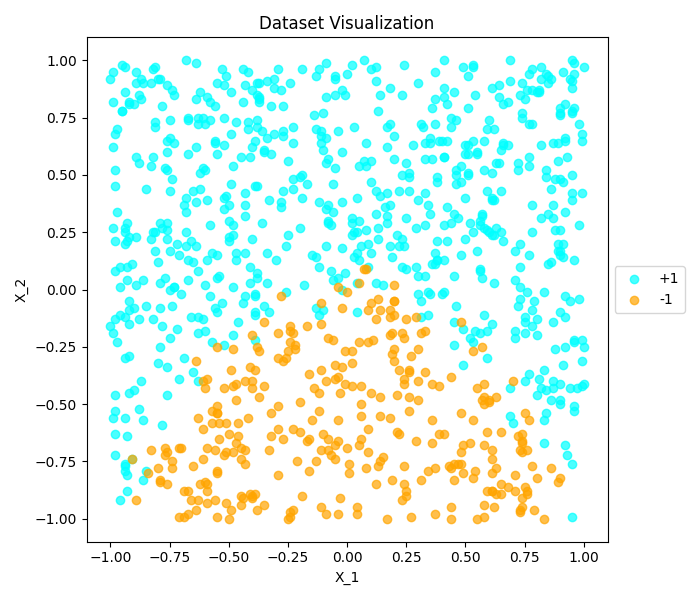
**CS7CS4 Machine Learning**

**The assigned dataset is: # id:16-16-16**

**Part A**

(i) The provided dataset consists of two numerical features (dubbed and ) and a binary target label (that was -1 or +1). The features were plotted into a scatter plot, with the X-axis representing the feature and the Y-axis representing the feature. Data points with label as +1 are represented using cyan circles, while those with label as -1 as orange circles for clarity. Here is the graph produced:



*Figure 1. Scatter plot of the given data*

The code segment that produced this is present under the plot\_given\_data() function.

(ii) The Logistic Regression model was trained using both the given features, which computes the probability of a point belonging to the +1 class using the sigmoid function and the -1 class’s probability would be , where represents the linear combination of the features.

After fitting, the feature coefficients that were produced were and , while the intercept is . The accuracy of this model compared to the actual data produced a score of 0.87888 i.e. approximately 87.9%. The interpretation of the coefficients is as follows:

* is negative but small, which implies that increasing feature would slightly decrease the probability of predicting +1.
* is positive and larger, which implies that increasing feature would increase the probability of predicting +1.
* As such, since , feature would have the most influence on prediction.

This analysis aligns with how the data is structured in Figure 1, where the data points appear to be vertically segregated into two distinct regions.

The code segment that produced this is present under the train\_log\_regr() and make\_predictions() functions.

(iii) The trained model’s predictions were plotted on top of the scatter plot produced by the provided data, with the predicted +1 labels appearing as red ‘+’ markers and the predicted -1 labels appearing as green ‘x’ markers to display a comparison between the two:



*Figure 2. Scatter plot representing given data and predictions of Logistic Regression model*

As mentioned in Part A (ii), the tendency of the predicted data showed an 87.9% overlap with the data points from the given dataset, with the predicted data being more distinctively separated into two sections thanks to the decision boundary (represented by the purple line).

The decision boundary was derived from the logistic regression parameters based on the formula mentioned in Part A (ii): by setting , which then allows us to solve for , which helps us define the decision boundary i.e. line that separates the +1 and -1 classes on the graph shown in Figure 2.

The code segment that produced this is present under the plot\_log\_regr\_predictions() function.

(iv) From Figure 2 and accuracy score of 87.9%, it becomes apparent that the predicted labels align closely with the true labels. The majority of the +1 and -1 data points lie on the correct sides of the decision boundary, with the few miscalculations on either side very close to the boundary line. My observation is that this method relied primarily on the feature on the vertical axis to separate the two classes, and since only two features were involved in modelling the decision process, the separation line acts as a proper split as compared to the gaussian separation of the given data set. Overall, I would say that the Logistic Regression model provided a good fit for the data with a clear linear decision boundary.

