ECGR 5105 Homework 4

Owen Bailey-Waltz (801488178)

GitHub Link: https://github.com/obaileyw-uncc/ecgr5105/hw04_supportvectormachines

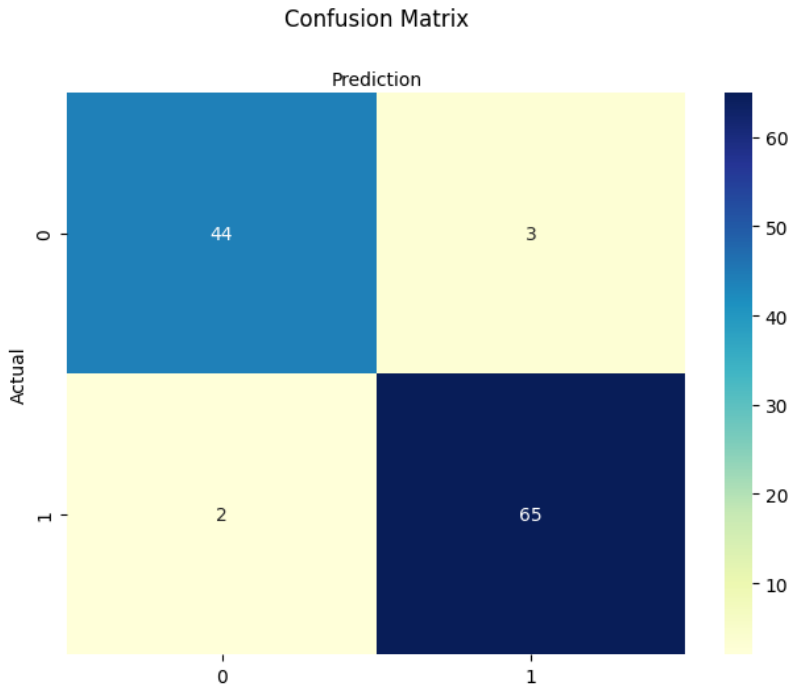
Problem 1: Breast cancer dataset

Linear kernel

Support vector regression

$$h(\boldsymbol{x}) = g(-0.486 + 2.367x_1 - 0.072x_2 + 0.531x_3 - 0.581x_4 - 1.230x_5 + 2.987x_6 - 2.293x_7 - 1.120x_8 - 0.331x_9 - 0.017x_{10} - 1.309x_{11} + 0.463x_{12} + 1.676x_{13} - 2.780x_{14} - 0.283x_{15} - 1.692x_{16} + 2.059x_{17} - 2.108x_{18} + 0.093x_{19} + 4.659x_{20} - 1.989x_{21} - 1.306x_{22} - 1.544x_{23} - 3.587x_{24} + 0.626x_{25} + 1.6\,83x_{26} - 1.716x_{27} + 0.411x_{28} - 0.443x_{29} - 3.678x_{30})$$

