ECGR 4105: Intro to Machine Learning HW 4

Axel Leon Vasquez

Student ID: 801182414

October 31st, 2024

Link to GitHub

Problem 1

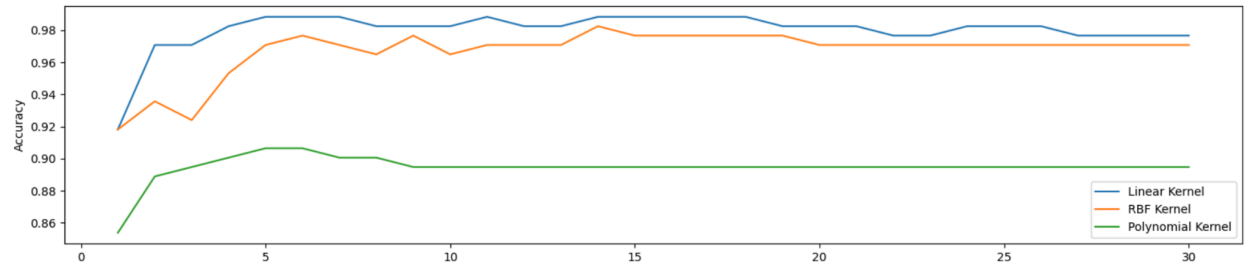
1. **Introduction:**
   This task aims to classify breast cancer types (malignant vs. benign) using the load_breast_cancer dataset, which includes various tumor features like mean radius, texture, and more. A Support Vector Machine (SVM) classifier is employed with Principal Component Analysis (PCA) to reduce dimensionality, and performance is measured across varying numbers of principal components (K). Three SVM kernels—linear, RBF, and polynomial—are tested to evaluate model accuracy and capture complex data patterns. Results are also compared with a Logistic Regression baseline for additional insights.

2. **Methodology:**
   The dataset was standardized to ensure uniform feature scaling, and PCA was applied to reduce dimensionality, varying the number of components (K) from 1 to the maximum feature count. For each K, the data was split 70% for training and 30% for testing. Three SVM models were built using linear, RBF, and polynomial kernels, with accuracy, precision, and recall recorded for each. A Logistic Regression model was trained on the same data split for a baseline comparison.

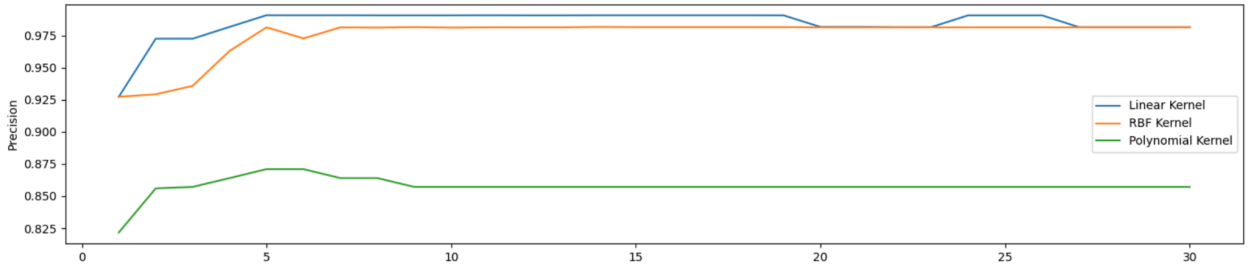
3. **Results:**
   The evaluation of the SVM classifier across different kernels yielded the following optimal number of PCA components (K) for achieving the highest accuracy:
   - **Linear Kernel:**
     - Optimal K: 5
     - Accuracy: 0.9883
   - **RBF Kernel:**
     - Optimal K: 14
     - Accuracy: 0.9825
   - **Polynomial Kernel:**
     - Optimal K: 5
     - Accuracy: 0.9064

   For comparison, the logistic regression model produced the following results:
   - **Logistic Regression Accuracy**: 0.9825
   - **Logistic Regression Precision**: 0.9907
   - **Logistic Regression Recall**: 0.9815

