Data Mining

It is a process of **extracting and discovering patterns** in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Companies use it to turn raw data into useful information.

Extracting knowledge from large amount of data:

- Cleaning, Integration, Selection, Transformation, Mining/Processing, Pattern Evaluation, Presentation

What kind of patterns can be mined/found by data mining?

- Characterization and Discrimination
- Frequent patterns, association, and correlations
- Classification and prediction
- Cluster analysis
- Outlier analysis
- Evolution analysis

Data Mining

Models

Predictive

Prediction of future

- Classification
- Regression
- Time Series Analysis
- Prediction

Descriptive

Description / Patterns of given data

- Clustering
- Summarization
- Association Rules
- Sequence Discovery

Knowledge Discovery in Database (KDD)

1. Selection
   a) Collection of data
2. Preprocessing
   a) Dear with incorrect / missing data
3. Transformation
   a) Common format and preprocessing
4. Data mining
   a) Algorithmic tools
5. Interpretation / Evaluation
   a) Presentation and visualization

Bayes' Theorem

Let A_1, A_2, ….,A_K be a collection of K mutually exclusive and exhaustive events with probability P (A_i) i = 1, 2, …., K

Then for any event B for which P(B) > 0,

$$P(A_j \cap B) = \frac{P(A_j \mid B)}{P(B)}$$

$$= \frac{P(B \mid A_j)\, P(Aj)}{\sum P(B \mid Aj)\, P(Aj)}$$

Statistics

Statistics is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data

Central Tendency: The central tendency is the descriptive summary of a data set. Through the single value from the dataset, it reflects the center of the data distribution. Moreover, it does not provide information regarding individual data from the dataset, where it gives a summary of the dataset. Generally, the central tendency of a dataset can be defined using some of the measures in statistics.

Central Tendency

- Mean
- Median
- Mode

$$mean = \frac{\sum x}{n}$$

mean [10, 15, 20, 26, 28] = 19.8

Median is the central value when data is in sorted order [10, 15, **20**, 26, 28] = 20

Median [10, 15, **20**, **26**, 28, 32] = (20 + 26) / 2 = **23**

Mode is the most frequent value [2, 5, 5, 2, 3, 2, 2, 2] = **2**

Mode is the most frequent value [2, 5, 5, 2, 3, 2, 2, 5, 2, 5, 5] = **2** or **5**

Pearson Mode Skewness

- If the mean is greater than mode, the distribution is positively skewed
- If the mean is less than mode, the distribution is negatively skewed
- If the mean is greater than median, the distribution is positively skewed
- If the mean is less than median, the distribution is negatively skewed