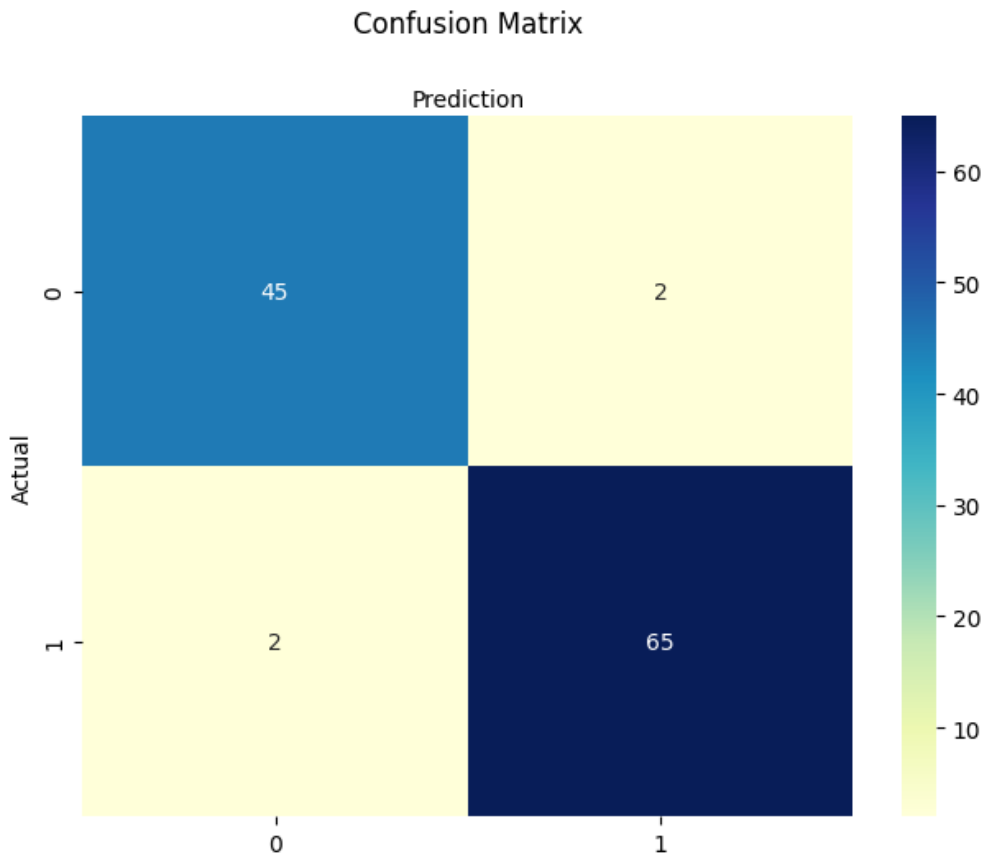
TRAINING ACCURACY 0.989010989010989
VALIDATION ACCURACY 0.956140350877193
PRECISION 0.9558823529411765
RECALL 0.9701492537313433
F1 SCORE 0.9629629629629629



Logistic regression from Homework 3 (no weight penalty)

$$h(\boldsymbol{x}) = g(0.248 - 0.351x_1 - 0.489x_2 - 0.341x_3 - 0.408x_4 - 0.192x_5 + 0.448x_6 - 0.669x_7 + 0.845x_8 - 0.338x_9 + 0.213x_{10} - 1.391x_{11} + 0.039x_{12} - 0.857x_{13} - 0.971x_{14} + 0.251x_{15} + 0.667x_{16} + 0.121x_{17} - 0.222x_{18} + 0.120x_{19} + 0.865x_{20} - 0.932x_{21} - 1.041x_{22} - 0.767x_{23} - 0.890x_{24} - 0.536x_{25} - 0.020x_{26} - 0.870x_{27} - 0.975x_{28} - 0.515x_{29} - 0.611x_{30})$$

TRAINING ACCURACY 0.9890
VALIDATION ACCURACY 0.9649
PRECISION 0.9701
RECALL 0.9701
F1 SCORE 0.9701



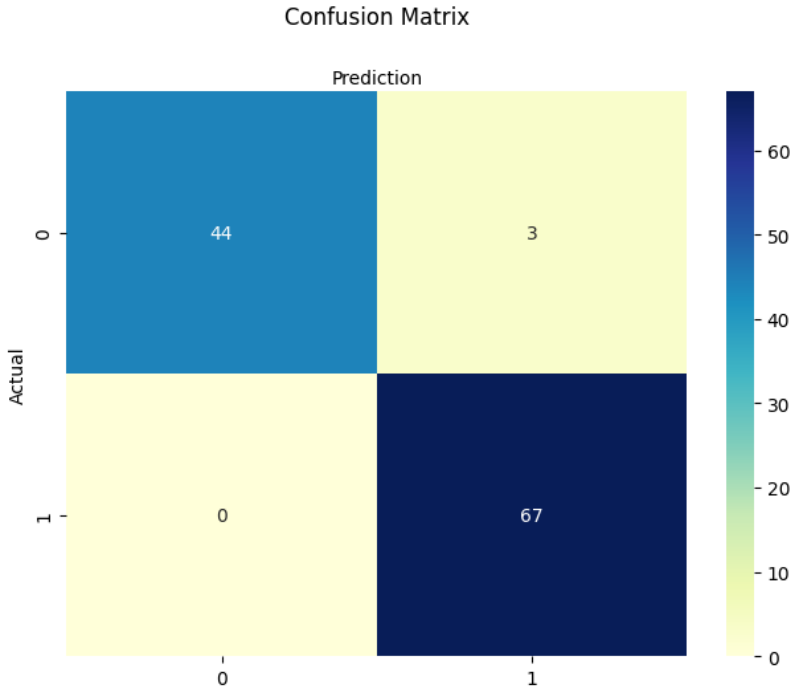
Polynomial kernel

Coefficients are not extractable for kernels other than linear.

