Optimized Kernel Selection for SVM Classification: A Case Study on the OJ Dataset

Project Explanation:

I conducted a comprehensive analysis of various Support Vector Machine (SVM) models, including linear, radial, and polynomial kernels, to classify data effectively. The analysis involved optimizing hyperparameters using cross-validation and evaluating the models' performance on both training and test datasets. Specifically, the I applied these techniques to the "OJ" dataset, aiming to minimize misclassification errors and determine the most effective kernel for the classification task.

OJ Dataset Overview:

The "OJ" dataset is available in the **ISLR** package in R, which accompanies the book *"An Introduction to Statistical Learning with Applications in R"* by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. The dataset is included as an example for various machine learning and statistical analysis techniques. The "OJ" dataset is derived from a study on orange juice sales. It contains data on purchases made by customers, with various features describing both the customers and the products. The key variables in this dataset include:

- **Purchase**: The response variable indicating whether a customer purchased the Citrus Hill (CH) brand or the Minute Maid (MM) brand of orange juice.
- **PriceCH and PriceMM**: Prices of the Citrus Hill and Minute Maid brands, respectively.
- **LoyalCH**: A measure of customer loyalty to the Citrus Hill brand.
- **StoreID**: The store identification number.
- **WeekofPurchase**: The week in which the purchase was made.

Brief Overview of Components:

- **Purchase Decision**: The key outcome of interest, where the goal is to predict whether a customer will purchase the Citrus Hill or Minute Maid brand.
- **Pricing and Discounts**: Several variables relate to the prices and discounts of both brands, which are critical factors influencing purchase decisions.
- **Customer Loyalty**: The variable *LoyalCH* reflects the degree of customer loyalty to the Citrus Hill brand, providing insights into brand preferences.
- **Store Information**: Variables like *StoreID* and *Store7* offer context on where the purchase was made, which can be relevant in understanding regional or store-specific trends.
- **Promotional Information**: Variables such as *SpecialCH*, *SpecialMM*, *PctDiscMM*, and *PctDiscCH* capture the impact of promotions and discounts on purchasing behavior.

The dataset includes a variety of other features that provide insight into the customer's purchasing behaviour, product preferences, and market conditions. The goal of the analysis is to use these features to accurately predict which brand a customer will purchase, using various SVM models to find the best classification performance.

Project Findings:

1. **Support Vector Classifier (SVC) Performance**:
    - The linear SVM model initially showed a training error rate of 16.6% and a test error rate of 18.1%. After tuning the cost parameter, the training error slightly decreased to 15.8%, but the test error slightly increased to 18.8%.
    - The linear SVM model used a substantial number of support vectors, indicating a high degree of model complexity but not necessarily better generalization to unseen data.
2. **Radial Basis Function (RBF) Kernel Performance**:
    - The RBF kernel SVM, which can capture non-linear relationships, initially had a training error rate of 14.5% and a test error rate of 17%. This model showed a slight improvement in classification accuracy compared to the linear SVM.
    - Tuning the RBF kernel did not significantly reduce the training or test error rates, suggesting that the default parameters were already near-optimal.
3. **Polynomial Kernel Performance**:
    - The polynomial kernel SVM initially showed a training error rate of 17.2% and a test error rate of 18.8%, similar to the linear kernel but with a different model complexity.
    - After tuning, the polynomial kernel's performance improved slightly, reducing both training and test errors, but it still did not outperform the RBF kernel.
4. **Overall Conclusion**:
    - Among the models tested, the **RBF kernel** SVM produced the lowest misclassification error on both the training and test datasets. This suggests that the RBF kernel was most effective in capturing the underlying patterns in the "OJ" dataset, particularly in handling the non-linear relationships between the features and the purchase decision.