Introduction to Data Science — Assignment 5

Instructors: **Dr. Bahrak, Dr. Yaghoobzadeh**

TA(s): Mohammad Javad Besharati

Deadline: Tuesday, Ordibehesht 18th, 11:59 PM

Introduction

In this assignment, you are expected to apply feature engineering techniques to a football-related dataset to analyze the likelihood of scoring a goal through a shot. Next, you will delve into regression and cross-validation concepts further by implementing multivariate regression and k-fold cross-validation from scratch and utilize them on a preprocessed dataset related to cars. Lastly, you will compare your outcomes with those attained using Python's built-in libraries. This task will strengthen your comprehension of these concepts and their practical implementation.

Dataset

The dataset you'll use for the preprocessing part pertains to football data (football.csv). It includes information about shots such as the timing, location (corner, penalty, etc.), and the outcome of the shots (saved by the keeper, blocked by defenders, missed, etc). Uncovering more intriguing facts about this dataset is up to you :)

For the implementation parts, you'll use a completely different, preprocessed dataset that contains information about cars (cars.csv). You should use this dataset to train your custom multivariate regression and k-fold models to predict the "Price in Thousands" and "Horsepower" columns.

Hint

The responsibility of splitting the data into training and testing sets lies with you, as no test data has been provided.

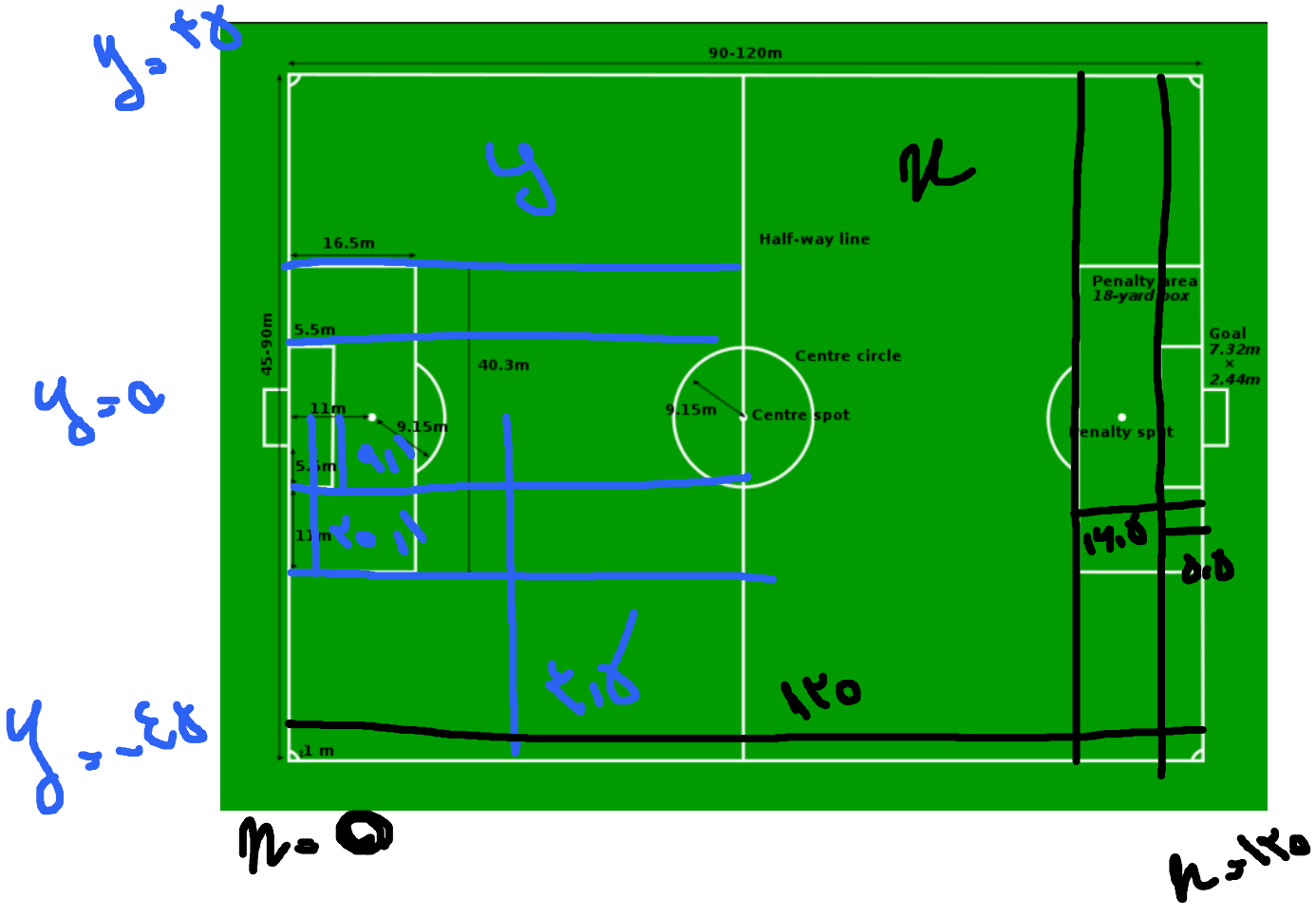
Tasks

1. Preprocessing

You are tasked with cleaning and analyzing the dataset, highlighting its statistical attributes and visualizing its features. Your goal is to identify the beneficial features and justify your conclusions convincingly. Additionally, you should employ feature engineering techniques to refine the dataset, either by removing or replacing less desirable features. To gain a deeper understanding of feature engineering, it is recommended to train an arbitrary but appropriate model and evaluate the outcomes before and after preprocessing. Furthermore, to assess the importance of each feature, utilize the mutual information method to create a pandas dataframe with two columns: one for features and the other for their importance. Subsequently, sort the dataframe in descending order based on importance and display the results.

Hints

- To gain better insights into a football pitch, you can refer to the following image:



