Machine Learning Techniques for Predictive Modeling and Classification

Project 6

OCT 11, 2024

ANKIT BHARTI

(CS 651 03) Cloud Computing & Big Data Analytic 2024 FA

Steven Thomas

Introduction

In this module, I used machine learning on Azure Machine Learning with the prepared data to build predictive models, execute classification, and take advantage of scalable, flexible computing in Azure. I have designed, optimized, and tested three machine learning models for their performance regarding accuracy and scalability. Ensuring high-performance training of the models was possible by using distributed computing frameworks like Apache Spark

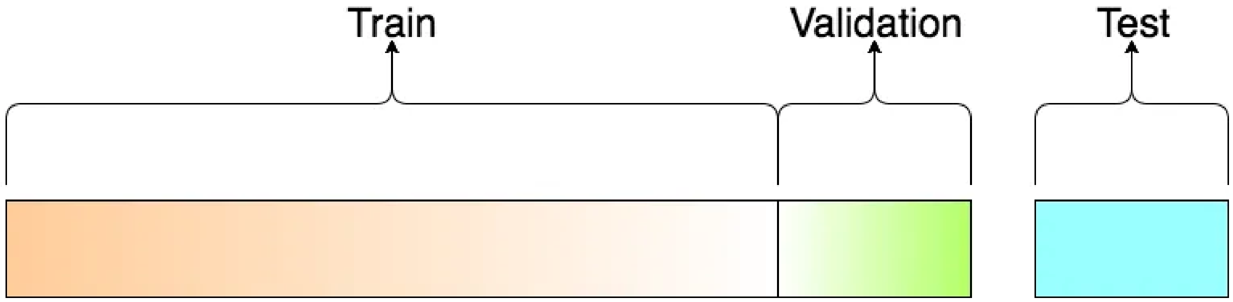
Data Preparation and Feature Engineering

I cleaned the data further and performed feature engineering on the data using Azure Machine Learning Studio to prepare it for machine learning. I selected key features according to their level of correlation with the target variable. Further, I treated missing values in the data using imputation techniques. I used one-hot encoding on categorical data to transform the data into forms suitable for machine learning algorithms.

Model Selection and Training

1. Linear Regression Predictive Model

Objective: To predict continuous outcome variables; for example, product sales or customer spending. I called the built-in linear regression algorithm within Azure Machine Learning to train the model. The data had been divided into 80% training and 20% test subsets. **Training data** is the subset of the dataset used to train the model. This is where the model learns patterns, relationships, and features of the data. **Validation data** is used to tune the model’s hyperparameters and to provide an unbiased evaluation of the model while tuning. Hyperparameters are the aspects of the model that are not learned from the data but are set before the training process begins (e.g., learning rate, number of layers in a neural network). **Test data** is the subset of the dataset used to provide an unbiased evaluation of a final model that fit on the training dataset. It is only used after the model has been trained (and validated).



2. Random Forest Classification:

Objective: Customer purchasing behavior regarding product usage and historical data. I have developed this model using the RandomForestClassifier algorithm in Azure. I tried a wide range regarding the number of decision trees and the depth of each tree to optimize for accuracy.

3. K-Means Clustering (Segmentation)

Objective: Segment customers into clusters based on purchasing behavior for targeted marketing. I'll implement the K-Means algorithm in Azure Machine Learning Studio, assuming beforehand the number of clusters, k. The Elbow method was utilized to find the number of clusters.

Model Evaluation and Fine-tuning

The model was evaluated against appropriate performance metrics for each of them:

1. Mean Squared Error (MSE) for regression model

2. Confusion Matrix and F1-Score for classification accuracy

3. Silhouette Score for clustering effectiveness.

To make sure that the solution is scalable, I employed distributed computing made available through Azure. For example, training jobs for all models were parallelized on several virtual machines to increase the speed of processing. Hence, it could handle a high volume and variety without any reduction in performance.

Challenges and Adjustments

One was tuning the hyperparameters for the Random Forest model without its overfitting. Iterating through the different combinations of parameters, with the help of Azure's HyperDrive, allowed me to automate this process and therefore get improved results with reduced effort from my end. Partitioning data optimally in Spark clusters so that the workload may be distributed uniformly across nodes was also required for scaling up the Linear Regression model.

Conclusion

Coupled with Apache Spark, Azure Machine Learning proved to be a powerful and scalable platform in building machine learning models pertaining to predictive analytics and classification tasks. After fine-tuning model parameters and the use of distributed computing, I was able to achieve high accuracy and scalability for each model.

References

- **Zaharia, M., et al. (2016). Apache Spark: The Definitive Guide. O'Reilly Media.**

- **Microsoft. (n.d.). Get started with Azure Machine Learning. Retrieved from https://learn.microsoft.com/en-us/azure/machine-learning/tutorial-azure-ml-in-a-day**
- **Zhang, Q., Yang, L. T., Chen, Z., & Li, P. (2018). Machine Learning on Big Data: Opportunities and Challenges. Neurocomputing, 275, 347-361.**