



Report - Week 4

🕒 Created @July 19, 2024 9:53 PM

Week 4 Report: Model Tuning and Hyperparameter Optimization

Overview

This week focused on model tuning and hyperparameter optimization using the Heart Disease UCI dataset. The main objectives were to understand the importance of hyperparameter tuning, apply different optimization techniques, and improve model performance. Here's a summary of the tasks performed and insights gained:

1. Importance of Hyperparameter Tuning

Hyperparameter tuning is crucial as it helps improve model performance and generalization. Proper tuning can significantly enhance the accuracy, reduce overfitting, and optimize the model's ability to generalize to new, unseen data.

2. Techniques for Hyperparameter Optimization

- **Grid Search:** Exhaustively searches through a specified subset of hyperparameters to find the optimal combination.
 - **Random Search:** Randomly samples hyperparameter values from a specified range or distribution. It is often more efficient than grid search when dealing with a large hyperparameter space.
 - **Bayesian Optimization:** Uses probabilistic models to model the performance of different hyperparameter combinations and iteratively refines the search based on past results.
-

3. Practical Implementation

1. Data Preparation

- **Data Cleaning:** Handled missing values, removed irrelevant columns, and replaced certain categorical values. Applied ordinal encoding for ordinal features and label encoding for binary features.
- **Feature Scaling:** Used MinMaxScaler to normalize features.
- **Feature Selection:** Applied mutual information and SelectKBest to choose the most significant features for model training.

2. Model Training and Evaluation

- **Logistic Regression:**
 - **Baseline Model:** Trained a logistic regression model and evaluated its accuracy.
 - **Grid Search for Hyperparameters:** Conducted a grid search to find the best hyperparameters for logistic regression. Best parameters and scores were identified.
- **Random Forest Classifier:**
 - **Baseline Model:** Trained a random forest model and evaluated its accuracy.
 - **Grid Search for Hyperparameters:** Performed a grid search for random forest hyperparameters to improve the model's performance.
- **XGBoost:**

- Trained an XGBoost classifier and evaluated its performance. This model is known for its robustness and efficiency in handling different types of data.
- **Decision Tree Classifier:**
 - Trained a decision tree classifier and assessed its performance.
- **K-Nearest Neighbors (KNN):**
 - Trained a KNN classifier and evaluated its accuracy.

3. Bayesian Optimization

- Implemented Bayesian Optimization for hyperparameter tuning of logistic regression using `BayesSearchCV`.
- Optimized hyperparameters and trained the final model using the best parameters found.

4. Results

- **Logistic Regression:** Achieved good accuracy, and hyperparameter tuning improved performance.
- **Random Forest:** Demonstrated strong performance with optimal hyperparameters.
- **XGBoost:** Provided competitive results and showed the effectiveness of advanced ensemble techniques.
- **Decision Tree:** Simpler model with reasonable performance.
- **KNN:** Performed well but required tuning of hyperparameters for better results.

Best Hyperparameters:

- **Logistic Regression:** `C` and `tol` values found through Bayesian Optimization led to improved accuracy.
- **Random Forest:** Optimal settings for `n_estimators`, `max_depth`, and other parameters were identified.

5. Conclusion

This week's work underscored the importance of hyperparameter tuning and various optimization techniques in enhancing model performance. The practical implementation of grid search, random search, and Bayesian optimization provided hands-on experience in improving model accuracy and understanding the impact of different hyperparameters on performance.
