Day 3

27 July 2024 19:19

Data types in statistics

Data table

In the context of machine learning, a data table, also known as a dataset or data matrix, refers to a structured arrangement of data

rows typically represent individual instances or observations,

columns represent attributes, features, or variables associated with those instances.

A data table is a fundamental building block in machine learning

example

columns represent attributes such as age, gender, height, weight

each row represents a person

ID	Age	Gender	Height	Weight	Class
1	25	Male	178	75	A
2	30	Female	162	58	B
3	22	Male	185	82	A
4	28	Female	170	63	B

class label indicating the group the person belongs to

names

Record

Columns

Record, samples, point, case, entity, instance, entry, Objects
Data points, Document, tuple, Transaction, feature vector

Attributes, Features, Variables
Field, Predictors, characteristics

Data science a study of 3 or 4 different disciplines. Hence there are a lot of vocab, often many for the same term!

FEATURE TYPES

- independent variable,
  - sometimes called an experimental or predictor variable,
  - is a variable that is being manipulated in an experiment in order to observe the effect …
- dependent variable, sometimes called an outcome/response/target variable.

Example

- Dependent Variable: Test Mark (measured from 0 to 100)
- Independent Variables:
  - Revision time (measured in hours),
  - Intelligence (measured using IQ score)

Sample dataset (temp forecasts)

year	month	day	week	Temp 2 days before	Temp 1 day before	Average temp	Actual temp on that day	Temp forecast by noaa	Temp Forecast by acc	Temp forecast by friend
2016	1	1	Fri	45	45	45.6	45	43	50	29
2016	1	2	Sat	44	45	45.7	44	41	50	61
2016	1	3	Sun	45	44	45.8	41	43	46	56
2016	1	4	Mon	44	41	45.9	40	44	48	53
2016	1	5	Tues	41	40	46	44	46	46	41
2016	1	6	Wed	40	44	46.1	51	43	49	40
2016	1	7	Thurs	44	51	46.2	45	45	49	38
2016	1	8	Fri	51	45	46.3	48	43	47	34
2016	1	9	Sat	45	48	46.4	50	46	50	47
...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...
2016	1	19	Tues	50	54	47.6	48	47	49	53
2016	1	20	Wed	54	48	47.7	52	44	52	61
2016	1	21	Thurs	48	52	47.8	52	43	51	57

Independent variables

Independent variables

Dependent variable

SAMPLE DATASET (automobiles)

KMs per liter	cylinders	displacement	horsepower	weight	acceleration	year	origin	Model name
18	8	307	130	3504	12	70	1	chevrolet chevelle malibu
15	8	350	165	3693	11.5	70	1	buick skylark 320
18	8	318	150	3436	11	70	1	plymouth satellite
16	8	304	150	3433	12	70	1	amc rebel sst
17	8	302	140	3449	10.5	70	1	ford torino
15	8	429	198	4341	10	70	1	ford galaxie 500
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
24	4	113	95	2372	15	70	3	toyota corona mark ii
22	6	198	95	2833	15.5	70	1	plymouth duster
18	6	199	97	2774	15.5	70	1	amc hornet
21	6	200	85	2587	16	70	1	ford maverick
27	4	97	88	2130	14.5	70	3	datsun pl510

Dependent variable ?

Sample data - jobs

Month	Total Filled Jobs
2004M07	1795610
2004M08	1792770
2004M09	1809590
2004M10	1815580
2004M11	1856360
2005M04	1871630
2005M05	1867870
2005M06	1857260
2005M07	1858360
2005M08	1856320
2005M09	1876270
2005M10	1866920
...	...
2011M10	1903630
2011M11	1940200
2011M12	1983070
2012M01	1865540
2012M02	1932380

Independent variables ?

Dependent variable ?

Sample data - text

Tweet d	Airline sentiment	Retweet count	text	tweet_location
570306133677760000	neutral	0	=@VirginAmerica What @dhepburn said.	
570301130888122000	positive	0	@VirginAmerica plus you've added commercials to the experience... tacky.	
570301083672813000	neutral	0	@VirginAmerica I didn't today... Must mean I need to take another trip!	Lets Play
570301031407624000	negative	0	@VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces & they have little recourse	
....	....	....	....	....
570300767074181000	negative	0	@VirginAmerica seriously would pay $30 a flight for seats that didn't have this playing. it's really the only bad thing about flying VA	
570300616901320000	positive	0	@VirginAmerica yes, nearly every time I fly VX this â€œear wormâ€[?]wonâ€™t go away :)	San Francisco CA
570300248553349000	neutral	0	@VirginAmerica Really missed a prime opportunity for Men Without Hats parody, there. https://t.co/mWpG7grEZP	Los Angeles
570285904809598000	positive	0	@VirginAmerica Thanks!	San Francisco, CA
570282469121007000	negative	0	=@VirginAmerica SFO-PDX schedule is still MIA.	palo alto, ca
570277724385734000	positive	0	@VirginAmerica So excited for my first cross country flight LAX to MCO I've heard nothing but great things about Virgin America. #29DaysToGo	west covina
570276917301137000	negative	0	@VirginAmerica I flew from NYC to SFO last week and couldn't fully sit in my seat due to two large gentleman on either side of me. HELP!	this place called NYC
570270684619923000	positive	0	I â[?]ï¸[?]flying @VirginAmerica. â˜ºï¸[?]Ÿ'[?]	Somewhere celebrating life.
570267956648792000	positive	0	@VirginAmerica you know what would be amazingly awesome? BOS-FLL PLEASE!!!!!!! I want to fly with only you.	Boston | Waltham
570265883513384000	negative	0	=@VirginAmerica why are your first fares in May over three times more than other carriers when all seats are available to select???	