When mean and median are equal, we call it a symmetric dataset

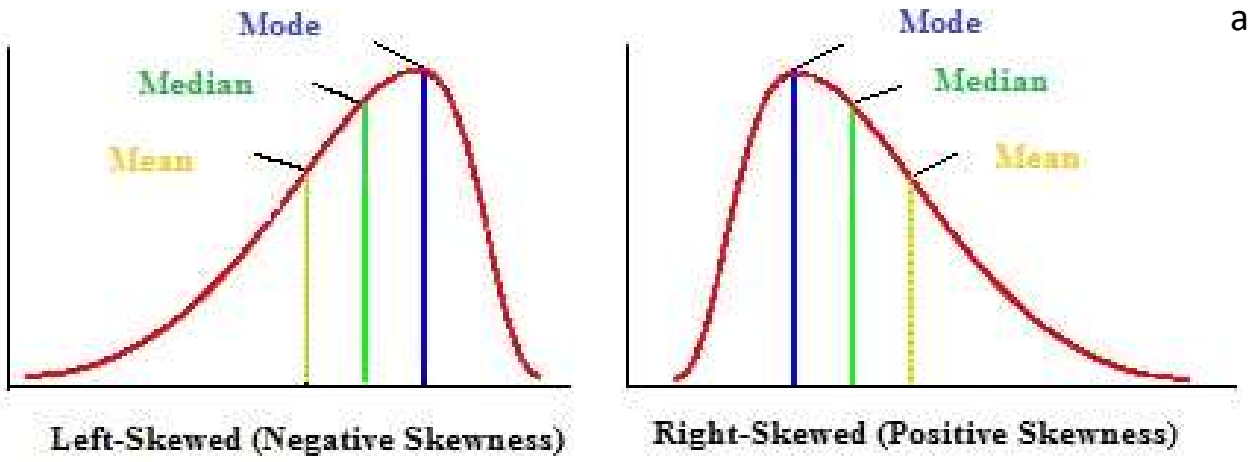


Image Source: https://www.statisticshowto.com/pearson-mode-skewness/

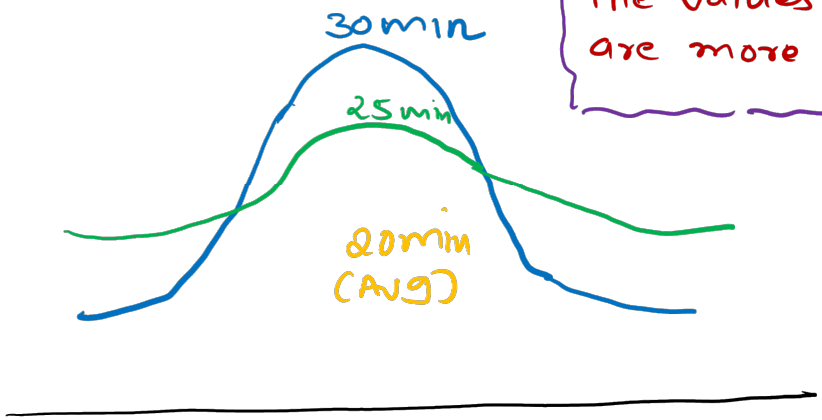
Measures of Variability

It refers to how "spread out" a group of values is.

When a distribution has lower variability, the values in a dataset are more consistent.

Example :- Pizza Delivery Time

	Shop S1	Shop S2
Avg Time	20 min	20 min
During Peak Traffic	25 min	30 min



Quiz 1

No. of students

2	6	5	4	3
5	6	7	8	9

Marks

$\frac{5+6+7+8+9}{5} = 7$ (Mean)

Quiz 2

2	4	3	3	2	3	3
4	5	6	7	8	9	10

$\frac{4+5+6+7+8+9+10}{7} = 7$ (Mean)

Measures of Variability

It refers to how "spread out" a group of values is.

- Range
- Variance and standard deviation
- Coefficient of variation
- Interquartile Range (IQR)

Range: Largest value – Smallest Value [10, 25, 12, 13, 15, 16, 23]. Range 25−10 = **15**

Variance: How far a set of numbers is spread out from their average values [δ^2]

Standard Deviation: Standard or typical difference between each data point and the mean [δ]

Population Standard Deviation (SD): $\sqrt{\frac{\sum(X-\bar{X})^2}{N}}$

Standard Deviation

www.hbpatel.in

Data: [2, 7, 3, 12, 9]

Mean = [2+7+3+12+9] / 5 = 6.6

$(2-6.6)^2 + (7-6.6)^2 + (3-6.6)^2 + (12-6.6)^2 + (9-6.6)^2$ = 21.16 + 0.16 + 12.96 + 29.16 + 5.76 = 69.2

Variance = δ^2 = 69.2 / 5 = 13.84

Standard Deviation = $\delta = \sqrt{13.84}$ = 3.72

	A	B	C
1		**Data Set 1**	**Data Set 2**
2		10	100
3		12	19
4		15	150
5		8	800
6		20	659
7		18	11
8		25	239
9		13	789
10		19	1000
11	**Variance (Popuated)**	26.02	134467.36
12	**Variance (Standard)**	29.28	151275.78
13	**Standard Deviation (P)**	5.10	366.70
14	**Standard Deviation (S)**	5.41	388.94
15			

```
=VAR.P(B2:B10)
=VAR.S(B2:B10)
=STDEV.P(B2:B10)
=STDEV.S(B2:B10)
```

Standard Deviation

Population Standard Deviation (for all values)= $\sqrt{\frac{\sum (x-\mu)^2}{N}}$, μ = Population Mean

Sample Standard Deviation (for sample/selected values) = $\sqrt{\frac{\sum (x-\bar{x})^2}{N-1}}$ $\bar{x}$ = Population Mean

