GENZTECHS

ASSIGNMENT 04

DATA SCIENCE INTERNSHIP

HAMZA ASLAM

1. Arithmetic Mean:

Arithmetic Mean ek average hoti hai jo kisi data set ke sab numbers ko add karke, unki total count (numbers ki tadaad) se divide karne se milta hai. Yeh aik central value hai jo data ke distribution ko summarize karne ke liye use hoti hai.

Example:

Aik class ke 5 student’s ke marks hain: 80, 70, 90, 85, 75

400 / 5 = 80

Arithemtic mean = 80

2. Weighted Mean:

Weighted Mean ek average ka advanced version hota hai, jo har value ko uski importance ya weight ke mutabiq adjust karta hai. Har value ko ek weight assign kiya jata hai, aur us value ko uske weight ke saath multiply karne ke baad, sab ko sum karke total weights se divide karte hain.

Example:

i) Education: Final grades based on tests, assignments, and participation weights.

ii) Finance: Portfolio returns where each asset has a different weight based on investment size.

iii) Statistics: Calculating averages for datasets where some observations are more significant.

3. Trimmed Mean:

Trimmed Mean ek average hai jo extreme values (outliers) ka effect kam karne ke liye design kiya gaya hai. Isme data set ke kuch highest aur lowest values ko remove kar diya jata hai, aur baqi values ka arithmetic mean liya jata hai. Yeh technique tab useful hoti hai jab data me extreme values ki wajah se average distort ho raha ho.

Example:

Sports: In gymnastics or diving, highest and lowest scores are removed before calculating the average.

4. Median:

Median ek data set ka central value hai, jo data ko ascending ya descending order me arrange karne ke baad middle position pe hoti hai. Median data ke extreme values (outliers) ka effect nahi leti, is liye yeh ek robust measure of central tendency hai.

Example:

Housing Prices: Median house price shows the central price, avoiding istortion by a few very expensive homes.

5. Mode:

Mode ek data set ka wo value hai jo sabse zyada baar repeat hoti hai. Ye measure of central tendency un scenarios me useful hota hai jahan aapko sabse common value jaan'ni ho.

Example:

If most students score 75 marks in an exam, 75 is the mode, representing the most common performance level.

6. Standard Deviation (SD):

Standard Deviation ek statistical measure hai jo data ke values ki spread (variability) ko quantify karta hai. Yeh batata hai ki data points mean (average) ke around kitni door ya nazdeek hain.

Example:

Scenario: Exam Scores

A teacher collects the following exam scores of 5 students: 70, 72, 68, 75, 65

Mean Score: 70

Standard Deviation: ≈3.16

Interpretation: The scores are closely clustered around the mean of 70.

Comparison with Higher SD:

If the scores were 50,90,70,30,100, the standard deviation would be higher, showing the scores are more spread out.

7. Regression:

Regression ek technique hai jo predict karne me madad karti hai ki ek variable (dependent variable) doosre variable(s) (independent variables) ke basis par kaise change hota hai. Iska main focus cause-and-effect relationship hai.

Points:

Regression ek line of best fit banata hai jo variables ke darmiyan relationship ko represent karta hai.

Linear Regression: Sirf ek independent variable ke liye use hoti hai.

Multiple Regression: Jab ek se zyada independent variables hoon.

Example:

Business: Predict karna ki advertising spend se kitna revenue generate hoga.

Revenue = a + b × Advertising Spend

8. Correlation:

Correlation ek statistical measure hai jo batata hai ki do variables ke darmiyan kis had tak aur kis direction me relationship hai. Lekin ye cause-and-effect ko explain nahi karta.

Points:

Correlation Coefficient (r) ki value: -1 se $+1$

$r=+1$: Strong positive relationship (jab ek variable barhta hai, doosra bhi barhta hai).

$r=-1$: Strong negative relationship (jab ek variable barhta hai, doosra ghatta hai).

$r=0$: No relationship.