Independent variables ?

Dependent variable ?

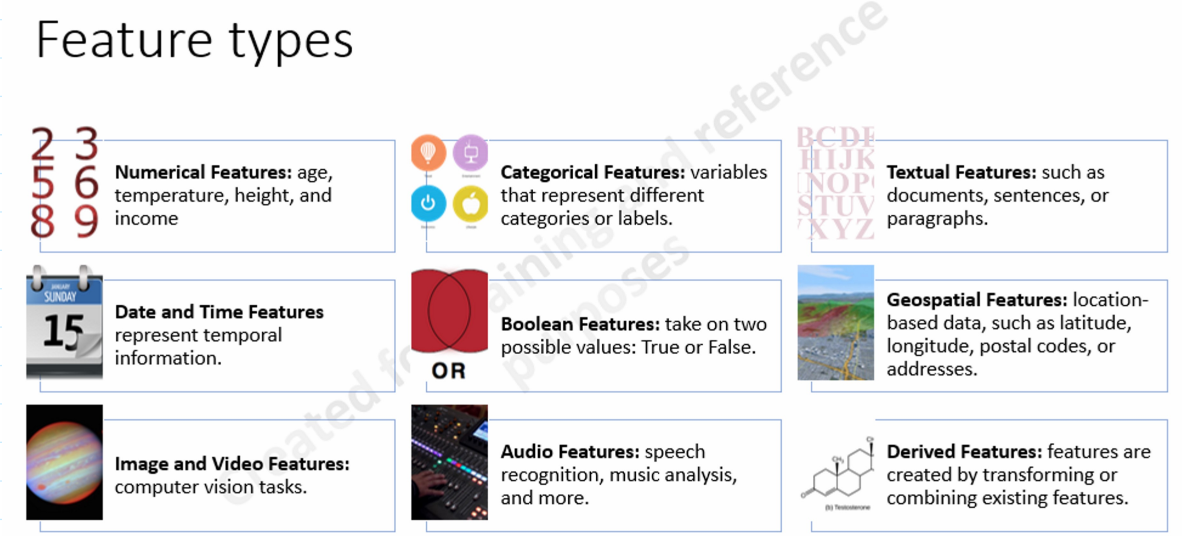
04-08-2024

Stock prices

10 - last 3 days prices = 12 , 9, 9

12
9
9
11
23
22

Feature types (different)



Numerical variables

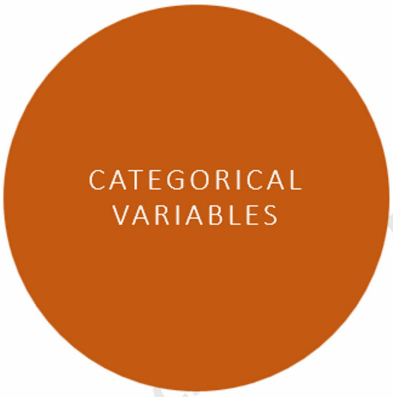
Continuous Numerical Variables:

- **Definition:** can take an infinite number of possible values within a given range.
- **Examples:**
  - **Height**: The height of a person can be any value within a certain range and is not restricted to specific discrete values.
  - **Temperature**: Temperature can be measured with great precision, allowing for an infinite number of possible values.

Discrete Numerical Variables:

- **Definition:** can only take specific, separate values, often integers.
- **Examples:**
  - Number of Siblings: The count of siblings is a discrete numerical variable, as it can only be a whole number.
  - Number of Cars in a Parking Lot: The count of cars is discrete, and you can't have a fraction of a car.

- a type of numerical variable that can only take specific integer values within a certain range.
- values are often counted, and there are gaps between them.
- can't be subdivided further.
- Examples of discrete variables:
  - The number of students in a classroom
  - The number of cars in a parking lot
  - The number of items purchased

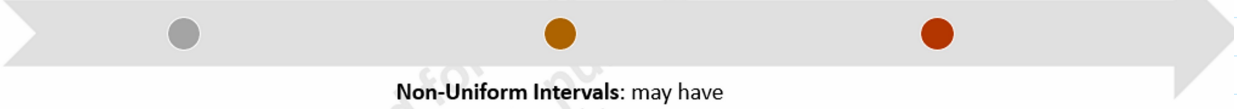


- also known as a **qualitative** variable, represents data in categories or labels that have no inherent numerical order or ranking.
- categories are usually distinct and don't have mathematical operations applied to them.
- characteristics or attributes that fall into different groups.
- **Examples** of categorical variables:
  - Colors (red, blue, green, etc.).
  - Types of fruits (apple, banana, orange, etc.).
  - Payment methods (credit card, cash, PayPal).

Ordinal variables - a type of categorical variable

Ordering: have a meaningful and logical order. Indicates greater or lesser than another, but the exact magnitude of the difference between categories may not be known.

Limited Arithmetic Operations: Ordinal variables can be ranked, but mathematical operations like addition, subtraction, multiplication, or division might not be meaningful or applicable.



Non-Uniform Intervals: may have uneven intervals between categories.

Examples of ordinal variables:

- Education Levels: (High School, Associate's, Bachelor's, Master's, PhD)
- Socioeconomic Status: (Low, Middle, High)
- Customer Satisfaction Ratings: (Poor, Fair, Good, Very Good, Excellent)
- Pain Intensity Levels: (Mild, Moderate, Severe)

Nominal variables - a type of categorical variable

Ordering: no inherent order or ranking between them.

