Small Project 1 - Logistic Regression vs SVM

Jacob D. Evans

University of West Georgia

CS 4275 Machine Learning Foundations

Dr. Md Shirajum Munir

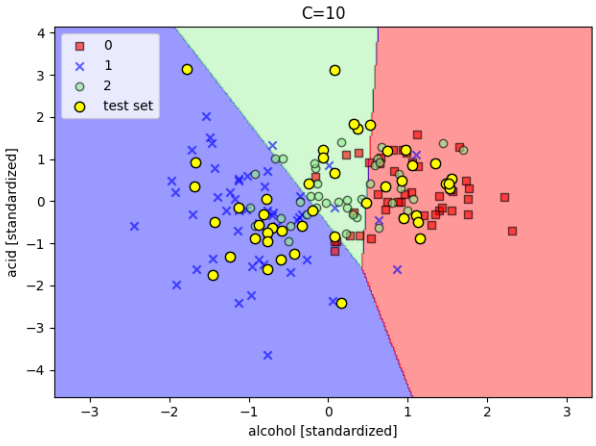
September 24, 2024

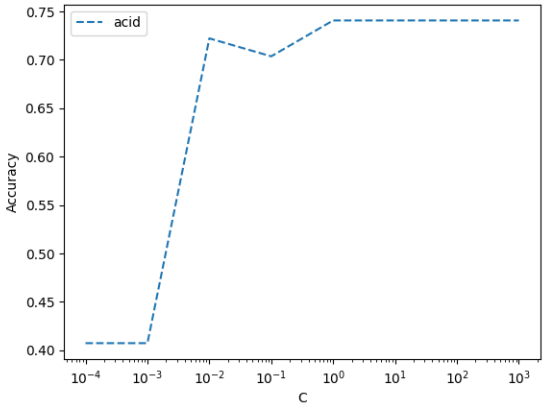
Abstract

In this report, Logistic Regression and Support Vector Machines (SVM) are compared using the wine dataset. We start by fitting and testing Logistic Regression with different values of the parameter C and examine how it affects decision boundaries and accuracy. We used values of 10, 100, 1000 and 5000 for C in our testing. Then, we do the same for SVM with an RBF kernel, testing two different gamma values. (gamma values 0.1 and 10) This comparison helps us understand which algorithm works better for this dataset and why. The following figures are generated to help display the decision regions and accuracy for each algorithm and C value.