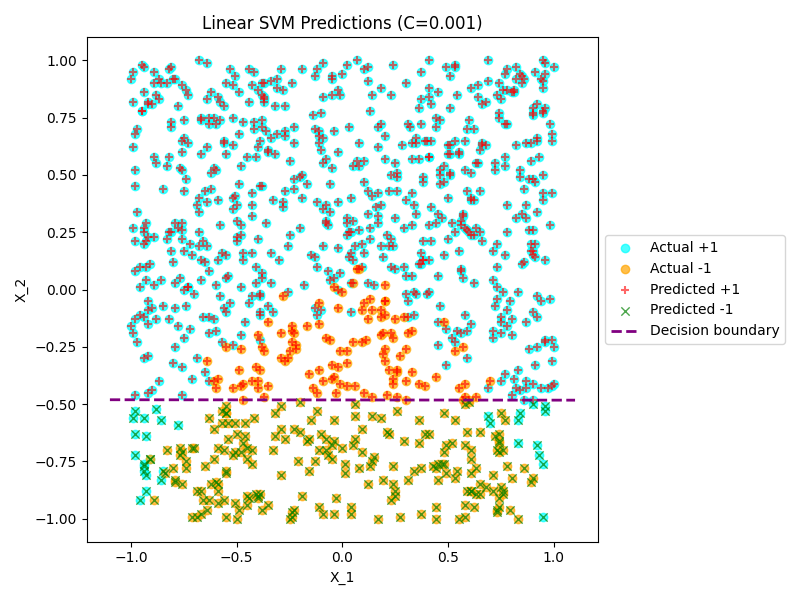
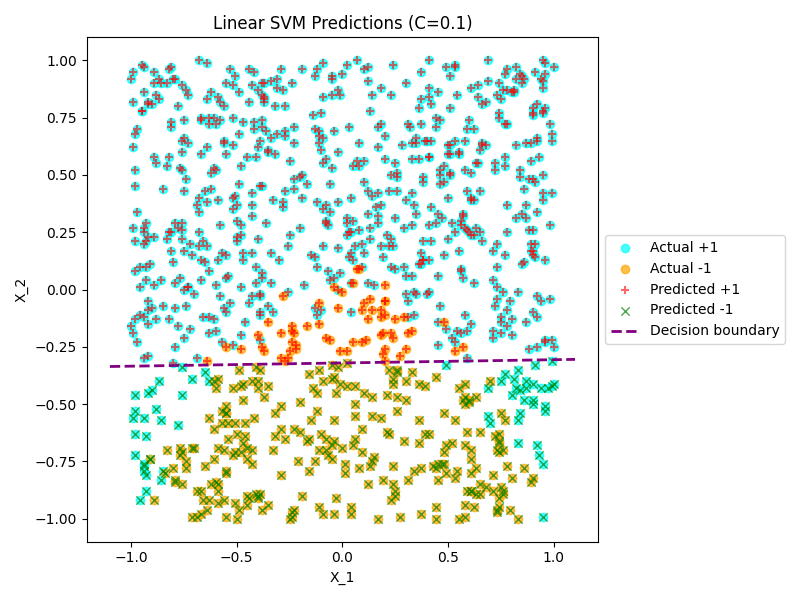
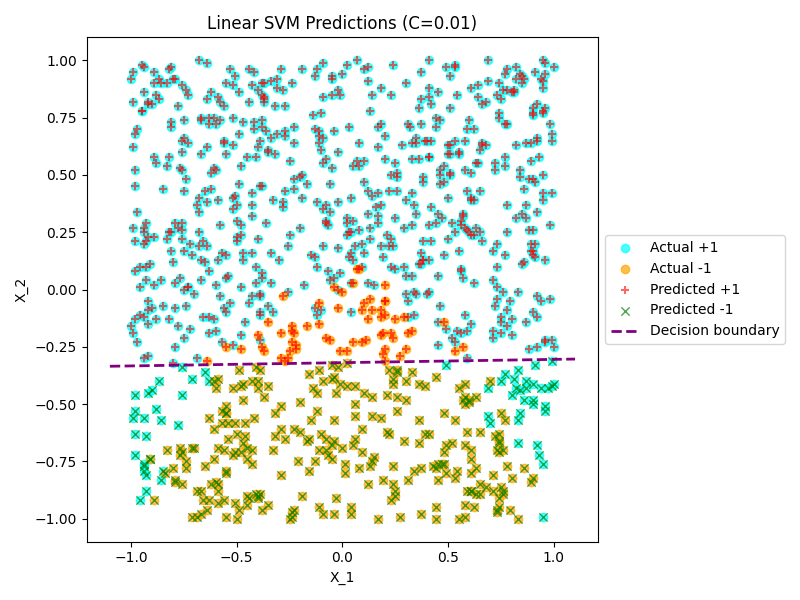