No Arithmetic Operations: do not support arithmetic operations like addition, subtraction, multiplication, or division, as the categories have no numerical meaning.

Non-Uniform Intervals: may have uneven intervals between categories.

Examples of nominal variables:

- Colors: (Red, Blue, Green, etc.)
- Types of Fruits: (Apple, Banana, Orange, etc.)
- Payment Methods: (Credit Card, Cash, PayPal)
- Countries: (USA, Canada, France, etc.)

binary

has only two categories or levels.

categories are often represented as "1" and "0," "Yes" and "No," or "True" and "False."

Examples of dichotomous variables:

- Gender: (Male, Female)
- Smoker: (Yes, No)
- Voter: (Voted, Did Not Vote)
- Married: (Married, Not Married)

Measures of central tendency

WHAT IS CENTRAL TENDENCY

refers to the measure that represents the central or typical value of a dataset.

helps provide a single value around which the data tends to cluster.

making it easier to understand the general characteristics of the dataset.

MEASURES OF CENTRAL TENDENCY:

Mean: The average value of a set of data points.

Median: The middle value when data points are arranged in order. It's less affected by extreme values compared to the mean.

Mode: The value that occurs most frequently in a dataset.

MEAN OR MEDIAN

The **median** is less sensitive to outliers (extreme scores) than the **mean** and thus a better measure than the mean for highly skewed distributions, e.g. family income.

For example mean of 20, 30, 40, and 990 is (20+30+40+990)/4 =270.

The median of these four observations is (30+40)/2 =35.

Here 3 observations out of 4 lie between 20-40.

So, the mean 270 really fails to give a realistic picture of the major part of the data. It is influenced by extreme value 990.

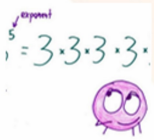
Types of mean

GEOMETRIC MEAN

Intuition

The geometric mean is used to find the central tendency of a set of values that are related multiplicatively, rather than additively.

useful when dealing with quantities that change or grow exponentially.



"Values that are related multiplicatively"

the growth or change in these values is better represented by multiplying them together rather than adding them.

This is often the case when dealing with quantities that experience exponential growth or decay.

EXAMPLES

Compound Interest

- investment that earns a fixed interest rate over several years. The value of the investment at each year is obtained by multiplying the previous year's value by (1 + interest rate).
  - Year 1: Initial value * (1 + interest rate)
  - Year 2: Year 1 value * (1 + interest rate)
  - Year 3: Year 2 value * (1 + interest rate)
  - ... and so on

Population Growth

Imagine a population of bacteria that doubles in size every hour.

Hour 1: Initial bacteria count * 2

- Hour 2: Hour 1 bacteria count * 2
- Hour 3: Hour 2 bacteria count * 2
- ... and so on

FORMULA (GEOMETRIC MEAN)

Geometric Mean = (x1 * x2 * x3 * ... * xn)^(1/n)

- Where:
  - x1, x2, x3, ..., xn are the individual numbers in the set.
  - n is the total number of values in the set.

multiply all the numbers together and then take the nth root of the product, where n is the number of values in the set.

KEY POINTS ABOUT THE GEOMETRIC MEAN

Multiplicative Relationships: is suitable for data where the values represent ratios, growth rates, or multiplicative factors.

Non-Negative Values: It doesn't work well with datasets that include negative values or zero.

Sensitivity to Small Values: A single very small value can significantly reduce the geometric mean.

Use Cases: calculating average investment returns, the average annual growth rate of populations or financial assets, the mean of values that represent relative change or ratios.

HOW TO BE SURE IF THE DATA IS EXPONENTIAL FORM

Visual Inspection: Exponential data often shows a characteristic curve that rises rapidly (for growth) or declines smoothly (for decay).

Logarithmic Transformation: Transform the data using a logarithm (e.g., natural logarithm) and examine if the transformed data appears more linear.

Examine Histograms: Plot histograms of your data. An exponential distribution often shows a long tail on one side. The histogram might resemble a decreasing exponential curve.

Quantile-Quantile (Q-Q) Plot: Create a Q-Q plot by plotting the quantiles of your data against the quantiles of an exponential distribution. If the points fall approximately on a straight line, your data might be exponential.

HARMONIC MEAN

useful in dealing with rates, ratios, or inversely proportional quantities. It's calculated using the formula:

Harmonic Mean = n / (1/x1 + 1/x2 + 1/x3 + ... + 1/xn)

- Where:
  - x1, x2, x3, ..., xn are the individual values in the dataset.
  - n is the total number of values in the dataset.

The harmonic mean places more weight on smaller values in the dataset.

EXAMPLE: NETWORK SPEEDS

Suppose you're analyzing the network speeds (in Mbps) of various internet service providers in a region.

Provider A: 100 Mbps

Provider B: 50 Mbps

Provider C: 20 Mbps

find the **average speed** that reflects the overall network performance more accurately.

CONTD.

Using the arithmetic mean, you would calculate (100 + 50 + 20) / 3 = 56.67 Mbps.

Harmonic Mean = 3 / (1/100 + 1/50 + 1/20) ≈ 35.09 Mbps

the harmonic mean takes into account the reciprocal nature of network speeds (higher speed values are better) and gives more weight to the slower speeds.

useful when you're concerned about the performance of the slowest provider and want to avoid bias towards the faster providers.

provide a more balanced average that reflects the underlying relationships between values.