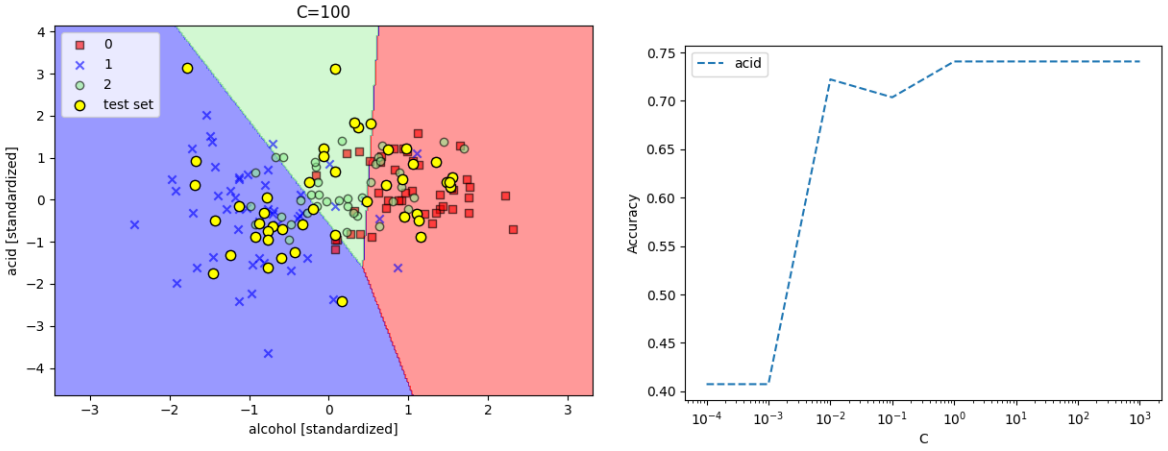
Logistic Regression

C=10 decision regions and Accuracy

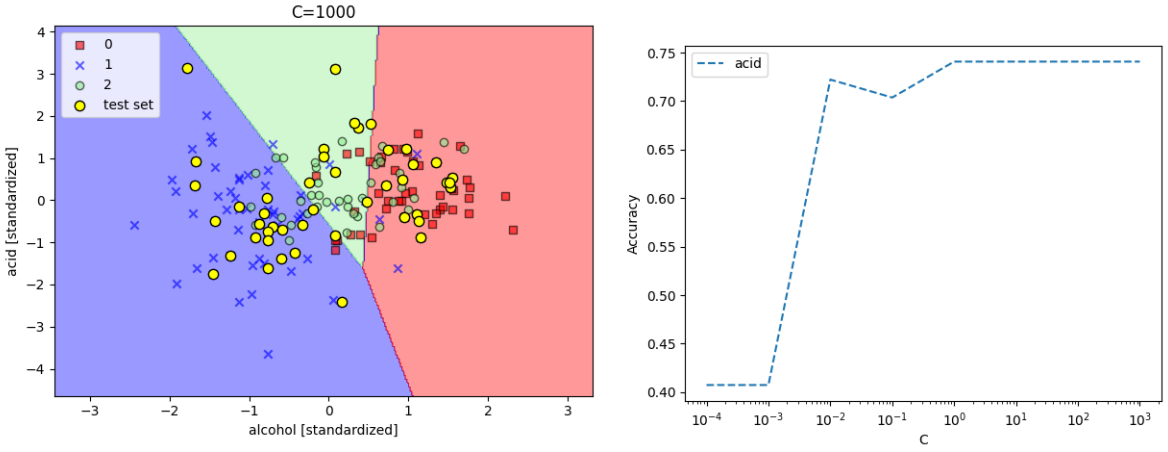




C=100 decision regions and Accuracy

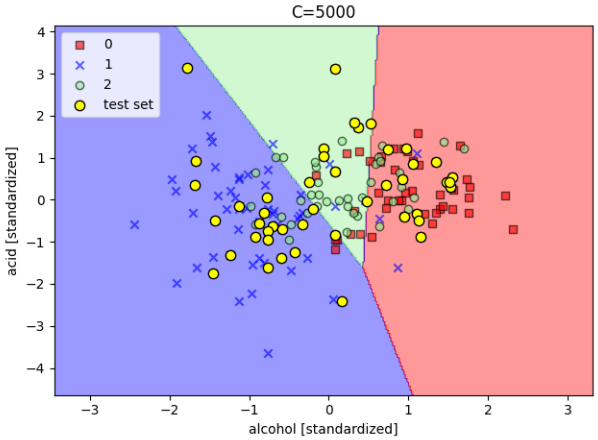


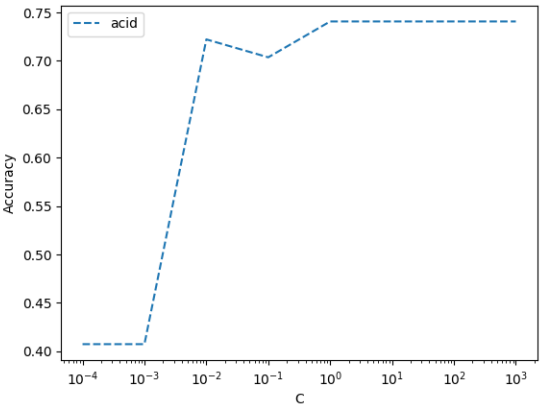
C=1000 decision regions and Accuracy



As the C value increases, the decision boundary becomes more defined, but there's a limit to how much improvement we see.

C=5000 decision regions and Accuracy



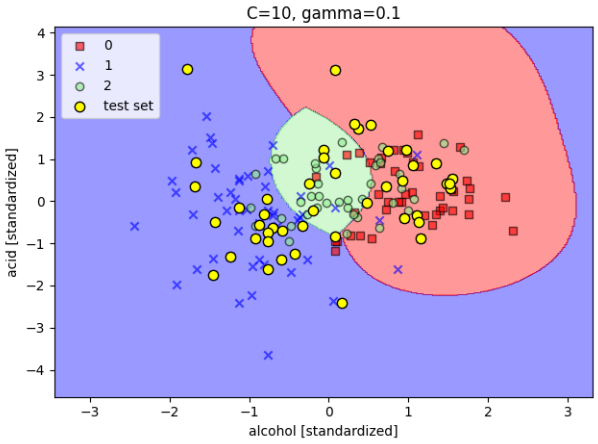


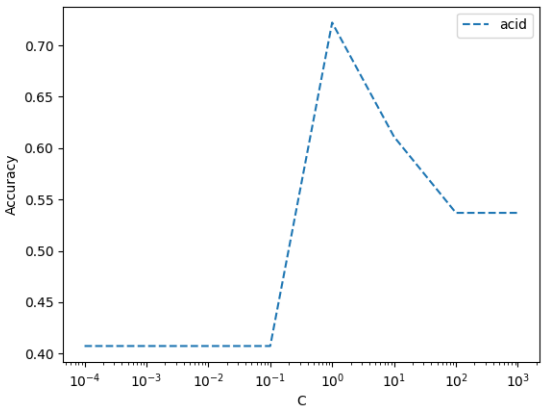
At C=5000, we barely see any improvement from the previous iteration.

