Day 5

19 May 2024 22:44

Is it possible to hv infinite values of pdf???

- PDF describes
  - Relative likelihood, not a probability
    - A var taking a certain value (any value)
  - Density
    - Densities can exceed 1
- Probability
  - Area under curve
  - Integral
    - Take a point
    - Take small epsilon (interval)
    - AUC -> probability
      - Cannot be more than 1

Z-score

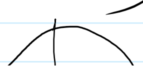
- Data col
  - Quantify
    - How much a data point is away from the mean
      - In terms of std dev

- {1,3,5,7,9,11,13,15,17,19}

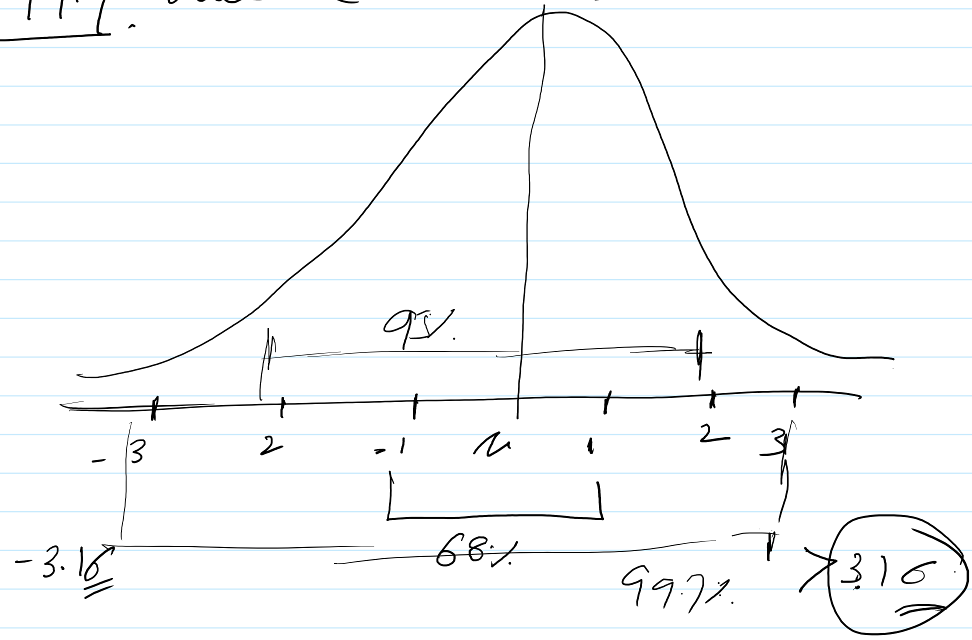
1. For $X = 1$:
   $Z = \frac{1-10}{5} = -1.8$
2. For $X = 3$:
   $Z = \frac{3-10}{5} = -1.4$
3. For $X = 5$:
   $Z = \frac{5-10}{5} = -0.8$
4. For $X = 7$:
   $Z = \frac{7-10}{5} = -0.6$
5. For $X = 9$:
   $Z = \frac{9-10}{5} = -0.2$
6. For $X = 11$:
   $Z = \frac{11-10}{5} = 0.2$
7. For $X = 13$:
   $Z = \frac{13-10}{5} = 0.6$
8. For $X = 15$:
   $Z = \frac{15-10}{5} = 1$
9. For $X = 17$:
   $Z = \frac{17-10}{5} = 1.4$
10. For $X = 19$:
    $Z = \frac{19-10}{5} = 1.8$

Use cases of Z-scores
- Outlier detection **
- normalization

68-95-99.7 rule (3 σ rule)



68-95-99.7 rule (3 σ rule)



Relations in DS
- How one var impacts the other
  - PATTERNS (data science)
- Methods
  - Viz
  - Numeric

- Types of relations
  - Linear correlations for <u>continuous</u> data
    - Pearson correlation coefficient
  - Rank based correlations
    - Spearman
    - Kendall Tau

Covariance

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

Correlation
- Normalized COV
- In the range of -1 to +1

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

-

- Why the range of -1 and +1??