In Machine Learning

- Tests
  - Metric1 (ratio) = .90
  - Metric2 (ratio) = .40

WINSORIZING

Winsorizing is a data preprocessing technique used in statistics to mitigate the effects of outliers by capping extreme values.

It involves replacing extreme values with less extreme values, such as the highest and lowest values within a certain range.

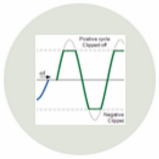
EXAMPLE 1: INCOME DISTRIBUTION ANALYSIS

- Suppose you are analyzing the income distribution of a population and have collected income data.
- However, due to various reasons, some individuals have extremely high incomes that are outliers.
- These outliers can distort the analysis and statistical measures such as the mean and standard deviation.

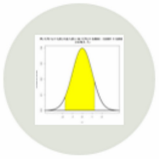
Original Income Data:
$20,000
$30,000
$35,000
$40,000
$1,000,000 (outlier)
$2,000,000 (outlier)

Winsorized Income Data:
$20,000
$30,000
$35,000
$40,000
$40,000 (capped outlier)
$40,000 (capped outlier)

RECOMMENDATION



Winsorizing is a useful technique to consider when you want to reduce the influence of outliers without discarding them completely.



PS : Winsorizing alters the data distribution, and its application should be based on domain knowledge and the goals of your analysis.

Determining the percentage of values

1. **Context and Purpose**: Consider the context of the analysis and the purpose of calculating the mean. If precision and accuracy are paramount, you may opt for a lower percentage. However, if robustness to outliers is more important, a higher percentage may be preferable.
2. **Empirical Studies or Guidelines**: Consult empirical studies or guidelines in your field of study, if available, to determine commonly used percentages for trimming or truncation. These guidelines may provide insights based on previous research or best practices.

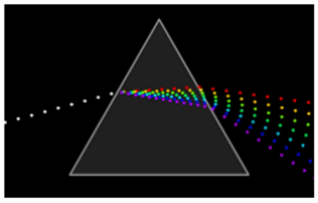
Determining the percentage of values

1. **Distribution of Data**: Assess the distribution of the data. If the data is symmetrically distributed and does not contain many outliers, you may choose a lower percentage (e.g., 5%) for trimming or truncation. However, if the data is skewed or contains outliers, a higher percentage (e.g., 10% or more) may be more appropriate.

2. **Outlier Sensitivity**: Consider the sensitivity of your analysis to outliers. If outliers have a significant impact on the mean and you want to minimize their influence, you may choose a higher percentage for trimming or truncation. Conversely, if outliers contain important information or are of interest, you may choose a lower percentage.

3. **Sample Size**: For smaller datasets, using a higher percentage for trimming or truncation may result in too much data loss, reducing the reliability of the mean estimate. Conversely, for larger datasets, a higher percentage may be more feasible without significant loss of information.

Dispersion of data (techniques)



Dispersion

- **Range:**
  - difference between the maximum and minimum values in a data series.
  - gives an idea of the span of the data, but it can be sensitive to outliers.
- **Variance:**
  - measures the average squared deviation of each data point from the mean of the dataset.
  - gives an idea of how much individual data points deviate from the mean and provides a more comprehensive measure of dispersion.
  - squaring the differences, which can make it harder to interpret.
- **Standard Deviation:**
  - square root of the variance.
  - a widely used measure of dispersion that indicates the average amount of deviation from the mean.
- **Interquartile Range (IQR):**
  - range between the first quartile (25th percentile) and the third quartile (75th percentile) of a dataset.
  - measures the spread of the middle 50% of the data and is less affected by extreme values.

Range

- **Definition:**
- calculated as the difference between the maximum and minimum values.
- Range = Maximum Value – Minimum

Limitations and Considerations

Sensitivity to Outliers:

range is sensitive to extreme values

Limited Information:

While the range provides a basic understanding of spread, it does not capture the distribution of values within the dataset.

8/4/2024

range of data can be useful

9

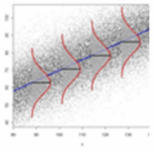
Data Exploration: Range provides a quick overview of the spread of the data. It helps in understanding the distribution and scale of the dataset.

Feature Selection: In feature engineering, knowing the range of different features can help you decide which features are relevant and which ones may need normalization or scaling.

Data Cleaning: Identifying outliers is crucial in data cleaning. The range can help you determine what values might be considered outliers or errors.

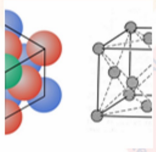
Normalization: If your data features have different ranges, normalization techniques such as Min-Max scaling or Z-score normalization can be applied to bring them to a similar scale, which is often useful in machine learning algorithms.

DESCRIPTION



Variance is calculated by taking the average of the squared differences between each data point and the mean of the dataset.

A high variance indicates that the data points are more spread out from the mean,



a low variance indicates that the data points are clustered closer to the mean.

COMPUTE VARIANCE

POPULATION

$$\sigma^2 = \frac{\sum(x_i-\mu)^2}{N}$$

Where:

- x_i represents each value in the population.
- μ is the mean of the population.
- N is the total number of values in the population.

SAMPLES

For a sample, the formula is slightly different:

$$s^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$$

Where:

- x_i represents each value in the sample.
- $\bar{x}$ is the sample mean.
- n is the number of values in the sample.