SVM

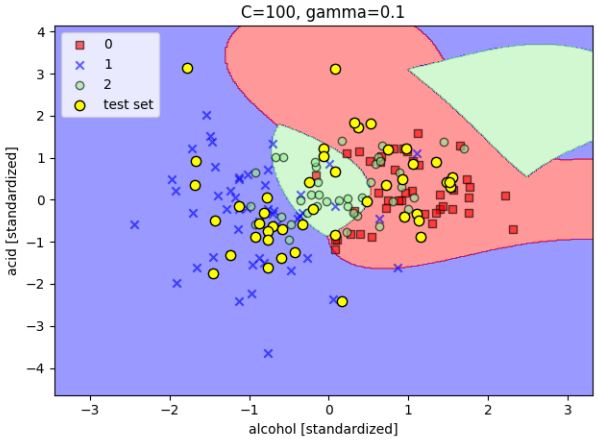
Now let us look at how an SVM model will react to the wine dataset. The following results are split into two gamma values, 0.1 and 10.

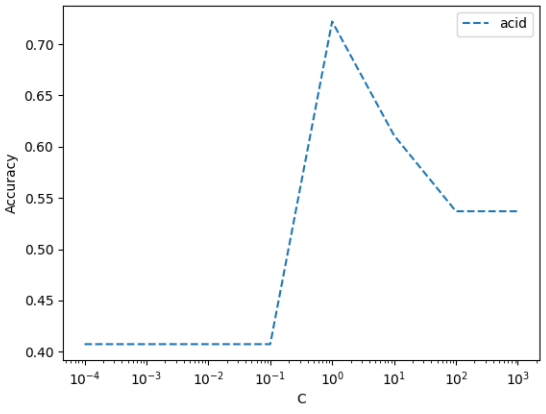
C=10 decision regions and Accuracy (gamma = 0.1)



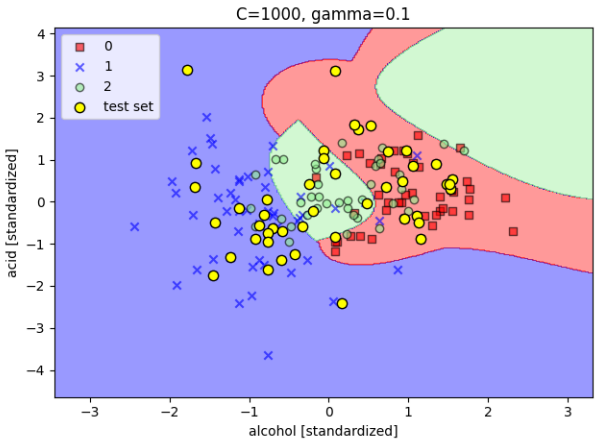


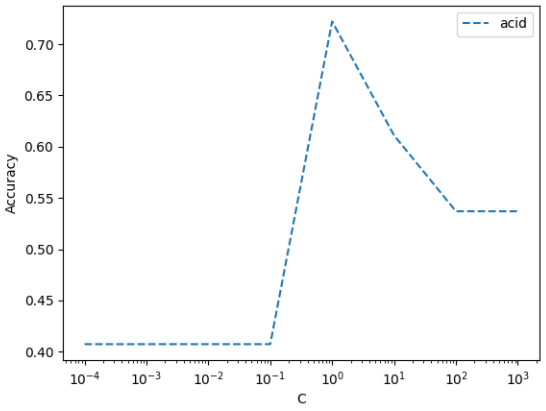
C=100 decision regions and Accuracy (gamma = 0.1)