Perfect Positive Correlation (+1):	When ρ_{XY}=+1, it indicates a perfect positive linear relationship. this happens when every deviation in X is matched by an equivalent deviation in Y, resulting in maximum positive covariance relative to the product of their standard deviations.
Perfect Negative Correlation (-1):	When ρ_{XY}=−1, it indicates a perfect negative linear relationship. this happens when every deviation in X is matched by an equivalent but opposite deviation in Y, resulting in maximum negative covariance relative to the product of their standard deviations.
No Correlation (0):	When ρ_{XY}=0, it indicates no linear relationship between the variables. The variables do not move together in any consistent linear pattern.

- Correlation from ML point of view

X_1 (# area), X_2 (# room), X_3 (# ...), X_4 (...) → y house price.

$x_1 | y$
$x_2 | y$
$x_3 | y$
$x_4 | y$

$x_1 | x_2$
$x_2 | x_3$
$x_1 | x_2$
⋮

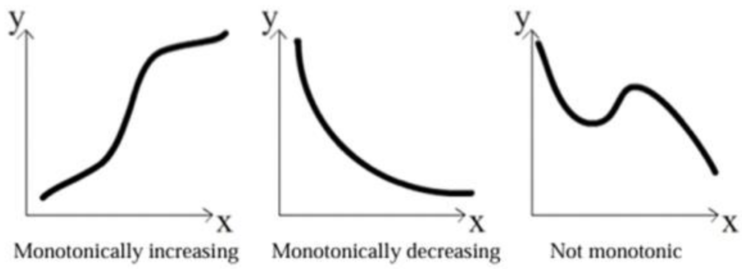
multi collinearity.
✗ ✗ ✗
monster prob.

Limitations of Pearsons corr method

- Linear
- Sensitive to outliers
- Numeric data

Spearman rank corr
- Useful
  - Ordinal
  - Non-linear

- Monotonic function



Monotonically Increasing (Non-decreasing):

- As one variable increases, the other variable also increases.
- Example: A person's age and experience level in their profession.

Monotonically Decreasing (Non-increasing):

- As one variable increases, the other variable decreases.
- Example: The time spent on social media and remaining time for other activities.

Monotonic increasing:-

Example 1

- individuals gain more experience and advance in their careers (increase in age), their salary tends to increase.
- it may not be strictly linear due to factors such as job changes, promotions, and industry differences.

Example 2

- students spend more time studying for an exam (increase in hours of study), their exam scores tend to improve.
- may not be strictly linear due to factors such as study methods, understanding of the material, and test-taking skills.

Example 3

- amount of fertilizer applied to crops increases, the crop yield typically increases.
- may not be strictly linear due to factors such as soil quality, weather conditions, and crop varieties.

Decreasing

Example 1

- more physical exercise (increase in hours of exercise), they tend to lose weight.
- might not be strictly linear due to factors like diet, metabolism, and individual differences.

Example 2

- price of a product increases, the demand for that product generally decreases.
- may not be strictly linear due to factors such as consumer

Decreasing

Example 1

- more physical exercise (increase in hours of exercise), they tend to lose weight.
- might not be strictly linear due to factors like diet, metabolism, and individual differences.

Example 2

- price of a product increases, the demand for that product generally decreases.
- may not be strictly linear due to factors such as consumer preferences, substitutes, and income levels.

Example 3

- altitude increases (elevation above sea level), the temperature tends to decrease.
- it may not be strictly linear due to atmospheric conditions, geographic factors, and local climate variations.

Spearman's Rank Correla

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Kendall Tau Corr coeff

- Monotonic relations
- Similar to spearman
- -1 to + 1
- Non linear data

Why it is called "REGRESSION"

Galton's Theory and Regression in Machine Learning

Galton's Observation:

- In the context of Galton's work on heredity and the heights of parents and children, he observed that extreme values (very tall or very short heights) tended to move or "regress" toward the average height of the population in subsequent generations.

Concept of Regression in Machine Learning:

- In machine learning, regression refers to the statistical method used to model the relationship between variables.
- aims to understand how changes in one variable (dependent variable) are associated with changes in another or multiple variables (independent variables).

Connection:

- connection lies in the idea that both Galton's observation and regression in machine learning involve understanding the tendency of values to move toward some central point (average or mean).

Example: Predicting Heights (Machine Learning Perspective)

Suppose we want to predict a child's height (dependent variable) based on the heights of their parents (independent variables) using a machine learning regression model.

collect data on the heights of parents and their children and train a regression model to learn the relationship.

trained model might reveal that, on average, the child's height is a weighted combination of the parents' heights.

In this way, the model captures the regression-like behavior observed by Galton.