Variance vs std dev

Units of Measurement:	Scale:	Consistency with Mean:	Normalization:
• Variance is in squared units (e.g., square meters, square dollars), making interpretation difficult. • Standard deviation, being the square root of variance, is in the same units as the original data, making it more interpretable.	• Standard deviation provides a measure of the typical deviation from the mean in the original scale of the data, making it more intuitive for practical interpretation.	• Since standard deviation is the square root of variance, both measures provide a sense of how spread out the data is relative to the mean.	• Standard deviation is often preferred when comparing variability across different datasets or populations because it avoids the issue of the squared units in variance.

Mean absolute deviation

is a measure of the average absolute deviation of a set of values from their mean.

provides a measure of the variability or dispersion of a dataset.

calculate MAD

1. **Calculate the Mean**: Find the arithmetic mean (average) of the dataset.
2. **Calculate the Absolute Deviation**: For each data point, find the absolute difference between the data point and the mean.
3. **Calculate the Mean of the Absolute Deviations**: Find the average of these absolute differences.

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \text{mean}|$$

Where:

- n is the number of data points in the dataset.
- x_i represents each individual data point.
- mean is the mean of the dataset.

when to use MAD

Robust Measure of Dispersion: MAD is less sensitive to outliers compared to other measures of dispersion like the standard deviation.

Non-Normally Distributed Data: If your data does not follow a normal distribution or if you're unsure about the distribution of your data

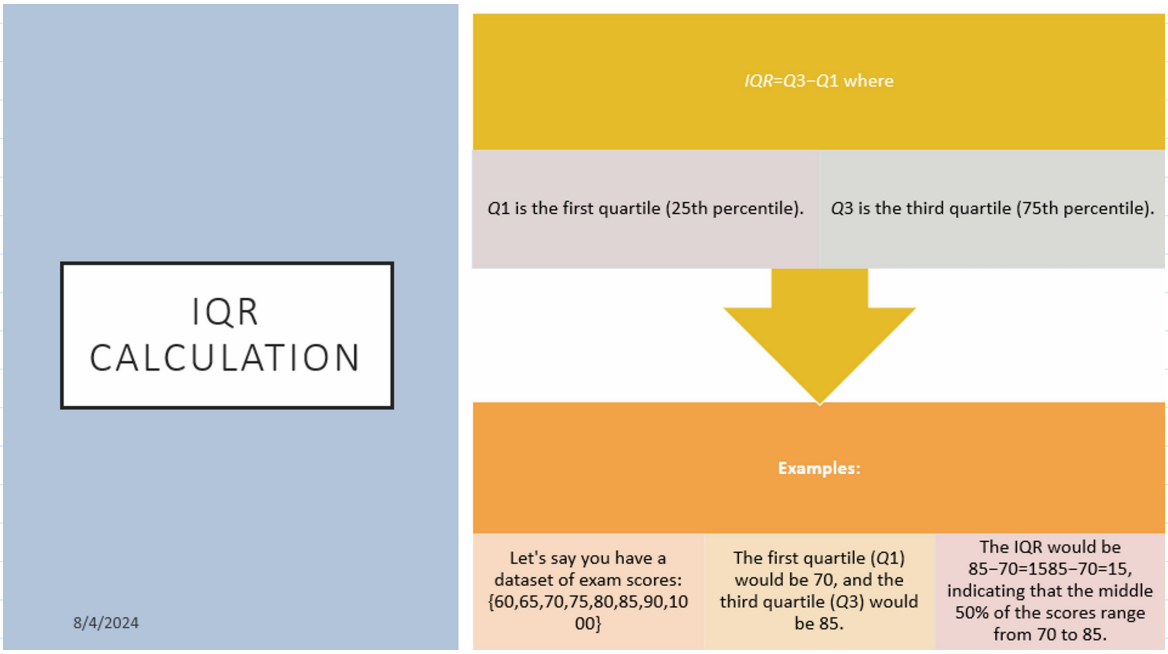
Comparing Variability Across Groups: when the groups have different distributions or when outliers are present.

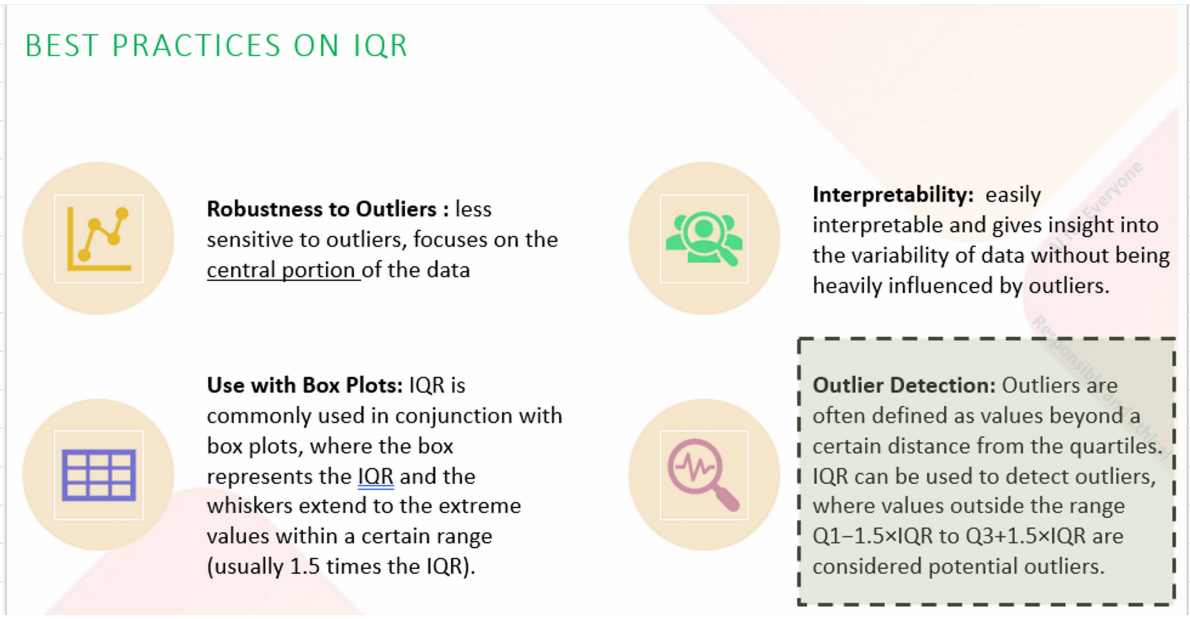
IQR (inter quartile range)