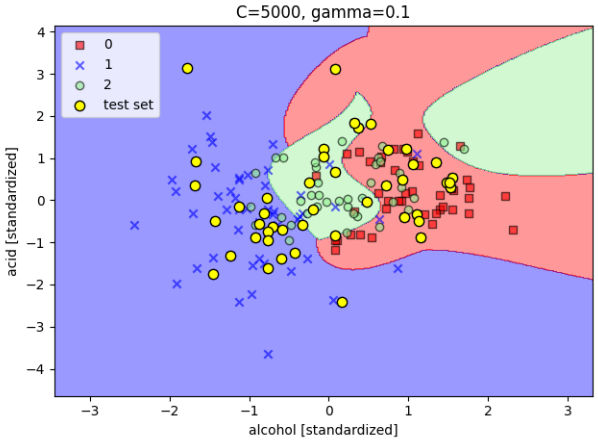


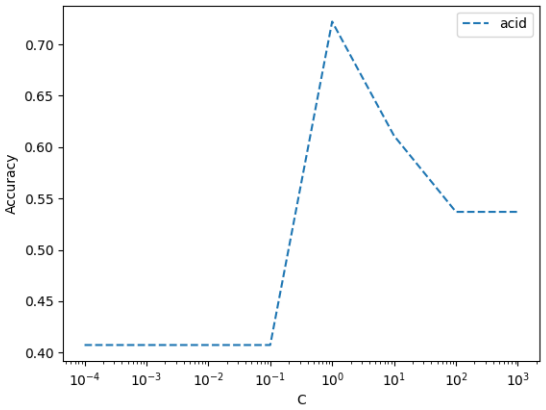
C=1000 decision regions and Accuracy (gamma = 0.1)



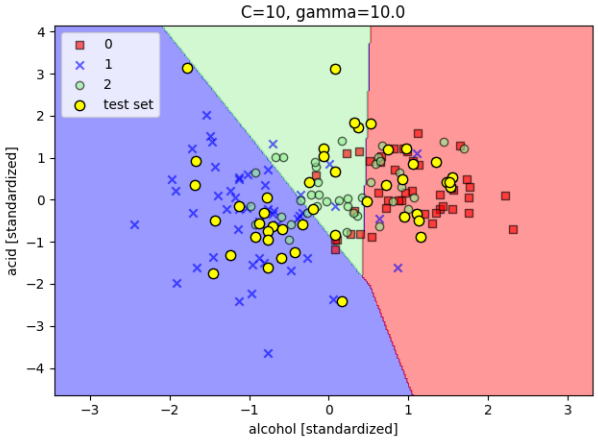


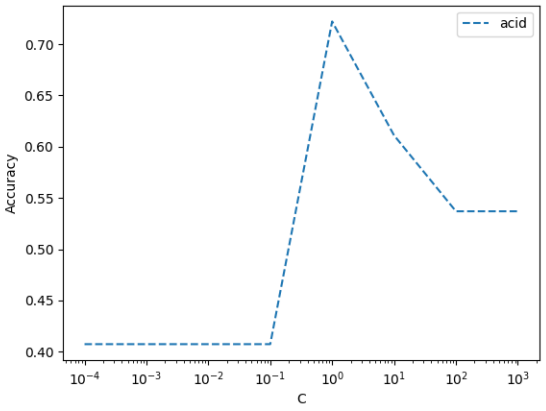
C=5000 decision regions and Accuracy (gamma = 0.1)



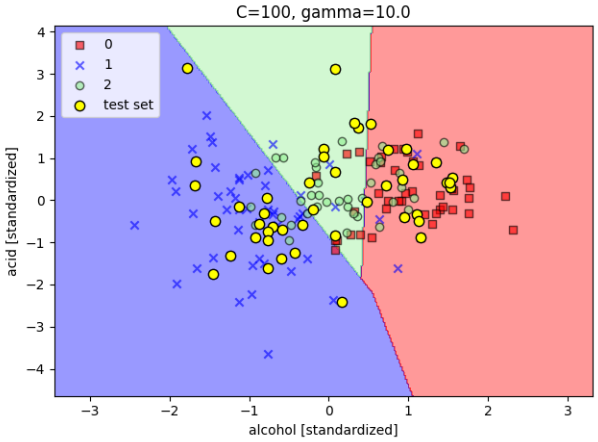


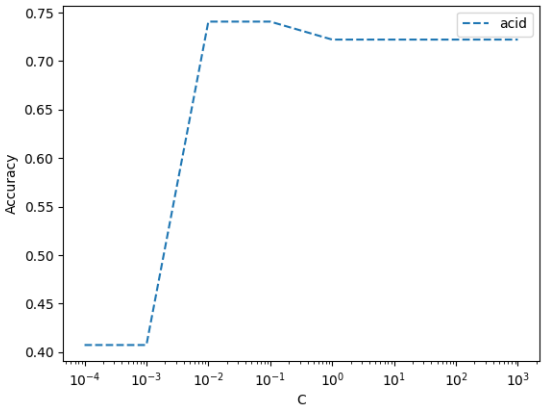
C=10 decision regions and Accuracy (gamma = 10)