TRAINING ACCURACY 0.9868131868131869
VALIDATION ACCURACY 0.9736842105263158
PRECISION 0.9571428571428572

RECALL 1.0

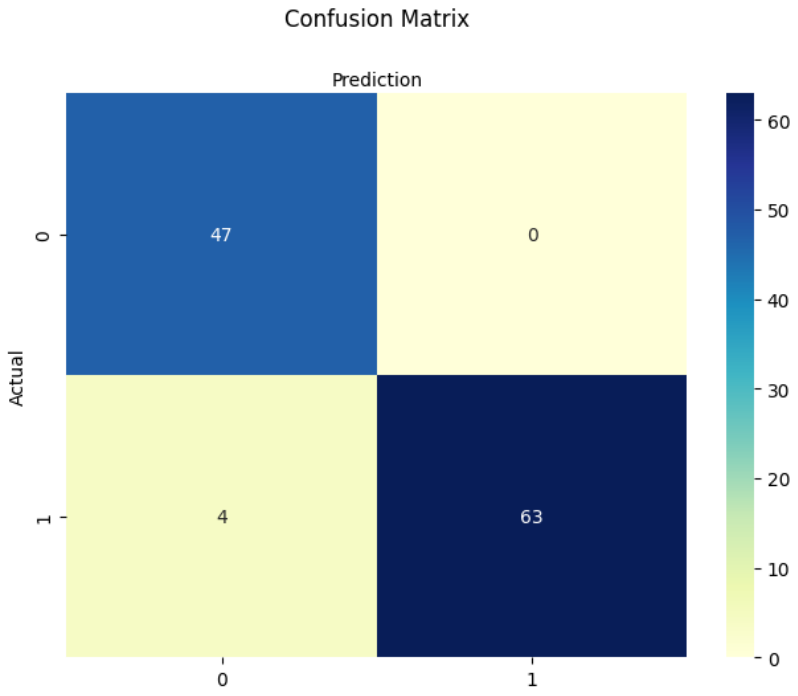
F1 SCORE 0.9781021897810219



Radial basis function kernel

Coefficients are not extractable for kernels other than linear.

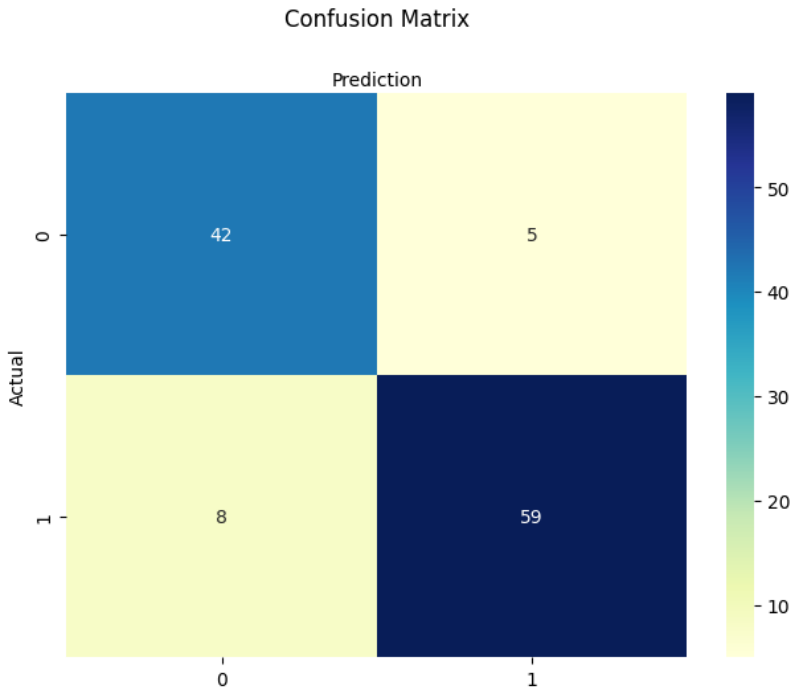
TRAINING ACCURACY 0.9956043956043956
VALIDATION ACCURACY 0.9649122807017544
PRECISION 1.0
RECALL 0.9402985074626866
F1 SCORE 0.9692307692307692



Sigmoid kernel

Coefficients are not extractable for kernels other than linear.