a measure of statistical dispersion that describes the spread of data within the middle 50% of a dataset.

difference between the third quartile (Q3) and the first quartile (Q1).

represents the range of values that cover the central portion of the data, excluding the most extreme values.





Data Symmetry

WHY MEASURE SYMMETRY IN DATA

01

because it provides valuable insights into the underlying distribution of the data

02

helps us understand how the data is spread around its central tendency

03

it has implications for the validity of statistical methods, the accuracy of predictive models, and the interpretation of results

METHODS TO CHECK ASYMMETRY

Visual Inspection: Plotting histograms, box plots, and density plots can provide a visual indication of the data's symmetry.

Bowley's Skewness: This method uses quartiles to determine skewness based on the position of the median relative to the quartiles.

Kurtosis and Skewness Diagram: This graphical method involves plotting the skewness on one axis and kurtosis (peakness or flatness of the distribution) on another.

kurtosis

Kurtosis is a statistical measure that quantifies the shape of a probability distribution's tail and peak relative to a normal distribution.

understand the heaviness of the tails and the central peak of a distribution compared to the normal distribution, which is often referred to as the "bell curve."

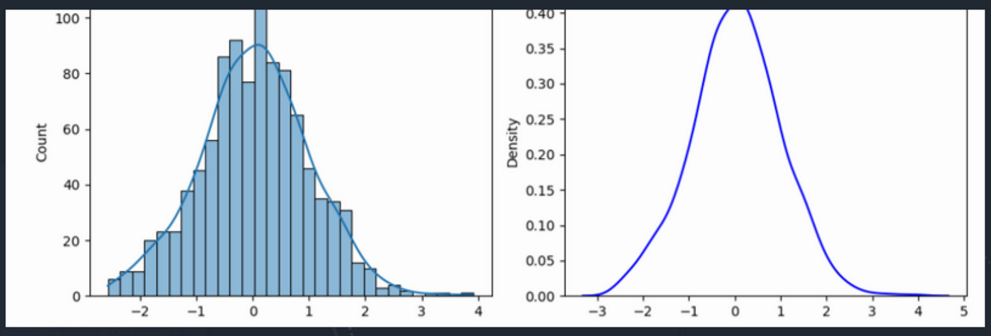
Intuition

Kurtosis is a measure of the "tailedness" of a probability distribution.

Intuitively, it tells us about the relative amount of extreme values (outliers) in a dataset compared to a normal distribution.

If a distribution has positive kurtosis, it means it has fatter tails and more extreme values than a normal distribution. This suggests that there are more outliers or extreme values present.

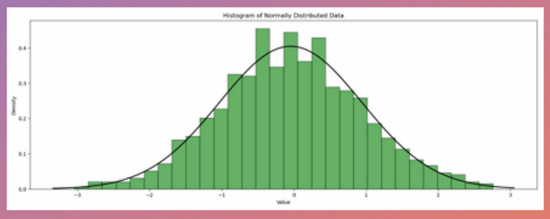
Conversely, negative kurtosis indicates thinner tails and fewer extreme values compared to a normal distribution.



Mesokurtic Distribution (Kurtosis = 3)

- The tails of the distribution are neither too heavy nor too light.
- the distribution has tails and a peak similar to that of a normal distribution.
- moderate tendency to produce outliers or extreme values.

Let us take a normal distribution



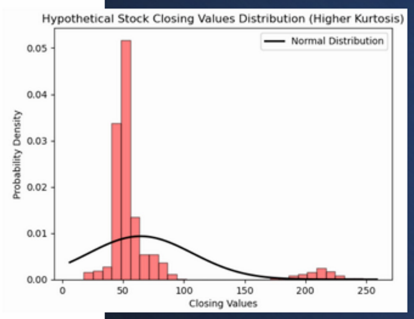
```
# Generate normally distributed data
np.random.seed(0)  # for reproducibility
data = np.random.normal(loc=0, scale=1, size=1000)
```

```
print("Kurtosis (manual calculation):", computed_kurtosis)
```

```
Kurtosis (manual calculation): 2.953233675521671
```

Leptokurtic Distribution (Positive Kurtosis)

- substantially greater than 0
- the distribution has heavier tails, leading to a more pronounced peak
- Positive kurtosis values indicate that the distribution has heavier tails and a sharper peak compared to a normal distribution.
- extreme values are more likely to occur, and the data might exhibit more variability, volatility



Relations and groups

Relations

Relations in data science and machine learning refer to the connections or associations between different data points or groups.

Understanding these relations is crucial for building models that can make accurate predictions and provide insights.

covariance

Covariance measures the degree to which two variables change together.