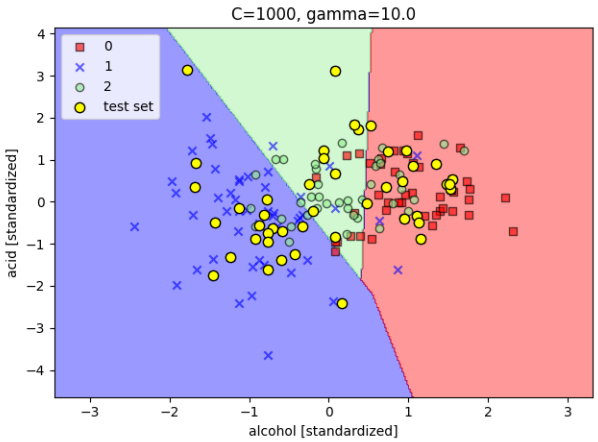


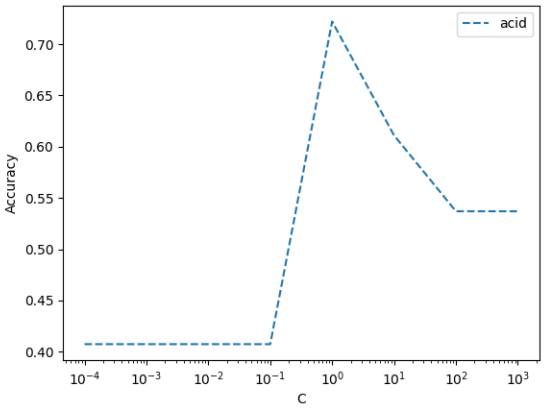
C=100 decision regions and Accuracy (gamma = 10)



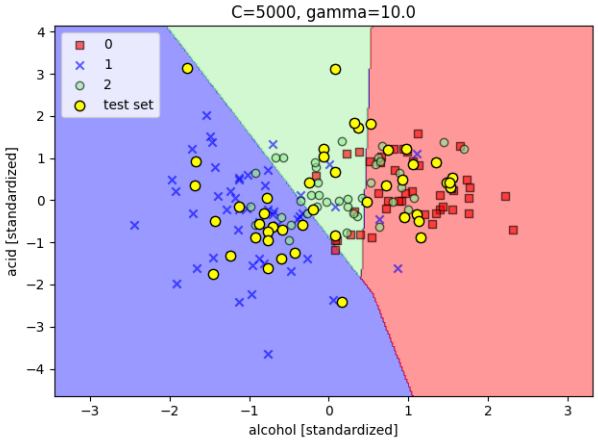


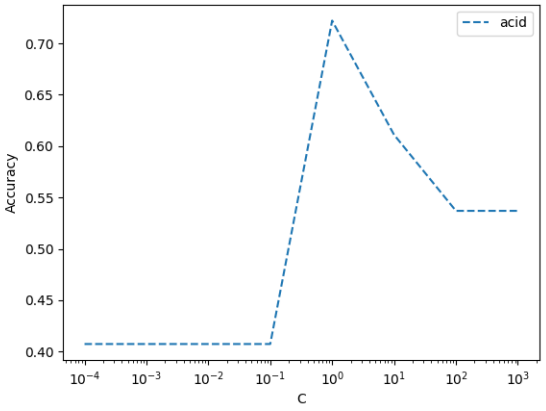
C=1000 decision regions and Accuracy (gamma = 10)





C=5000 decision regions and Accuracy (gamma = 10)





When looking at the two gamma values for SVM, we can see some important differences. With gamma = 0.1, the decision boundaries are smoother and more general because each point affects a larger area. This makes the model simpler but can miss important details in the data. With gamma = 10, the boundaries are much tighter, meaning the model pays more attention to each

point, capturing more complex patterns. However, this can make the model overfit by focusing too much on the specific data points. Overall, gamma = 0.1 is more general, while gamma = 10 is more detailed but riskier.

Results

Both Logistic Regression and SVM performed well on the wine dataset, but each has its strengths. Logistic Regression gave smoother decision boundaries as C increased, but didn't improve much after a point. SVM, with its RBF kernel, handled non-linear patterns better, especially with higher gamma values, making it more flexible. While Logistic Regression is easier to understand, SVM can give better accuracy when tuned correctly. The best choice depends on how complex the data is and whether accuracy or simplicity is more important.