CBSE CLASS XI ARTIFICIAL INTELLIGENCE (SUBJECT CODE 843)

Pearson Correlation Coefficient (Pearson's r)

Unit 6: Machine Learning Algorithms — Linear Relationship Analysis for NCERT Curriculum

NCERT & CBSE Curriculum | Includes Formula, Step-by-Step Calculation, Assumptions & Practical Code

What is Pearson's r & Why is it Critical in AI?

1. Fundamental Definition

• Standardized metric measuring the strength and direction of a linear relationship between two continuous variables.

• Symbol: Denoted as 'r' (sample statistic) or 'ρ' (rho, population parameter).

• Scaled Range: Bounded strictly between -1.0 and +1.0.

• Origin: Developed by Karl Pearson in 1896 based on Galton's regression theory.

2. Prerequisite for Machine Learning

• Linear Regression Gateway: Evaluates if a linear fit is appropriate before training a regression model ($y = mx + c$).

• Feature Selection in ML: Identifies features with high correlation to the target variable and low inter-correlation.

• Rule of Thumb: If $|r| < 0.3$, linear relationship is weak, making simple linear regression unreliable.

3. Continuous Scale Requirement

• Data Type Requirement: Demands continuous variables measured on an Interval or Ratio scale (e.g., marks, study hours, height).

• Linear Sensitivity: Captures straight-line relationships only; non-linear (curved) patterns may yield $r \approx 0$ despite strong association.

• Bounded Measure: Unitless statistic derived by dividing covariance by standard deviations.

Types of Correlation & Interpretation Scale (-1 to +1)

Directional Types of Correlation

- **Positive Correlation (r > 0)**
Both variables move in the SAME direction. As X increases, Y increases.
Example: Hours studied vs. Examination marks scored.

- **Negative Correlation (r < 0)**
Variables move in OPPOSITE directions. As X increases, Y decreases.
Example: Sleep loss / absences vs. Examination performance.

- **Zero Correlation (r ≈ 0)**
NO linear relationship exists between variables. Changes in X do not predict Y.
Example: Student shoe size vs. Mathematics examination score.

Range of r	Strength of Relationship	Variance Explained (r^2)
r = +1.0 / -1.0	Perfect Positive / Negative	100% (r^2 = 1.00)
0.8 ≤ |r| < 1.0	Very Strong Linear	64% - 99%
0.6 ≤ |r| < 0.8	Strong Linear	36% - 63%
0.4 ≤ |r| < 0.6	Moderate Linear	16% - 35%
0.2 ≤ |r| < 0.4	Weak Linear	4% - 15%
0.0 ≤ |r| < 0.2	Very Weak or No Correlation	0% - 3%

Mathematical Formula & CBSE Solved Calculation Example

PEARSON CORRELATION FORMULA (Raw-Score Computational Form):

$$r = [\, n(\Sigma xy) - (\Sigma x)(\Sigma y) \,] \;/\; \sqrt{[\, (n\Sigma x^2 - (\Sigma x)^2) \times (n\Sigma y^2 - (\Sigma y)^2) \,]}$$

Algebra (x)	Trig (y)	x^2	y^2	xy
15	18	225	324	270
16	11	256	121	176
12	10	144	100	120
10	20	100	400	200
8	17	64	289	136
$\Sigma x = 61$	$\Sigma y = 76$	$\Sigma x^2 = 789$	$\Sigma y^2 = 1234$	$\Sigma xy = 902$

Step-by-Step Substitution (n = 5):

1. Numerator Computation:
   5(902) - (61)(76) = 4510 - 4636 = -126

2. Denominator Terms:
   $n\Sigma x^2 - (\Sigma x)^2 = 5(789) - (61)^2 = 3945 - 3721 = 224$
   $n\Sigma y^2 - (\Sigma y)^2 = 5(1234) - (76)^2 = 6170 - 5776 = 394$

3. Calculation:
   $r = -126 / \sqrt{(224 \times 394)} = -126 / \sqrt{88256}$
   $r = -126 / 297.08 = -0.424$

Conclusion: r = -0.424 indicates a Moderate Negative Correlation between algebra & trig marks.

Key Statistical Assumptions & Cautions in AI

1. Interval or Ratio Scale

Both variables must be continuous numerical metrics. Ordinal data (rankings, letter grades) requires Spearman's rank correlation instead of Pearson's r.

2. Linearity Assumption

The relationship must follow a straight line. Always inspect a Scatter Plot first; non-linear (curved) associations cannot be evaluated using Pearson's r.

3. Absence of Outliers

Extreme data points drastically distort r. A single outlier can shift r by 0.3 or more, creating artificial correlation or destroying a real trend.

4. Correlation ≠ Causation

High correlation confirms linear co-movement, NOT cause and effect! Confounding variables or coincidence can drive correlation without causality.

Practical Implementation & Board Exam Summary

Practical Hands-On Tools (Syllabus)

📊 MS Excel Function (Practical Syllabus):
• Formula: =CORREL(array1, array2)
• Example: =CORREL(A2:A8, B2:B8)
• Excel Chart: Add Scatter Plot → Add Linear Trendline → Check 'Display R-squared value on chart'.

🐍 Python Code (Advanced Learners):

```
import numpy as np
from scipy.stats import pearsonr

x = np.array([1, 2, 3, 4, 5])
y = np.array([2, 4, 5, 4, 5])
r, p_val = pearsonr(x, y)
print(f'Pearson r: {r:.3f}')
```

CBSE Class XI Exam Summary Checklist

✓ Definition: Pearson's r measures strength and direction of linear association between 2 continuous variables.

✓ Range: Bounded strictly between -1.0 (perfect negative) and +1.0 (perfect positive); 0 means no linear correlation.

✓ Formula: Know the raw-score formula for manual calculation in theory paper.

✓ Link to Regression: Linear regression equation $y = mx + c$ should only be fitted when correlation $|r| \geq 0.3$.

✓ Key Distinction: Correlation measures linear dependence; Causation proves one event drives another.