A positive covariance indicates that as one variable increases, the other tends to increase as well.

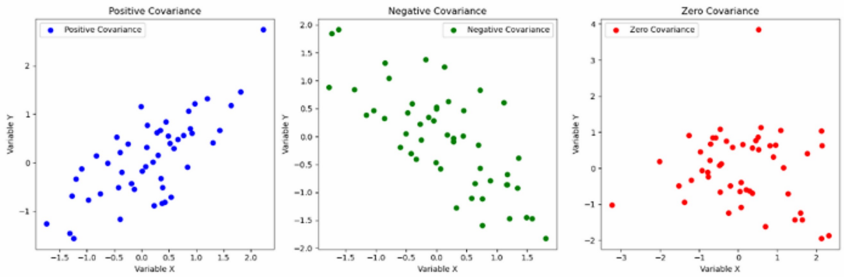
A negative covariance indicates that as one variable increases, the other tends to decrease.

Covariance is influenced by both the units of the variables and their magnitudes, making it hard to interpret on its

often normalized to give the correlation coefficient.

Col1	Col2	Col1- mean1	Col2 - mean2	(col1-mean1) * (col2-mean2)

COV - types



Positive Covariance: results in a scatter plot where points tend to align with a positive slope.

Negative Covariance: results in a scatter plot where points tend to align with a negative slope.

Zero Covariance: results in a scatter plot where points are randomly scattered without a clear alignment.

Calculate Covariance - COV

- COV values ranges from -∞ to +∞
- With covariance, there is no minimum or maximum value, so the values are more difficult to interpret
- a covariance of 50 may show a strong or weak relationship
- this depends on the units in which covariance is measured

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

use case of COV in DS /ML

Dimensionality Reduction:

- Techniques like Principal Component Analysis (PCA) use covariance to find orthogonal directions (principal components) of maximum variance in high-dimensional data.

Multivariate Analysis:

- In multivariate analysis, covariance matrices are used to characterize relationships between multiple variables. For example, in discriminant analysis, covariance matrices can help distinguish between classes.

04-08-2024

Dim Reduction : detailed discussion to follow in this course … FE topic

What is correlation

Correlation measures the strength and direction of the linear relationship between two variables.

It's a normalized version of covariance that ranges from -1 to 1.

close to 1 indicates a strong positive linear relationship,

close to -1 indicates a strong negative linear relationship.

close to 0 suggests a weak or no linear relationship.

CORRELATION

$$\rho_{xy} = \frac{Cov(r_x, r_y)}{\sigma_x \sigma_y}$$

- Correlation is defined as covariance normalized by the product of standard deviations, so the correlation between X and Y
- Correlation coefficients are standardized.. value is always between –1 and 1
- Correlation does not have units.

formula

x	y
92	6.3
145	7.8
30	3.0
70	5.5
75	6.5
105	5.5
110	6.5
108	8.0
45	4.0
50	5.0
160	7.5
155	9.0
180	8.6
190	10.0
63	4.2
85	4.9
130	6
132	7

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
92	6.3	−14.94	−0.11	1.64
145	7.8	38.06	1.39	52.90
30	3	−76.94	−3.41	262.37
70	5.5	−36.94	−0.91	33.62
75	6.5	−31.94	0.09	−2.87
105	5.5	−1.94	−0.91	1.77
110	6.5	3.06	0.09	0.28
108	8	1.06	1.59	1.69
45	4	−61.94	−2.41	149.28
50	5	−56.94	−1.41	80.04
160	7.5	53.06	1.09	58.07
155	9	48.06	2.59	124.68
180	8.6	73.06	2.19	160.00
190	10	83.06	3.59	298.19
63	4.2	−43.94	−2.21	97.11
85	4.9	−21.94	−1.51	33.13
130	6	23.06	−0.41	−9.45
132	7	25.06	0.59	14.79
$\bar{x} = 106.94$	$\bar{y} = 6.41$			$Sum = 1,357.06$
$s_x = 47.28$	$s_y = 1.86$			

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Key points

Predictors (input vars)

response (output vars)

X1 #rooms	X2 #income	X3 #parks	X4 #banks	...	...	Y House_price

- Relationship between the predictors and response variables should be strong

- Relationship amongst the predictors – indicates multi-collinearity and redundancy
- Is a HUGE practical problem
- Inflates the coefficients of the prediction
- How to handle?
  - Feature selection to restrict the columns
  - Use of advanced ML techniques (like Ridge, Lasso, XGB etc ..)

L1 and L2 distances - Key points

Define

Euclidean distance is a measure of straight-line distance between two points in Euclidean space.

It is the most common distance metric and is widely used in various applications, including clustering, classification, and dimensionality reduction.

distance between two points in Euclidean space.

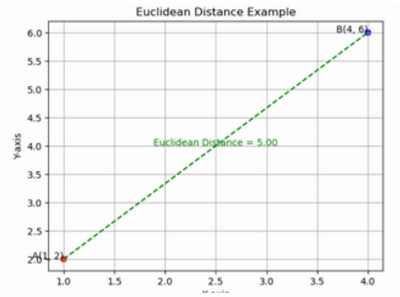
including clustering, classification, and dimensionality reduction.

Formula

For two data points/ rows/ samples/documents ...

