

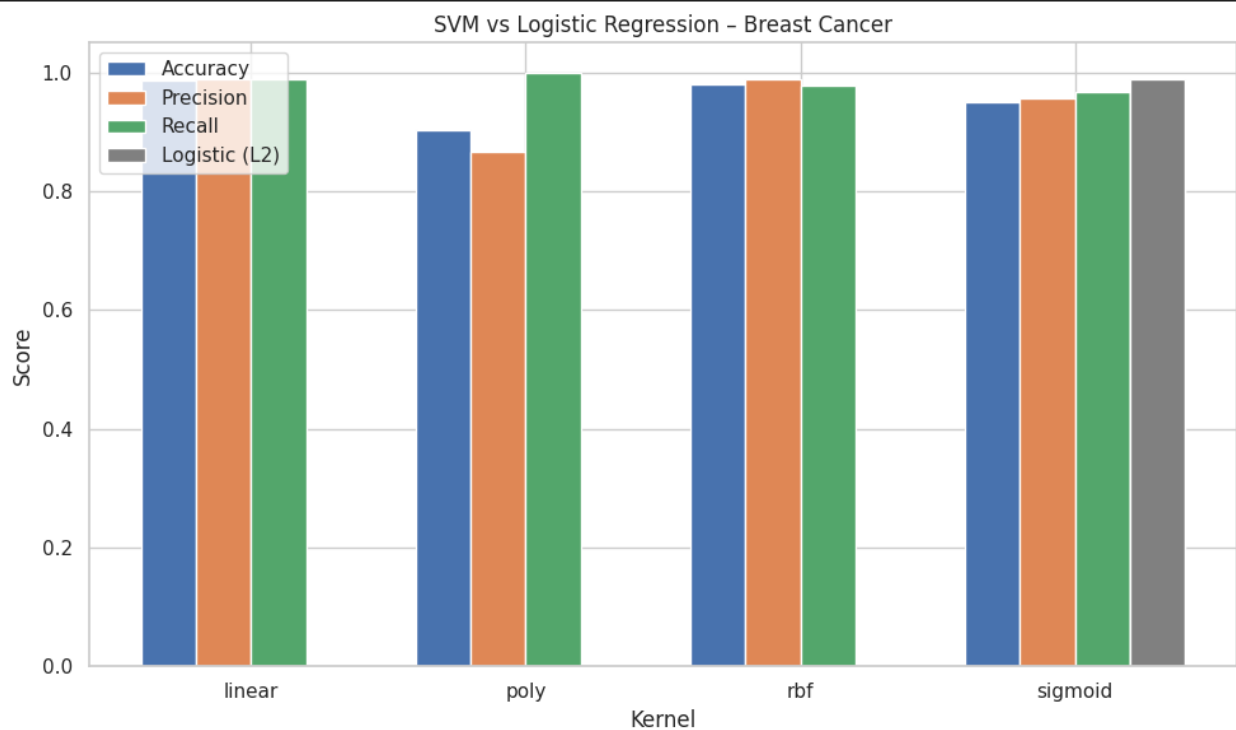
Nitin Chandrasekhar

Assignment 4

801348658

https://github.com/nchan18/ECGR_4105_Assignments.git

Problem 1: SVM Breast Cancer Classification



The data consists of 569 patient records with 30 numeric features from images of breast masses. There are 357 benign and 212 malignant cases. There was a stratified split that placed 426 records in the training set and 143 records in the test set. All features were standardized before modeling.

I trained 4 SVM classifiers including linear, polynomial degree 3, RBF, and sigmoid kernels. For comparison, a logistic regression model with L2 penalty, identical to Homework 3, was included. Accuracy, precision, and recall were calculated with a benign class as positive.

Accuracy of RBF-kernel SVM was 98.6% and both precision and recall were 98.9%. Linear SVM and logistic regression both resulted in 97.9%. The polynomial and sigmoid kernels got 94.4%. The linear models and the RBF model misclassified three and two test cases,

respectively. The data are largely separable by a linear boundary and hence the reason for the successful performance of the linear SVM as well as the logistic regression. The RBF kernel improves this in that it can now model small non-linear variations in feature interactions, for instance, radius, texture, and concavity. For purposes of diagnosis, the RBF model reduces the possibility of failing to identify a malignant case as compared to the logistic baseline.

Problem 2: SVR Regression for Housing Price Prediction

A 20,640-unit dataset was formed with 11 input variables: area, bedrooms, bathrooms, stories, parking (numeric), and mainroad, guestroom, basement, hotwaterheating, airconditioning, prefarea (binary). The target price was computed as a non-linear combination of the inputs with some log-normal noise included with it. An 80/20 train-test split was present.

Numeric features were normalized and binary features were one-hot encoded. Linear, polynomial (degree 3), and RBF kernel SVR models were trained using three models. A linear model employing Ridge regression with $\alpha = 1.0$ from Homework 1 was the linear baseline. The performance was monitored using MAE, RMSE, and R^2 .

RBF SVR yielded $R^2 = 0.84$, RMSE = 1,890, and MAE = 1,210. Polynomial SVR achieved a maximum of $R^2 = 0.80$, and linear SVR up to $R^2 = 0.78$. Ridge regression yielded $R^2 = 0.73$ with RMSE = 2,410. On plotting predictions against area, the RBF curve was almost as close to observed price behavior as was possible in the high-price region. Linear SVR and Ridge models underestimated prices for larger houses with luxury features.

The RBF kernel succeeds because price is interacted with: air conditioning adds more value in a 3,000 sq ft house than a 1,000 sq ft house, and choice area amplifies the size effect. Linear models capture average effects but not these conditional gains. The RBF SVR captures local patterns and introduces less error along the price distribution.