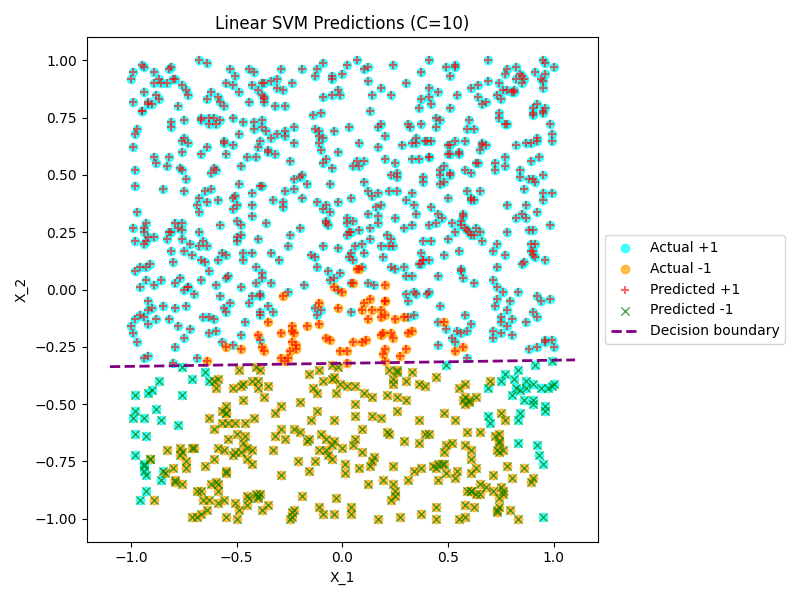
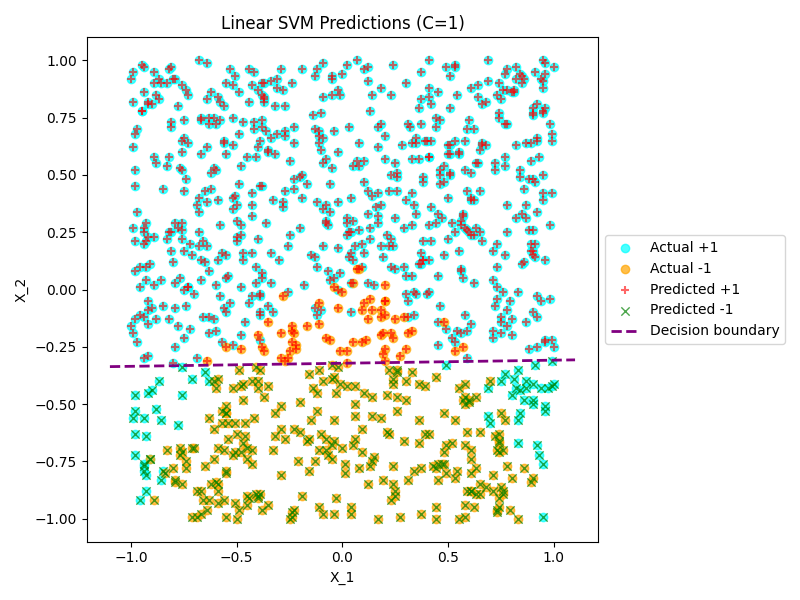
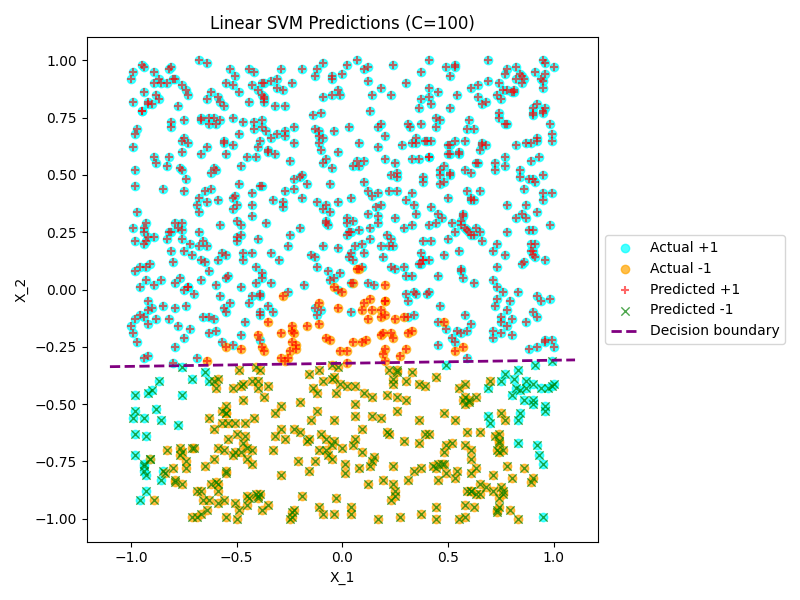
**Part B**