Example:

Health: Smokers aur lung cancer ke darmiyan correlation ko measure karna. Agar r=0.85, toh iska matlab strong positive correlation hai, lekin yeh nahi batata ki smoking cancer cause karta hai.

9. Box Plot:

Box Plot ek graphical representation hai jo data ke spread ko samajhne ke liye use hota hai. Isme ek "box" aur do "whiskers" hote hain jo data ke summary statistics ko show karte hain, jaise median, quartiles, aur outliers. Data visualization ke liye use hota hai.

Example:

Exam Scores a teacher wants to understand the distribution of exam scores for a class. The data might look like this:50, 55,60 ,70 ,75 ,80 ,85 ,90 ,95 ,100

A box plot can help show the spread of scores, how many students scored in the lower or higher quartile, and if there are any outliers (e.g., an unusually low or high score).

10. Euclidean Distance:

Euclidean Distance ek metric hai jo do points ke beech ka straight-line distance measure karta hai, jo geometrically 2D ya 3D space me points ke darmiyan hota hai. Ye sabse common distance metric hai jo distance-based algorithms, jaise K-Nearest Neighbors (KNN) aur K-means clustering, me use hota hai.

Example:

Computer Vision: Images ke feature vectors ke beech distance measure karne ke liye use hota hai.

11. Z-Score:

Z-Score ek statistical measure hai jo kisi data point ki position ko indicate karta hai relative to the mean of the data, in terms of standard deviations. Z-score batata hai ki ek specific data point mean se kitna door hai aur ye positive ya negative ho sakta hai depending on whether the point is above or below the mean.

Example: Test Scores

Suppose a school is analyzing the scores of students in a math exam.

The average score (mean) for the exam is 75 with a standard deviation of 5.

A student scores 80.

To calculate the Z-score: Z = (80 - 75)/5 = 1

This means that the student's score is 1 standard deviation above the mean.

12. Manhattan Distance:

Manhattan Distance (jo Taxicab Distance ya City Block Distance ke naam se bhi jana jata hai) ek distance measure hai jo do points ke beech ki distance ko calculate karta hai, lekin yeh Euclidean distance se farq rakhta hai. Manhattan distance sirf horizontal aur vertical movements ko consider karta hai (jaise ke ek taxi jo city ke block-wise streets par chal rahi hoti hai).

Example:

Manhattan distance ko real-world mein city layouts mein use kiya jata hai, jahan aapko ek block se doosre block tak travel karna hota hai, bina diagonal cutting kiye. Misaal ke taur par: Agar ek taxi driver ko ek block se doosre block tak jaana hai, toh woh sirf horizontal aur vertical streets par travel karega, aur Manhattan distance use karke uska total distance measure kiya jaa sakta hai.

13. Minkowski Distance:

Minkowski Distance ek generalized distance metric hai jo Euclidean aur Manhattan distances ko ek common framework me combine karta hai. Iska use different types ke distances ko calculate karne ke liye kiya jata hai, depending on the value of the parameter p.

Example:

Data Normalization: In high-dimensional datasets, Minkowski distance can help normalize the data and standardize the distances between data points.

14. Supremum Distance:

Supremum Distance, jise Chebyshev Distance bhi kaha jata hai, ek distance metric hai jo do points ke beech ka distance calculate karta hai, aur yeh distance maximum coordinate-wise difference ko consider karta hai.

Example:

Supremum distance ko real-world mein un scenarios mein use kiya jata hai jahan ek object ko grid-based space mein move karte hue sabse zyada distance ka pata chalana ho, jaise: Chessboard: Agar aapko chessboard par king ke movement ka distance calculate karna ho, toh aap supremum distance use kar sakte hain, kyun ke king ek time mein ek square move kar sakta hai aur kisi bhi direction mein jaa sakta hai (horizontal, vertical, or diagonal).