	A	B	C
1		**Data Set 1**	**Data Set 2**
2		10	100
3		12	19
4		15	150
5		8	800
6		20	659
7		18	11
8		25	239
9		13	789
10		19	1000
11	**Variance (Popuated)**	26.02	134467.36
12	**Variance (Standard)**	29.28	151275.78
13	**Standard Deviation (P)**	5.10	366.70
14	**Standard Deviation (S)**	5.41	388.94

```
=VAR.P(B2:B10)
=VAR.S(B2:B10)
=STDEV.P(B2:B10)
=STDEV.S(B2:B10)
```

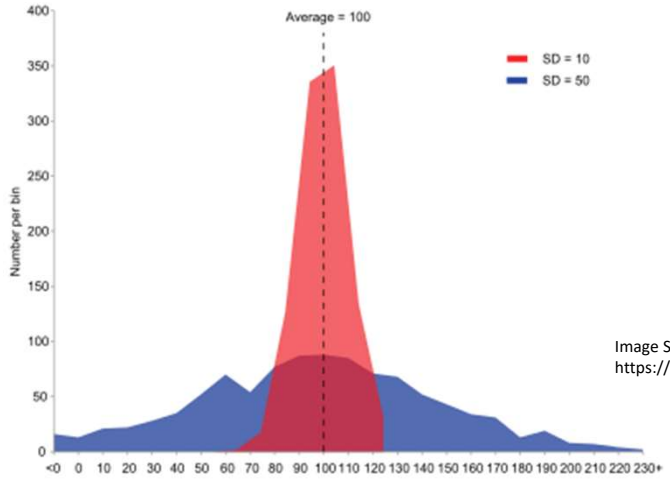


Image Source:
https://en.wikipedia.org/wiki/Variance

Example of samples from two populations with the same mean but different variances. The red population has mean 100 and variance 100 (SD=10) while the blue population has mean 100 and variance 2500 (SD=50).

Coefficient of Variation

- To compare two datasets (Relative Standard Deviation)

Coefficient of variation = $\frac{Standard\ Deviation}{Mean}$ x 100

	X	(X-μ)	$(X-\mu)^2$
	7	2	4
	2	-3	9
	7	2	4
	6	1	1
	5	0	0
	6	1	1
	2	-3	9
Total	35		28
Average μ	5		

Population Standard Deviation = $\sqrt{\frac{\sum (x-\mu)^2}{N}} = \sqrt{\frac{28}{7}} = 2$

Sample Standard Deviation = $\sqrt{\frac{\sum (x-\mu)^2}{N-1}} = \sqrt{\frac{28}{6}} = 2.16$

Coefficient of variation = $\frac{Standard\ Deviation}{Mean}$ x 100

= $\frac{2}{5}$ x 100 = 40%

Sample Coefficient of variation = $\frac{Populatio\ Standard\ Deviation}{Mean}$ x 100 = $\frac{2.16}{5}$ x 100 = 43.2%

Measures of Variability

- Interquartile Range (IQR)

The interquartile range defines the difference between the third and the first quartile. Quartiles are the partitioned values that divide the whole series into 4 equal parts. So, there are 3 quartiles. First Quartile is denoted by Q1 known as the lower quartile, the second Quartile is denoted by Q2 and the third Quartile is denoted by Q3 known as the upper quartile. Therefore, the interquartile range is equal to the upper quartile minus lower quartile. The difference between the upper and lower quartile is known as the interquartile range.

