Mohammad Bagher Soroush - 400130273 - ML 2 Report

1. Explanation of Work Steps

This project follows a systematic approach to preprocess and train an SVM (Support Vector Machine) model using the Iris dataset. The main steps include:

- Data Loading and Preprocessing: The Iris dataset is loaded and processed. Since the target labels in the Iris dataset are categorical (representing species types), encoding techniques are applied.
- Label Encoding: The categorical target labels are transformed into numeric values for model compatibility.
- Feature and Target Definition: The feature variables (X) and the encoded target variable (y) are defined.
- Model Selection and Hyperparameter Tuning: The SVM model is trained using different hyperparameters, including variations in the penalty parameter (C) and kernel functions (linear, rbf, poly, sigmoid), to identify the best configuration.
- Cross-Validation: A 5-fold cross-validation is applied for each combination of parameters to ensure the reliability of results.
- Model Evaluation and Comparison: After identifying the optimal parameters, the best-performing model is selected and evaluated on test data.

2. Purpose of Using Label Encoding

Label encoding is used to convert categorical labels (species types) into numerical values, allowing machine learning algorithms to interpret and process the target variable. Since SVMs and other machine learning models work with numeric data, converting the categorical data into numeric form is essential for enabling the model to learn from the target values effectively.

3. Cross-Validation Explanation and Reason for Use

Cross-validation is a technique that divides the dataset into several "folds" to evaluate the model on different subsets of data. In this project, K-Fold Cross-Validation with 5 splits is used. Each split involves training on four folds and testing on the remaining fold, rotating through all folds to ensure each subset is used as a test set once.

Reason for Use: Cross-validation helps mitigate overfitting by ensuring the model generalizes well across different subsets of data. By averaging the results over multiple folds, cross-validation provides a more reliable measure of model performance than a single train-test split.

4. Comparison and Interpretation of Trained SVM Models

The two SVM models are trained using different kernel functions and values of C to find the optimal configuration based on model accuracy. Through the cross-validation process, the average accuracy for each combination of hyperparameters was calculated, allowing for comparison.

- Model Performance Comparison: Each model's performance, particularly in terms of accuracy, is compared based on cross-validation scores. The model with the highest cross-validation accuracy is selected as the optimal model.

- Interpretation: A higher cross-validation accuracy suggests that the model is better suited for generalization on unseen data, indicating its ability to handle data variability effectively.

5. Kernel Function Used in the Optimal SVM Model

The kernel function used in the optimal model is the polynomial kernel. Kernel functions in SVM allow the algorithm to fit a model in a higher-dimensional space, enabling it to capture complex relationships within the data. The polynomial kernel, in particular, can represent interactions between features to varying degrees (as specified by the degree of the polynomial).

- Interpretation: Using the polynomial kernel in this context allows the SVM model to classify the Iris dataset classes with nonlinear boundaries, which is beneficial when the classes are not linearly separable in the original feature space.

6. Interpretation of the C Value in the SVM Algorithm

The C parameter in SVM represents the regularization parameter, which controls the trade-off between maximizing the margin and minimizing classification error.

- High C Value: When C is high, the model prioritizes classifying all training points correctly, potentially leading to overfitting.
- Low C Value: A smaller C value allows some misclassifications but results in a larger margin, potentially improving generalization on unseen data.