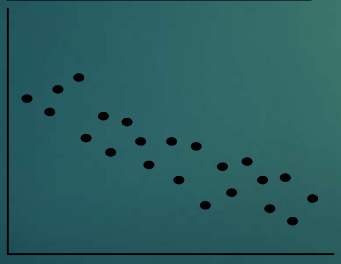
Even set of numbers

- 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

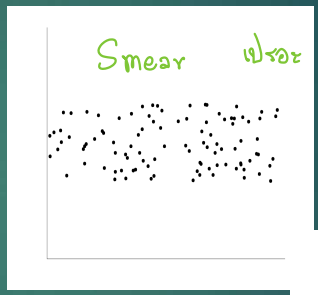
Positively

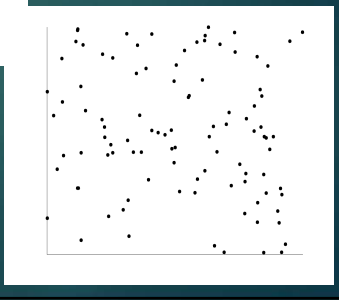




Negatively

Uncorrelated Data





Regression

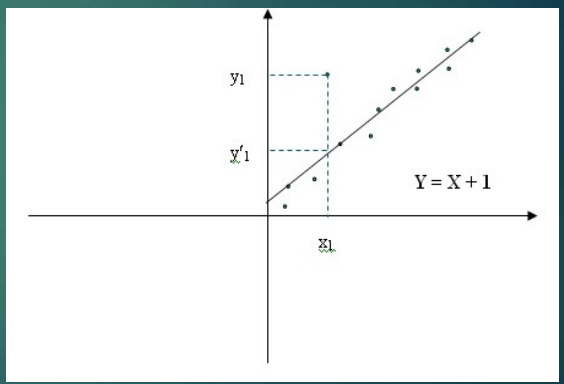
- Linear Regression

  $Y = \alpha + \beta X$

- Multiple Linear Regression

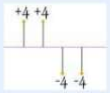
  $Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_m X_m$

- Smooth out noise
- Fill in missing value



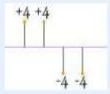
***Footnote: Why *square* the differences?**

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



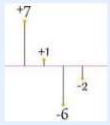
$$\frac{4 + 4 - 4 - 4}{4} = 0$$

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



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

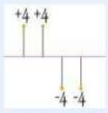
That looks good (and is the Mean Deviation), but what about this case:



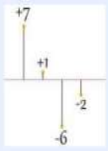
$$\frac{|7| + |1| + |-6| + |-2|}{4} = \frac{7 + 1 + 6 + 2}{4} = 4$$

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):



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



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

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

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

Data Transformation

- make the learning algorithms understanding easier, and achieving better results
  - Standardization (Scaling) / Normalization
    - Min-max Scaling
    - Log Transform
    - Z-score/ Zero-mean
    - Sigmoidal
  - Data Type Conversion

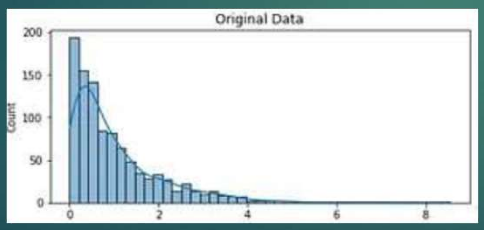
https://towardsdatascience.com/how-to-differentiate-between-scaling-normalization-and-log-transformations-69873d365a94

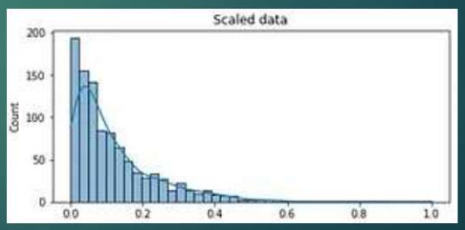
Scaling vs. Normalization

* very similar, confusing
* sometimes used interchangeably
* **numeric variables** are transformed in both cases
* what's the difference?

Scaling

- Numeric variables
- Change **range** of data 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 feature
- By scaling, variables are compared on equal footing





From Kaggle source

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

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

Scaling: case study

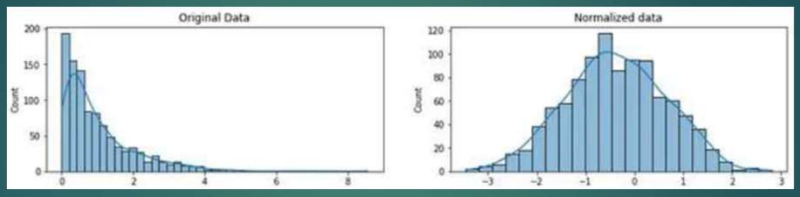
- Purpose: Change the values of numeric columns to a common scale
- Example: *age*(x1) ranges 0-100; *income*(x2) ranges 0-1,000,000
- Observing *income* will influence the result more due to its larger value
- Example of two deep neural network models w/ and w/o data scaling, accuracy = 88.93%, 48.80%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

Elevation	Aspect	Slope	Horizontal_D	Vertical_Dist	Horizontal_D	Hillshade_9a	Hillshade_Nc	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	5573

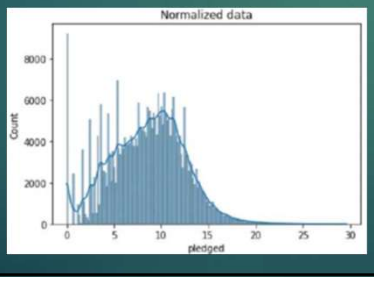
https://medium.com/@urvashilluniya/why-data-normalization-is-necessary-for-machine-learning-models-681b65a05029

Normalization

- Change **shape of distribution**
- Change the observations so that they can be described as **Normal** distribution, also known as **Gaussian** distribution



From Kaggle source



Box-Cox Transformation