TRAINING ACCURACY 0.9252747252747253
VALIDATION ACCURACY 0.8859649122807017
PRECISION 0.921875
RECALL 0.8805970149253731
F1 SCORE 0.9007633587786259



Reflections

Based on the F_1 score, the best fit achieved from this dataset comes from the polynomial kernel function, which achieved roughly a 97.8% F_1 after optimizing the error parameter. The next best fit in this case came from the RBF, which achieved a 96.9% fit. The linear regression fit followed at 96.3% and the sigmoid kernel only achieved 90.1% F_1.

Comparing the linear kernel SVM with the linear logistic regression classifier from the previous homework reveals that the optimal SVM solution produced a roughly equivalent model to the optimal gradient descent solution, with only one wrong classification more in the SVM solution leading to a slightly lower precision value than gradient descent with no weight penalty.

Problem 2: Housing dataset

Linear kernel

Support vector regression

$$h(\boldsymbol{x}) = 4709918.122 + 527776.788x_1 + 86534.126x_2 + 441136.556x_3 + 444871.952x_4 + 130501.495x_5 + 164751.775x_6 + 112646.049x_7 + 185023.418x_8 + 342796.483x_9 + 164643.058x_{10} + 318915.305x_{11}$$

$$J(\boldsymbol{\Theta}) = 6.149 \times 10^{11} \qquad J_{\text{val}}(\boldsymbol{\Theta}) = 4.720 \times 10^{11}$$

TRAINING R^2 0.6625543331345585
VALIDATION R^2 0.6724225377064308

Standardized model from Homework 2

$$h(\boldsymbol{X}) = 4749014 + 534348x_1 + 97970x_2 + 431243x_3 + 428637x_4 + 130230x_5 + 162189x_6 + 177745x_7 + 206256x_8 + 419006x_9 + 266853x_{10} + 282602x_{11}$$

$$J(\boldsymbol{\Theta}) = 5.459 \times 10^{11} \qquad J_{\text{val}}(\boldsymbol{\Theta}) = 6.871 \times 10^{11}$$

Polynomial kernel

Coefficients are not extractable for kernels other than linear.

$$J(\boldsymbol{\Theta}) = 7.196 \times 10^{11} \qquad J_{\text{val}}(\boldsymbol{\Theta}) = 7.296 \times 10^{11}$$

TRAINING R^2 0.6050643907278781
VALIDATION R^2 0.4935883588772938

Radial basis function kernel

Coefficients are not extractable for kernels other than linear.

$$J(\boldsymbol{\Theta}) = 3.920 \times 10^{11} \qquad J_{\text{val}}(\boldsymbol{\Theta}) = 6.326 \times 10^{11}$$

TRAINING R^2 0.7848586847431868
VALIDATION R^2 0.5609442218228046

Sigmoid kernel

Coefficients are not extractable for kernels other than linear.

$$J(\boldsymbol{\Theta}) = 7.157 \times 10^{11} \qquad J_{\text{val}}(\boldsymbol{\Theta}) = 5.486 \times 10^{11}$$

TRAINING R^2 0.6071944779624667
VALIDATION R^2 0.6192539162620936

Reflections

The `score()` function for the SVR API in SciKit Learn gives the R^2 value for model quality as opposed to the loss function value, so validation loss was computed using the function created in earlier homeworks. Well-fitting models have a validation R^2 closer to 1 and a low validation loss. Based on these metrics, the best fit for this model comes from the linear kernel support vector regression. The next best fit came from the sigmoid kernel, followed by the RBF kernel and the polynomial kernel.

Comparing the linear kernel SVR with the linear logistic regression classifier from the previous homework reveals that the optimal SVR solution produced a better fitting model than the optimal gradient descent solution. The validation loss in the SVR case was 4.720×10^{11} while the validation loss for the gradient descent case was 6.871×10^{11}. This lower validation loss level indicates better generalization and therefore a higher-quality model with SVR when compared to gradient descent.