Interquartile range = Upper Quartile – Lower Quartile
= Q3 – Q1

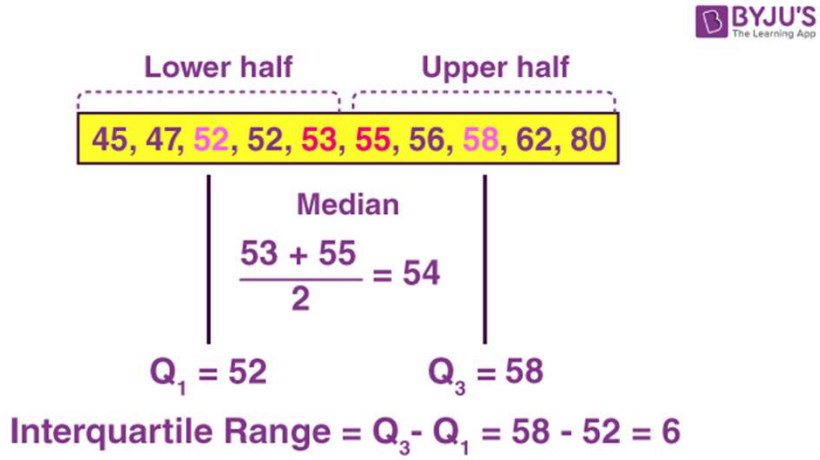


Image Source: https://byjus.com/

Measures of Variability: Take away

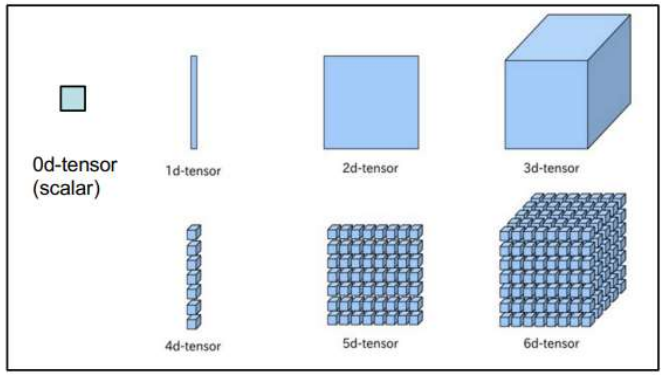
- Symmetric data with no serious outliers then range and SQ
- Skewed data and/or have serious outliers then IQR
- Comparison across two datasets, then coefficient of variation

Linear Algebra for Machine Learning

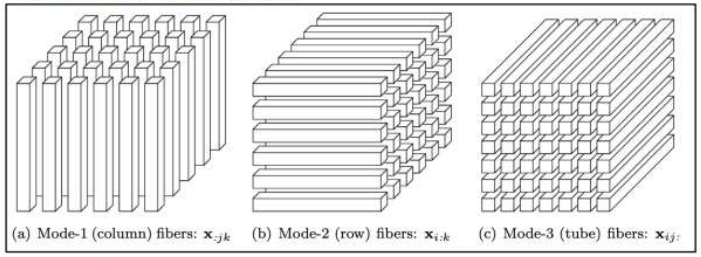
- Linear Algebra is a branch of mathematics concerning linear equations such as $a_1x_1 + a_2x_2 + \ldots + a_nx_n = b$
- Linear Algebra is fundamental to geometry, for defining objects such as lines, planes and rotations
- Linear Algebra is based on continuous math rather than discrete math
- Linear Algebra is essential for understanding of ML algorithms
- Topics
  - Scalars, Vectors, Matrices, Tensors
  - Linear dependence, Span
  - Norms…..

- Scalar
  - Single number (unlike arrays of numbers)
  - x Є R (Real value), n Є N (natural value)
- Vector
  - Array of numbers arranged in orders
- Matrix
  - 2D arrays of numbers
- Tensors
  - Array with more than two axes (E.g. RGB color has three axes)
  - A tensor is an array of numbers arranged on regular grid with variable number of axes

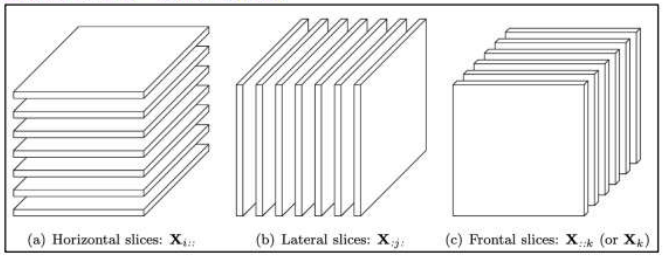
Linear Algebra for Machine Learning



Fibers of a 3rd order tensor



Slices of a 3rd order tensor



1	5	2	7	11	24	25	12

One dimensional Tensor

Collection of one dimensional tensors gives two dimensional tensor

1	5	2	7	11	24	25	12
2	3	35	7	14	0	2	15
5	25	3	1	13	28	3	16

Two dimensional tensor

Collection of two dimensional tensors gives two dimensional tensor

1	5	2	7	11	24	25	12
2	3	35	7	14	0	2	15
5	25	3	1	13	28	3	16
1	5	2	7	11	24	25	12
2	3	35	7	14	0	2	15
5	25	3	1	13	28	3	16

Three dimensional tensor

Image Source: https://cedar.buffalo.edu/~srihari/CSE676/2%20LinearAlgebra.pdf