Module 8 Quiz and Answer

1. What is the main purpose of feature scaling in machine learning?

a) To remove missing values from the dataset
b) To bring all features to a similar numerical range
c) To encode categorical variables into numbers
d) To split data into train and test sets

Ans: b) To bring all features to a similar numerical range

ব্যাখ্যা: Machine learning model-এ যদি একটি feature-এর range ০–১ হয় এবং আরেকটির range ০–৫০০ হয়, তাহলে model বড় range-এর feature-কে বেশি গুরুত্ব দেয়। Scaling সব feature-কে একই range-এ এনে এই সমস্যা solve করে। Learning: Scaling ensures no single feature dominates due to its large numerical range।

2. Which of the following is the correct formula for Standard Scaling?

a) z = (x - min) / (max - min)
b) z = (x - median) / IQR
c) z = (x - mean) / standard deviation
d) z = x / max

Ans: c) z = (x - mean) / standard deviation

ব্যাখ্যা: Standard Scaling-এ প্রতিটি value থেকে mean বাদ দিয়ে standard deviation দিয়ে ভাগ করা হয়। এর ফলে scaled data-এর mean হয় ০ এবং standard deviation হয় ১। Learning: Standard Scaling transforms data to have mean=0 and std=1।

3. After applying Standard Scaling to a feature, what will be the approximate mean of the scaled values?

a) 1
b) 100
c) 0
d) -1

Ans: c) 0

ব্যাখ্যা: Standard Scaling-এর পর data-এর mean সবসময় ০ হয়, কারণ প্রতিটি value থেকে mean বিয়োগ করা হয়। এই property-কে centering বলে এবং এটি distance-based algorithm-এর জন্য খুবই গুরুত্বপূর্ণ। Learning: Standard Scaler always centers data around mean=0।

4. What is the range of output values produced by MinMax Scaling?

a) -1 to 1
b) 0 to 100
c) 0 to 1
d) -3 to 3

Ans: c) 0 to 1

ব্যাখ্যা: MinMax Scaling formula হলো (x - min) / (max - min)। এই formula ব্যবহার করলে সবচেয়ে ছোট value হয় ০ এবং সবচেয়ে বড় value হয় ১। সব value এই দুটির মাঝে compress হয়ে যায়। Learning: MinMax Scaler compresses all values into the [0, 1] range।

5. Which scaling technique is most affected by outliers in the data?

a) RobustScaler
b) MinMaxScaler
c) No scaler is affected by outliers
d) All scalers handle outliers equally

Ans: b) MinMaxScaler

ব্যাখ্যা: MinMaxScaler formula-তে min এবং max ব্যবহার হয়। একটি extreme outlier থাকলে min বা max অনেক বড় বা ছোট হয়ে যায়, ফলে বাকি সব normal values একটি খুব ছোট range-এ চাপা পড়ে যায়। এই কারণে outlier-এর উপস্থিতিতে MinMaxScaler সবচেয়ে বেশি ক্ষতিগ্রস্ত হয়। Learning: MinMaxScaler is highly sensitive to outliers because it relies on min and max values।

6. What makes RobustScaler different from StandardScaler?

a) RobustScaler uses mean and standard deviation
b) RobustScaler uses median and IQR instead of mean and std
c) RobustScaler only works on binary features
d) RobustScaler removes outliers from the dataset

Ans: b) RobustScaler uses median and IQR instead of mean and std

ব্যাখ্যা: RobustScaler formula হলো (x - median) / IQR। Median এবং IQR outlier দ্বারা প্রভাবিত হয় না, কারণ এগুলো middle 50% data দিয়ে calculate হয়। তাই outlier-heavy data-এ RobustScaler অনেক বেশি reliable। Learning: RobustScaler is the best choice when data contains significant outliers।

7. What is the purpose of ColumnTransformer in sklearn?

a) To drop unnecessary columns from a DataFrame
b) To apply different transformations to different columns at once
c) To merge two DataFrames into one
d) To convert all columns to float type

Ans: b) To apply different transformations to different columns at once

ব্যাখ্যা: ColumnTransformer একটি powerful tool যেটি দিয়ে একসাথে বিভিন্ন column-এ বিভিন্ন transformation apply করা যায়। যেমন numeric column-এ StandardScaler এবং categorical column-এ OneHotEncoder একই step-এ apply করা সম্ভব। Learning: ColumnTransformer allows applying multiple transformations in a single, clean step।

8. Why is it important to fit the scaler only on training data and not on the entire dataset?

a) Because scalers cannot process large datasets
b) To prevent information from the test set leaking into the training process
c) Because test data does not need to be scaled
d) Because fitting on full data makes the model faster

Ans: b) To prevent information from the test set leaking into the training process

ব্যাখ্যা: যদি পুরো dataset-এ scaler fit করা হয়, তাহলে test set-এর statistical information (mean, std, min, max) training process-এ চলে আসে। এটাকে data leakage বলে এবং এতে model-এর real-world performance overestimate হয়। Learning: Always fit scalers on training data only to avoid data leakage।

9. A developer fits and transforms the test data separately using its own statistics. What problem does this cause?

a) The test data will have more features than training data
b) The model will train faster but predict slower
c) The scaled test data will be on a different scale than training data, causing wrong predictions
d) The test data will lose all its rows

Ans: c) The scaled test data will be on a different scale than training data, causing wrong predictions

ব্যাখ্যা: Model training-এর সময় যে scale ব্যবহার হয়েছে, prediction-এর সময়ও একই scale দরকার। Test data আলাদাভাবে scale করলে model ভুল input পায় এবং prediction ভুল হয়। তাই train set থেকে fit করা scaler দিয়েই test set শুধু transform করতে হয়। Learning: Use the same fitted scaler on both train and test data for consistent predictions।

10. Which of the following best describes when to use MinMax Scaling over Standard Scaling?

a) When the data has a Gaussian distribution and no outliers
b) When you need output values strictly between 0 and 1, such as for neural networks
c) When the dataset has too many missing values
d) When features are already on the same scale

Ans: b) When you need output values strictly between 0 and 1, such as for neural networks

ব্যাখ্যা: Neural network-এ input values ০ থেকে ১-এর মধ্যে রাখা ভালো কারণ activation function-গুলো এই range-এ ভালো কাজ করে। MinMaxScaler ঠিক এই কাজটি করে। তবে data normally distributed হলে এবং outlier না থাকলে StandardScaler বেশি উপযুক্ত। Learning: Use MinMaxScaler when a bounded [0,1] output is required, especially for neural networks।