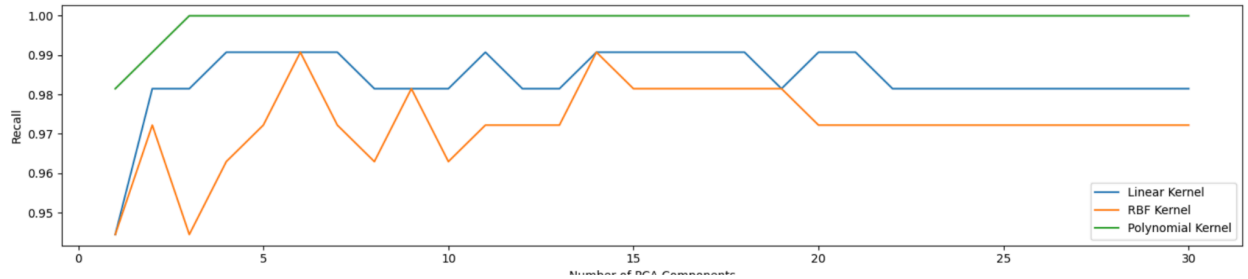
Accuracy Graph



Precision Graph



Recall Graph



4. **Summary of Findings**

   The analysis of SVM classifiers using PCA revealed that the linear kernel was the most effective model, achieving the highest accuracy of 0.9883 with 5 PCA components. The RBF kernel followed closely with an accuracy of 0.9825 at 14 components, while the polynomial kernel had a lower accuracy of 0.9064 at 5 components.

   In comparison, logistic regression achieved an accuracy of 0.9825 with high precision (0.9907) and recall (0.9815). Although logistic regression provided strong results, the SVM models, especially the linear kernel, demonstrated superior performance in classifying cancer types. This highlights the effectiveness of SVM classifiers with PCA in enhancing predictive accuracy for cancer classification.

5. **Comparison to Homework 3 Logistic Regression**

   The logistic regression model from Homework 3 achieved an accuracy of 97.37% with a precision of 97.22% and recall of 98.59% when classifying tumors as benign or malignant. The

model demonstrated strong performance, successfully identifying malignant cases while minimizing false positives.

In contrast, the SVM models in this analysis outperformed the logistic regression results, particularly the linear kernel, which achieved an accuracy of 98.83% with 5 PCA components. The RBF kernel also performed well with an accuracy of 98.25% at 14 components. Additionally, the precision and recall metrics for the linear kernel were impressive, indicating the model's ability to capture the complexities of the dataset more effectively.

Compared to the logistic regression results from Homework 3, SVM, especially with PCA feature extraction, demonstrated superior accuracy, highlighting its efficacy in classifying cancer types. This underscores the advantages of using SVM for complex datasets, where capturing non-linear relationships can significantly enhance classification performance.

