Understanding Correlation and Error Metrics

ISE4132 : AI Application System

AMIN AL (아민알)
Integrated System Engineering (ISE)
010-6853-6648
alaminanik@inha.ac.kr

What is Correlation Coefficient?

Symbol: r
Range: -1 to +1
Shows strength and direction of relationship between two variables.

r = +1 → Perfect positive relation (as X ↑, Y ↑).
r = -1 → Perfect negative relation (as X ↑, Y ↓).
r = 0 → No relation.

Example:
Study hours vs Exam score → likely positive correlation.
Exercise vs Weight → likely negative correlation.

Mean Absolute Error (MAE)

Measures average size of errors (ignores direction).

Formula: $MAE = \frac{1}{n} \sum |y_i - \hat{y}_i|$

y_i = actual value.

$\hat{y}_i$ = predicted value.

$|y_i - \hat{y}_i|$ = absolute error (distance between prediction and reality).

Example:

If prediction errors are 2, -3, 4 → MAE = (|2| + | -3| + |4|)/3 = 3.

Mean Absolute Error (MAE)

Student	Actual Score (y_i)	Predicted Score ($\hat{y}_i$)	Error ($y_i - \hat{y}_i$)	Absolute Error $\mid y_i - \hat{y}_i \mid$
1	90	85	5	5
2	75	70	5	5
3	60	65	-5	5
4	80	70	10	10
5	95	100	-5	5

Now, calculate **MAE**:

Root Mean Squared Error (RMSE)

Similar to MAE but gives more weight to large errors.

Formula: $RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$

$(y_i - \hat{y}_i)^2$ = squared error. Squaring removes negatives and makes big errors count more.

Average the squared errors → Mean Squared Error (MSE).

Useful when large mistakes are more serious.

Example:

Errors = 2, -3, 4 → RMSE = $\sqrt{((2^2 + (-3)^2 + 4^2)/3)}$ = $\sqrt{((4+9+16)/3)}$ = $\sqrt{9.67}$ ≈ 3.1.

Root Mean Squared Error (RMSE)

Student	Actual Score (y_i)	Predicted Score ($\hat{y}_i$)	Error ($y_i - \hat{y}_i$)	Squared Error
1	90	85	5	
2	75	70	5	
3	60	65	-5	
4	80	70	10	
5	95	100	-5	

Sum of Squared Errors:

Mean Squared Error (MSE):

Root Mean Squared Error (RMSE):

Relative Absolute Error (RAE)

Compares prediction errors with a simple baseline model.

Formula: $$RAE = \frac{\sum |y_i - \hat{y}_i|}{\sum |y_i - \bar{y}|}$$

Where, $\bar{y}$ mean of actual values.

Example:

RAE < 1 → Our model is better than just using the average.

RAE = 1 → Same as guessing the average.

RAE > 1 → Worse than average guessing!

Relative Absolute Error (RAE)

Student	Actual Score (y_i)	Predicted Score ($\hat{y}_i$)	Error ($y_i - \hat{y}_i$)	Absolute Error $\mid y_i - \hat{y}_i \mid$
1	90	85	5	5
2	75	70	5	5
3	60	65	-5	5
4	80	70	10	10
5	95	100	-5	5

Total Baseline Error=10+5+20+0+15=50

Now, calculate **RAE**:

Root Relative Squared Error (RRSE)

Similar to RAE, but uses squared errors.

Formula: $$RRSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}}$$

Example:

RRSE < 1 → Good model.

RRSE = 1 → Same as average guess.

RRSE > 1 → Worse than average.

Practical Example

Predicting house prices:

Actual prices: [200, 220, 250]

Predicted: [210, 215, 240]

Compute:

MAE → average absolute difference.

RMSE → square-root of average squared difference.

RAE, RRSE → compare against mean of actual prices.

Code

```python
import numpy as np
from sklearn.metrics import mean_absolute_error,
mean_squared_error

# Actual and predicted house prices
actual = np.array([200, 220, 250])
predicted = np.array([210, 215, 240])

# Correlation Coefficient
correlation = np.corrcoef(actual, predicted)[0, 1]

# Mean Absolute Error (MAE)
mae = mean_absolute_error(actual, predicted)

# Root Mean Squared Error (RMSE)
rmse = np.sqrt(mean_squared_error(actual, predicted))

# Relative Absolute Error (RAE)
mean_actual = np.mean(actual)
rae = np.sum(np.abs(actual - predicted)) /
np.sum(np.abs(actual - mean_actual))

# Root Relative Squared Error (RRSE)
rrse = np.sqrt(np.sum((actual - predicted) ** 2) /
np.sum((actual - mean_actual) ** 2))

# Print results
print(f"Correlation coefficient (r): {correlation:.3f}")
print(f"Mean Absolute Error (MAE): {mae:.3f}")
print(f"Root Mean Squared Error (RMSE): {rmse:.3f}")
print(f"Relative Absolute Error (RAE): {rae:.3f}")
print(f"Root Relative Squared Error (RRSE): {rrse:.3f}")
```

Class Activity 1

Task: Understanding Correlation

Given the following datasets:

Study Hours: [2, 4, 6, 8, 10]
Exam Scores: [50, 55, 65, 70, 85]

- Plot the data on a scatter plot.
- Guess whether the correlation is positive, negative, or none.
- Calculate the correlation coefficient.

Hint: Use np.corrcoef(x, y) in Python.

Class Activity 2

Task: Calculate MAE and RMSE

Actual values: [100, 120, 150, 130]
Predicted values: [110, 115, 140, 128]

- Calculate errors = actual - predicted

- Find the absolute errors → calculate MAE.

- Square the errors → calculate RMSE.

Discussion Question:
Which metric is more sensitive to larger mistakes?

Class Activity 3

Task: Pick the Best Model

Two models predict house prices. Compare them using MAE and RMSE.

Actual: [300, 310, 320, 330]
Model A: [305, 315, 300, 325]
Model B: [295, 325, 310, 340]

- Calculate MAE for both models.
- Calculate RMSE for both models.

Discuss: Which model is better and why?

Class Activity 4

Task: Mini Project (Visualization of Errors)

Take any small dataset (e.g., weather: actual vs predicted temperatures).

- Calculate correlation between actual and predicted.
- Compute MAE, RMSE, RAE, RRSE.
- Create a bar chart comparing error values.
- Present findings in 2–3 sentences.

Discussion Question:

If correlation is high, does it always mean the model is accurate? Why/why not?