A(*x*1, *y*1) and *B*(*x*2, *y*2) in a 2D space, the Euclidean distance is calculated using the formula:

$$d(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



Let's consider two points in a 2D space: *A*(1, 2) and *B*(4, 6)

$$d(A,B) = \sqrt{(4-1)^2 + (6-2)^2} = \sqrt{9+16} = \sqrt{25} = 5$$

Strengths of Euclidean Distance

Intuitive Geometric Interpretation:
- Euclidean distance corresponds to the straight-line distance between two points in Euclidean space, which is a familiar geometric concept.

Simple and Computationally Efficient:
- The calculation of Euclidean distance is straightforward and computationally efficient. It involves basic arithmetic operations (subtraction, squaring, and square root),

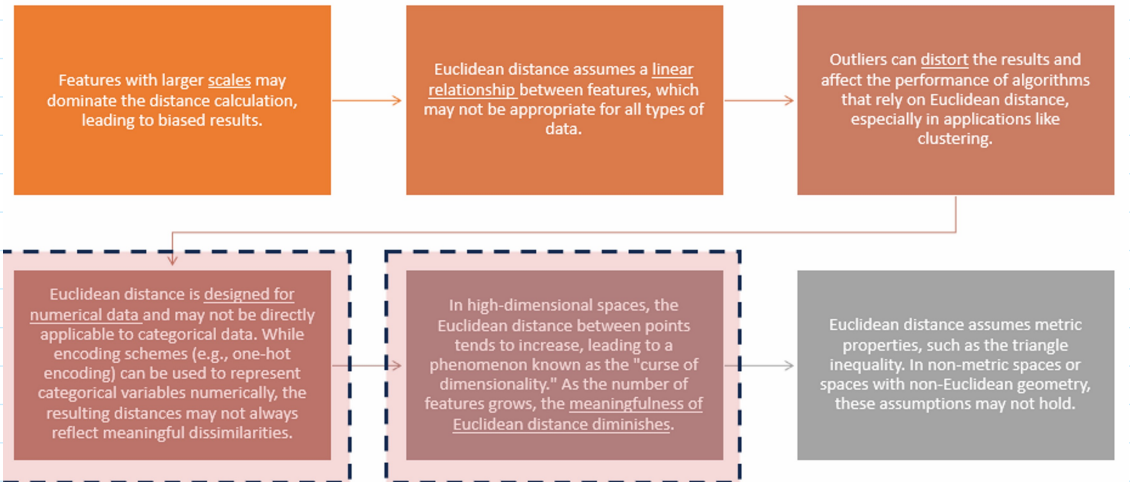
Applicability to Various Data Types:
- Euclidean distance can be applied to a wide range of data types, including numerical, categorical (with appropriate encoding), and spatial data.

Well-Suited for Clustering:
- Euclidean distance is commonly used in clustering algorithms, such as k-means, where it helps define the concept of compact and well-separated clusters.

Relationship to Similarity Measures:
- In many cases, a high similarity corresponds to a low Euclidean distance and vice versa.

Widely Adopted and Understood:
- Euclidean distance is a well-established and widely adopted metric in various fields, including mathematics, physics, and computer science.

weaknesses of Euclidean distance



Effect of Scale Differences on Euclidean Distance

- Euclidean distance is sensitive to the scale of features, meaning that the magnitudes of individual features can influence the overall distance between points.

- **Example:** Consider two points in a 2D space:
  - Point A: (2, 3)
  - Point B: (6, 8)

$$
\begin{aligned}
d_{AB} &= \sqrt{(6-2)^2 + (8-3)^2} \\
&= \sqrt{4^2 + 5^2} \\
&= \sqrt{16 + 25} \\
&= \sqrt{41} \approx 6.4
\end{aligned}
$$

scale the second feature

- (y-coordinate) by a factor of 10
  - Point A: (2, 3)
  - Point B: (6, 80)

- The new Euclidean distance (d'_{AB}) is calculated as:

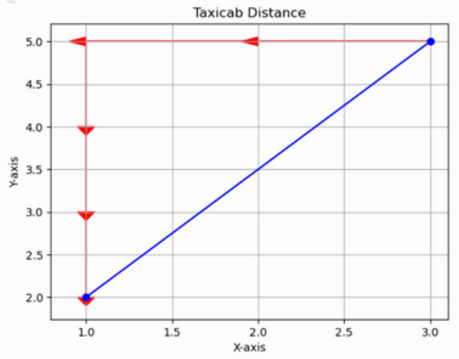
$$
\begin{aligned}
d'_{AB} &= \sqrt{(6-2)^2 + (80-3)^2} \\
&= \sqrt{4^2 + 77^2} \\
&= \sqrt{16 + 5929} \\
&= \sqrt{5945} \approx 77.1
\end{aligned}
$$

Define

Taxicab distance, also known as Manhattan distance or L1 distance, is a measure of the distance between two points in a grid-based system (like a city grid) measured along the grid lines.

Mathematically, the Taxicab distance between two points $(x1,y1)$ and $(x2,y2)$ in a 2D space is given by:

$$\text{Taxicab distance} = |x1-x2| + |y1-y2|$$



type of data Manhattan distance is suitable

Grid-Based or Categorical Data:

- well-suited for grid-based systems, such as city layouts, game boards, or any situation where movement is restricted to horizontal and vertical steps.
- suitable for categorical data

Sparse Data:

- Manhattan distance can be more effective than Euclidean distance.
- As Manhattan distance only considers the non-zero dimensions, making it less sensitive to the presence of zeros.

Robustness to Outliers:

- Manhattan distance is less sensitive to outliers compared to Euclidean distance.

Feature Importance is Equal:

- When all dimensions (features) are considered equally important, Manhattan distance may be appropriate.