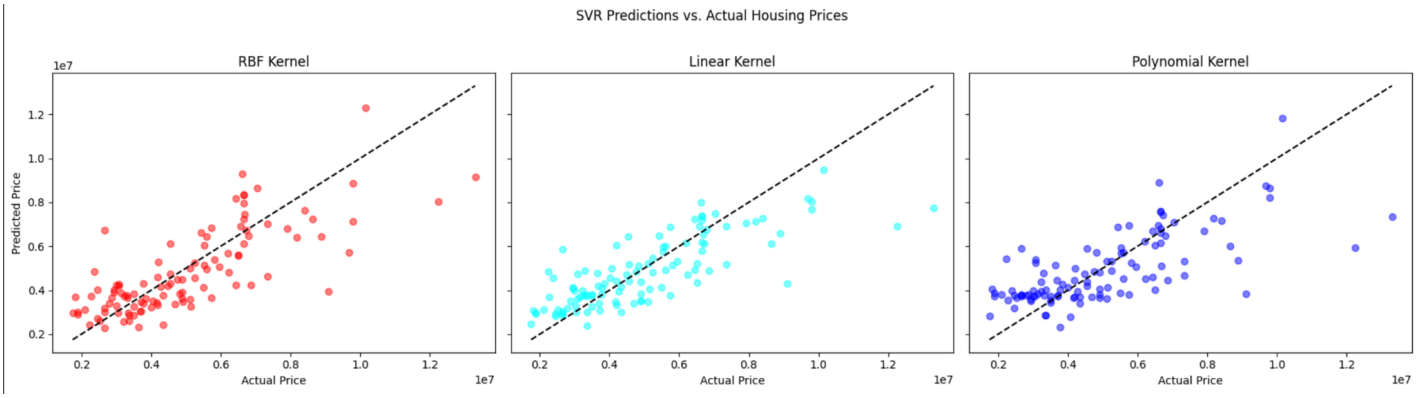
Problem 2

1. **Introduction:**
   In this assignment, an SVR model is developed to predict housing prices based on various features, including area, bedrooms, bathrooms, stories, and amenities like guestrooms and air conditioning. The objectives are to explore different kernel functions within the SVR model, apply PCA for optimal feature dimensionality, and compare SVR performance with a regularized linear regression model created in Homework 1.

2. **Methodology:**
   The dataset was standardized, and PCA was applied to reduce dimensionality by varying the number of components (K) from 1 to the maximum feature count. For each K, the data was split 70% for training and 30% for testing. Three SVR models were developed using linear, RBF, and polynomial kernels, with MSE recorded for each. The results of the SVR models were compared against a linear regression model with regularization from Homework 1 for baseline performance evaluation.

3. **Results:**
   The performance of the SVR models was evaluated using MSE as a metric. Among the different kernels tested, the linear kernel achieved the lowest MSE, indicating that it provided the most accurate predictions of housing prices. The RBF kernel followed closely, while the polynomial kernel had the highest MSE, suggesting it may not have captured the underlying relationships in the data as effectively as the other models.

   - **RBF Kernel MSE:** 0.0160
   - **Linear Kernel MSE:** 0.0145
   - **Polynomial Kernel MSE:** 0.0190



4. **Summary of Findings**

   The analysis aimed to predict housing prices using SVR with various kernels and PCA for dimensionality reduction. Key findings include:

   - **Kernel Performance:** The linear kernel demonstrated the best performance with the lowest MSE of 0.0145, indicating its effectiveness in capturing the linear relationships in the data.
   - **Non-linear Models:** The RBF kernel achieved a competitive MSE of 0.0160, suggesting it effectively modeled some non-linearities in the dataset, but it was slightly less accurate than the linear kernel. The polynomial kernel performed the least well, with an MSE of 0.0190, indicating it may have overfitted or inadequately represented the data patterns.
   - **PCA Impact:** Applying PCA allowed for a reduction in dimensionality while retaining significant variance, which improved model training efficiency and performance.

   Overall, the findings highlight the effectiveness of linear SVR in this context, while also demonstrating the potential of RBF kernels for more complex data relationships. The polynomial kernel, despite its theoretical advantages, may require further tuning or feature engineering to enhance its predictive accuracy.

5. **Comparison with Homework 1**

   In Homework 1, a linear regression model was utilized, achieving a minimum cost of 0.73846424 with a learning rate of 0.10. This analysis focused on optimizing parameters across three explanatory variables (X1, X2, and X3), highlighting the effects of different learning rates on convergence.

   In contrast, Homework 4 employed SVR, testing various kernels—linear, RBF, and polynomial—along with dimensionality reduction through PCA. This approach yielded lower MSEs, notably 0.0145 for the linear kernel, showcasing enhanced predictive accuracy over the linear regression model from Homework 1. The integration of PCA allowed for effective feature extraction, improving model performance while reducing complexity. Overall, Homework 2 not only provided deeper insights into the data relationships but also demonstrated significant advancements in predictive capabilities compared to the first assignment.