Mohammad Yaser Azad | 810800003
Parham Javan | 810800008

summery

1. Preprocessing the Football Dataset

- Clean the data: Handle missing values, correct inconsistencies, and standardize formats.

- Analyze the dataset: Calculate statistical measures, visualize data distributions, and relationships between features.

- Feature engineering: Create new features, such as shot distance and angle using provided formulas.

- Feature selection: Identify and retain important features, possibly using mutual information to rank feature importance.

- Justify choices: Explain why certain features were kept or discarded.

2. Multivariate Regression Implementation

- Implement from scratch: Write code for multivariate regression using the gradient descent algorithm.

- Training and validation: Split the car dataset into training and test sets, train the model, and validate predictions against actual values.

- Visualization: Plot predicted vs. actual values for "Price in Thousands" and "Horsepower".

- Learning curve: Show how model performance improves over iterations of training.

3. Manual K-Fold Cross Validation Implementation

- Implement K-Fold from scratch: Write code to split the dataset into K parts, train and validate the model on each fold.

- Statistical validation: Use metrics like mean squared error to evaluate model performance across folds.

- Learning curve: Display how model performance evolves during training for each fold.

4. Comparison with Built-in Python Libraries

- Repeat the tasks: Use libraries like `scikit-learn` to perform multivariate regression and k-fold cross-validation.

- Compare results: Evaluate the differences in performance, ease of implementation, and computational efficiency between your custom implementations and the library functions.

Question 1:

Describe your strategy for addressing challenges such as handling missing values and categorical features. Could you also elaborate on your feature selection metrics and explain the rationale behind them?

Handling Missing Values: Strategies like mean imputation, median imputation, or more advanced techniques like K-nearest neighbors (KNN) can be employed to handle missing values. The choice of imputation method depends on the nature of the data and the extent of missingness.

Handling Categorical Features: Categorical features can be encoded using techniques like one-hot encoding or label encoding, depending on the nature of the data and the algorithm being used. One-hot encoding is preferred when there is no ordinal relationship among categories, while label encoding can be used when there is ordinality.

Feature Selection Metrics: Feature selection metrics such as correlation analysis, mutual information, or feature importance from tree-based models can be used to select the most relevant features for the model. The rationale behind these metrics is to retain features that have the most significant impact on the target variable while removing redundant or irrelevant ones, thus improving model performance and reducing overfitting.

Question 2:

Why didn't we use regression to predict whether a shot would result in a goal?

Predicting whether a shot would result in a goal typically involves classification rather than regression because the outcome is categorical (goal, own goal, saved, etc.), not continuous. Regression predicts continuous values, making it unsuitable for this type of prediction task.

Question 3:

How would you go about verifying the accuracy of the given formula used to calculate the shot angle in the preprocessing section?

To verify the accuracy of the shot angle calculation formula, you could manually calculate the angles for a few sample shots using basic trigonometry and compare them with the angles obtained from the formula provided. Additionally, you could cross-reference with known shot angles from official match data or use domain knowledge experts to validate the formula's accuracy.

Question 4:

Discuss the advantages and disadvantages of k-fold cross-validation. Can you also explain other types of cross-validation methods that could address the limitations and issues associated with k-fold cross-validation?

Advantages of K-Fold Cross-Validation: Provides a more accurate estimate of model performance, reduces bias, and utilizes the entire dataset for training and validation.

Disadvantages of K-Fold Cross-Validation: Computationally expensive, especially for large datasets, and may not be suitable for certain types of data distributions (e.g., time-series data).

Other Cross-Validation Methods:

Leave-One-Out Cross-Validation (LOOCV): In LOOCV, only one data point is used as the validation set, and the rest are used for training. This method can be computationally expensive but provides a less biased estimate of model performance.

Stratified K-Fold Cross-Validation: Ensures class balance in each fold, which is useful for imbalanced datasets where one class may dominate the data.

Time Series Split: Maintains the temporal order of data for time-series problems, ensuring that data from the future is not used for training when predicting past observations.

Question 5:

What metrics did you use to evaluate your manual implementations of multivariate regression and k-fold cross-validation, and why did you choose them?

For multivariate regression, metrics like Mean Squared Error (MSE) or R-squared can be used to evaluate the model's performance. MSE measures the average squared difference between predicted and actual values, while R-squared represents the proportion of variance explained by the model.

For K-Fold cross-validation, metrics like average MSE or average R-squared across all folds can be used to evaluate the overall performance of the model. These metrics provide a comprehensive measure of model accuracy and generalization to unseen data.