15. Chi-Square:

Chi-Square (χ^2) ek statistical test hai jo data ke observed aur expected frequencies ke beech ke differences ko analyze karta hai. Yeh test typically categorical data ke liye use hota hai, jisme hum check karte hain ki kya do categorical variables ke beech koi association hai, ya kisi expected pattern ke against koi deviation hai.

Example:

Maan lijiye ek survey conduct kiya gaya jisme 100 logon se unke Gender aur Smoking Habits ke baare mein pucha gaya. Hum yeh check karna chahte hain ki Gender aur Smoking ke beech koi association hai ya nahi.

16. Covariance matrix:

Covariance Matrix ek square matrix hai jo multiple variables ke beech covariance ko represent karta hai. Covariance, do variables ke beech ke relationship ko measure karta hai, aur yeh batata hai ki kya dono variables ek saath increase ya decrease hote hain. Covariance matrix, especially multiple variables ke analysis mein, isliye use hota hai kyunki yeh ek matrix format mein multiple variables ke covariance ko capture karta hai.

17. Hamming Distance:

Hamming Distance ek measurement hai jo do strings (ya binary sequences) ke darmiyan difference ko dikhata hai. Yeh count karta hai un positions ki, jahan do strings ke characters (ya bits) alag hain. Hamming distance sirf tab kaam karta hai jab do strings ya binary sequences same length ke hon.

Formula ke mutabiq:

Hamming Distance=Number of positions where the corresponding symbols are dif ferent

Example: Strings:

Agar aap ke paas do strings hain:

String 1: "karan"

String 2: "kamal"

Hamming distance count karte hain:

1. k vs k → Same

2. a vs a → Same

3. r vs m → Different

4. a vs a → Same

5. n vs l → Different

Hamming Distance = 2

18. Jaccard Distance:

Jaccard Distance ek measurement hai jo do sets ke darmiyan similarity aur dissimilarity ko quantify karta hai. Yeh Jaccard Index ya Jaccard Similarity Coefficient ka inverse hai. Jaccard Distance calculate karne ke liye: Jaccard Distance = 1− Jaccard Similarity

19. Gower Distance:

Gower Distance ek measurement hai jo heterogeneous data (numeric, categorical, binary, etc.) ke darmiyan similarity ya dissimilarity ko quantify karta hai. Yeh ek generalized metric hai jo tab use hoti hai jab data mein multiple types ke variables ho aur traditional distance metrics (e.g., Euclidean) use nahi kiya ja sakta. Gower distance har feature ke liye individual distances calculate karta hai aur unka normalized average leta hai.

Example:

Agar p=3 features hain:

1. Age (Numeric): $xi1$ = 25, $xj1$ = 35, Range: 15 to 60 (range1 = 45).

2. Gender (Categorical): $xi2$ = "Male", $xj2$ = "Female".

3. Married (Binary): $xi3$ = 1, $xj3$ = 0.

Individual distances:

1. Numeric: $d1\ (i, j)$ = (|25−35|)/45 = 10/45 ≈ 0.222.

2. Categorical: $d2$ (i, j) = 1 (Mismatch).

3. Binary: $d3$ (i, j) = 1 (Mismatch).sss

Gower distance:

D (i, j) = 1/3 (0.222+1+1) = 2.222/3 ≈ 0.741

20. Spearmen:

Spearman's Rank Correlation Coefficient (Spearman's ρ) ek statistical measure hai jo do variables ke darmiyan monotonic relationship ko quantify karta hai. Yeh correlation ka ek non-parametric method hai, jo continuous ya ordinal data ke liye use hota hai.

Spearman correlation un data values ko ranks mein convert karta hai aur phir Pearson correlation coefficient apply karta hai. Yeh measure karta hai ki kya ek variable ke barhne (ya ghatne) par doosra variable bhi barhta (ya ghatta) hai.

Example: Sports: Performance and Rankings

Scenario: A sports analyst evaluates whether players' match performances (e.g., goals scored) correlate with their rankings on a leaderboard. Data: Goals scored are continuous, while rankings are ordinal. Spearman's ρ helps quantify the relationship between performance and rank.