- You can apply various methods that you have learned throughout the course to fill missing values and manipulate categorical features in your data.
- You can consolidate similar features. For instance, you could treat "goal" and "own goal" as the same.
- Employ feature selection to exclude less significant features, thereby reducing the dimensionality and lowering computational costs.
- For a more thorough analysis, consider extracting new, more informative features from existing ones. For instance, calculate shot distance and angle using the following formulas and incorporate them into your analysis:

$$distance = \sqrt{x^2 + y^2}$$

$$angle = \begin{cases} rad2\deg(arctan(\theta)) & \arctan(\theta) \geq 0 \\ rad2\deg(arctan(\theta + \pi)) & \arctan(\theta) < 0 \end{cases}, \theta = \frac{7.32x}{x^2 + y^2 - \left(\frac{7.32}{2}\right)^2}$$

2. Multivariate Regression Implementation

Implement multivariate regression from scratch and use the gradient descent algorithm to update the weights. Validate the regression model by providing a visual comparison between the predicted and actual values for “Price in Thousands” and “Horsepower”. Additionally, plot the accuracy across different random states for a more robust verification. Finally, display a learning curve to illustrate the progression of the regression process.

3. Manual K-Fold Cross Validation Implementation

Implement K-Fold cross-validation from scratch. As in the previous section, use the gradient descent algorithm to adjust the weights. Then, validate your custom K-Fold implementation using statistical metrics. Finally, display a learning curve upon completion.

Hint

You can use some parts of the code you’ve implemented in the previous section, here.

4. Comparison with Built-in Python Libraries

Now, compare the results from your custom implementations in sections 2 and 3 with those obtained using built-in Python libraries, and report the findings.

Questions

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?

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

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

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?

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?

Notes

- Upload your work as a zip file in this format on the website: DS_CA5_[Std number].zip. If the project is done in a group, include all of the group members' student numbers in the name.
- If the project is done in a group, only one member must upload the work.
- We will run your code during the project delivery, so make sure your results are reproducible.