Report - Week 3

Created @July 12, 2024 4:54 PM

Week 3 Focus: Advanced Machine Learning Algorithms

Breast Cancer Wisconsin (Diagnostic) Data Set

Theory Covered

1. **Ensemble Methods**
   - **Boosting**: Techniques such as XGBoost and LightGBM.
   - **Bagging**: Including methods like Bagging Classifier.
2. **Support Vector Machines (SVM)**
   - Understanding SVMs and their application in classification tasks.
3. **Advanced Regression Techniques**
   - Exploring advanced regression methods relevant to classification and prediction.

Practical Implementation

Data Preprocessing:

1. **Loading and Cleaning Data**:
   - Removed unnecessary columns (`id`, `Unnamed: 32`).
   - Encoded categorical variables using LabelEncoder.
   - Dropped less relevant features (`texture_se`, `symmetry_se`, `fractal_dimension_se`).
2. **Splitting Data**:
   - Divided the dataset into training and testing sets with a 70-30 split.
3. **Feature Scaling**:
   - Standardized features using `StandardScaler`.

Models Implemented:

1. **XGBoost**:
   - Used `XGBClassifier` with parameters: `n_estimators=3`, `max_depth=4`, `learning_rate=0.8`.
   - Achieved satisfactory performance metrics (Accuracy, Precision, Recall, F1 Score).
2. **LightGBM**:
   - Utilized `LGBMClassifier` with parameters: `boosting_type='gbdt'`, `num_leaves=31`, `n_estimators=100`.
   - Evaluated performance using the same metrics.
3. **Bagging Classifier**:
   - Implemented Bagging with `LogisticRegression` as the base estimator.
   - Conducted cross-validation to determine mean accuracy.
4. **Stacked Model**:
   - Built a meta-model using predictions from base models (`DecisionTreeClassifier`, `ExtraTreesClassifier`, `AdaBoostClassifier`,

`RandomForestClassifier`).

- Employed XGBoost as the meta-model and evaluated using ROC-AUC scores.

5. **Voting Classifier**:
   - Combined `RandomForest`, `DecisionTree`, and `LogisticRegression` using `VotingClassifier`.
   - Used soft voting to enhance prediction accuracy.
6. **Support Vector Machine (SVM)**:
   - Implemented `SVC` with a linear kernel for classification.
   - Evaluated performance on the test set with metrics such as Accuracy, Precision, Recall, and F1 Score.

Feature Selection:

1. **Correlation-Based Feature Selection**:
   - Removed highly correlated features to reduce multicollinearity.
2. **Variance Thresholding**:
   - Eliminated features with low variance to ensure robust model performance.

Summary of Results

Model Performance:

- **XGBoost**: Demonstrated high accuracy and robust classification performance.
- **LightGBM**: Showed comparable results to XGBoost with efficient training times.
- **Bagging Classifier**: Provided stable accuracy through ensemble method.
- **Stacked Model/Blending**: Achieved high ROC-AUC score indicating strong predictive capability.
- **Voting Classifier**: Enhanced accuracy through a combination of multiple classifiers.

- **SVM**: Performed well with linear kernel, showcasing the power of SVMs in classification tasks.

Feature Selection Impact:

- Successfully reduced feature dimensionality without sacrificing model performance.
- Ensured models were not overfitted by removing irrelevant or redundant features.

Conclusion

During the third week of the internship at LMKR, the focus was on implementing advanced machine learning algorithms with a specific emphasis on ensemble methods, SVMs, and advanced regression techniques. The practical work involved rigorous, model training, and evaluation using the Breast Cancer Wisconsin dataset. The models implemented demonstrated strong performance, highlighting the effectiveness of ensemble methods and SVMs in classification tasks. Feature selection techniques further optimized the models, ensuring robust and reliable predictions.