(i) I trained a total of 7 Support Vector Machine classifiers on the same dataset, with the weighting parameter having values of 0.001, 0.01, 0.1, 1, 10, 100, 1000.

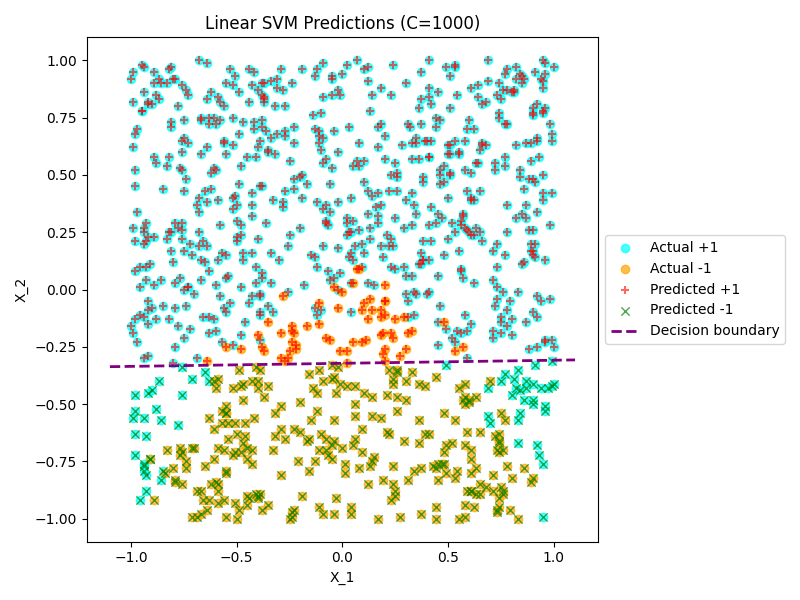
The decision function for just two features (and ) can be computed as , where the weights learned by the SVM are and , while represents the intercept or bias, and each SVM uses this to predict the class of the data point based on the sign of the calculated . The resulting parameters for all SVMs are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| C |  |  | bias | accuracy |
| 0.001 | 0.00040 | 0.48143 | 0.23188 | 0.85686 |
| 0.01 | -0.01701 | 1.19233 | 0.38027 | 0.87888 |
| 0.1 | -0.02430 | 1.72052 | 0.55092 | 0.87888 |
| 1 | -0.02498 | 1.86583 | 0.59995 | 0.87788 |
| 10 | -0.02524 | 1.88392 | 0.60616 | 0.87788 |
| 100 | -0.02527 | 1.88578 | 0.60679 | 0.87788 |
| 1000 | -0.02527 | 1.88597 | 0.60685 | 0.87788 |

(ii) The graphs generated using these are as follows:

    *Figures 3-8. Scatter plot representing given data and predictions of Support Vector Machine for C = 0.001, 0.01, 0.1, 1, 10, 100*

The code segment that produced Part B (i) is present under the train\_linear\_svm() function, and the one that produced these graphs is present under the plot\_svm\_predictions() function.



*Figure 9. Scatter plot representing given data and predictions of Support Vector Machine for C = 1000*

(iii) The weighting parameter is used in the cost function , where represents the regularization, that describes an inverse relationship where bigger makes the penalty less important, and as a result SVM prioritizes correctly classifying points, whereas a smaller would increase the regularization and result in a wider margin for potential misclassifications.

My interpretation of the table in Part B (i) and Figures 3-9 is as follows:

* Smaller values such as 0.001 and 0.01 (0.1 as well, to an extent) show smaller weights associated with the features and the biases, signifying the decision boundary to be shallower and some amount of misclassification. This is further represented by the accuracy of these models, especially to be lower than that of the other SVM models with higher values.
* Higher values ranging from 1 to 1000 display much less variance in the values of weights and stronger biases than when , and the accuracy scores also appear to converge into 0.87788 or approximately 87.9%. This shows how the parameters appear to stabilize after a certain point, indicating the data is well separated and further increasing has very little effect on how the SVM model computes its predictions.
* The decision boundary was computed using , which can be derived by setting , which then allows us to solve for . The data points that appear close to the decision boundary reflect how the change in variance in the different values can affect the predicted labels.
* Similar to the Logistic Regression model, , and once again feature seems to have a greater influence on prediction.
* To conclude, offers a unique influence in controlling the model’s complexity in an inverse relationship, where the more you increase it, the more the model minimizes misclassification and gives a more complex boundary.

(iv) Both the Logistic Regression model and the SVM models learn a linear decision boundary, and their orientation is very similar, especially for moderate to large values, as well as near identical accuracy, with Logistic Regression being incredibly slightly more accurate on my data.

Despite the SVM model showcasing smaller values for the weights associated with the features, as well as smaller intercepts than the Logistic Regression model, the trend of the weight associated with being larger and thus that feature having a greater influence on the prediction remained consistent.

As such, despite the numerical differences, both methods displayed consistency on this dataset for up to 87.9% accuracy, with Logistic Regression providing a probabilistic interpretation and SVM emphasizing margin-based decisions.

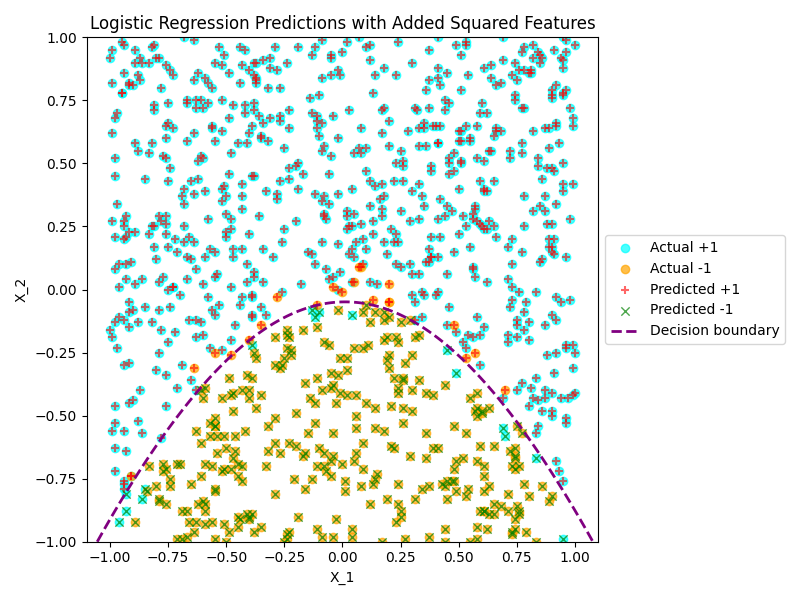
**Part C**

(i) The Logistic Regression model was trained with two additional features that were the squares of the original input features and to allow the model to learn using a quadratic combination of the features as

After fitting, the feature coefficients that were produced were , , and while the intercept is . The accuracy of this model compared to the actual data produced a score of 0.96296 i.e. approximately 96.3%.

The code segment that produced this is present under the train\_log\_regr\_with\_sq() function.

(ii) The trained model’s predictions were plotted on top of the scatter plot produced by the provided data, with the predicted +1 labels appearing as red ‘+’ markers and the predicted -1 labels appearing as green ‘x’ markers to display a comparison between the two:



*Figure 10. Scatter plot representing given data and predictions of Logistic Regression model involving squared features*

The code segment that produced this is present under the plot\_sq\_predictions() function.

The interpretation of the coefficients mentioned in Part C (i) is as follows:

* is negative but small, which implies that increasing feature would slightly decrease the probability of predicting +1.
* is positive and larger, which implies that increasing feature would increase the probability of predicting +1.
* is positive and larger, which implies that increasing feature would increase the probability of predicting +1.
* is negative but small, which implies that increasing feature would slightly decrease the probability of predicting +1.
* Based on the absolute value of the coefficients, feature would have the most influence on prediction, followed by

The key difference that the quadratic function introduced for the model to be more refined than the ones in Parts A and B was that the non-linear decision surface allows the classifier to better separate data that wasn’t perfectly linearly separable thanks to increased flexibility. This is also reflected in the values of the parameters associated with the weights being much larger, resulting in an improved accuracy score of 96.3% compared to the 87.9% of the previous models.

One key thing that I observed was how the squared features’ coefficients reflected an inverse relationship with their linear counterparts. However, it becomes more intuitive once visualizing them on the graph:

* Data points that are classified as -1 in the given data occur in a small region where and , and other points form the larger region of +1 labels.
* Points that are farther from the central point of are more likely to be classified as +1, which is reflected in the graph where the given data points at and are indeed labelled as +1.
* Points that are farther from the central point of are more likely to be classified as -1, which is reflected in the graph where the given data points at are mostly -1.

This also explains why we yield a curved decision boundary and gives model increased flexibility.

(iii) I coded a very trivial baseline predictor that always predicts the most common class in the training data, which was +1, and it produced an accuracy score of 0.68268 i.e. about 68.2%. This showed just how big a jump the current classifier provided in terms of accuracy, demonstrating how powerful it is to add non-linear features to a Logistic Regression model.

This code segment that produced this is present in the baseline\_accuracy() function

(iv) Using the formula stated in Part C (i), the decision boundary can be derived by setting to 0. We can solve for using the quadratic equation formula that would turn the equation into , which would give , but upon doing some research on how curves are plotted in matplotlib, I learned that it is done using contours by developing a mesh grid of the given data points and the process is more straightforward using inbuild functions.

This code segment is already present in plot\_sq\_predictions() function, and the decision boundary has been displayed in Figure 10 as a dashed purple curve.