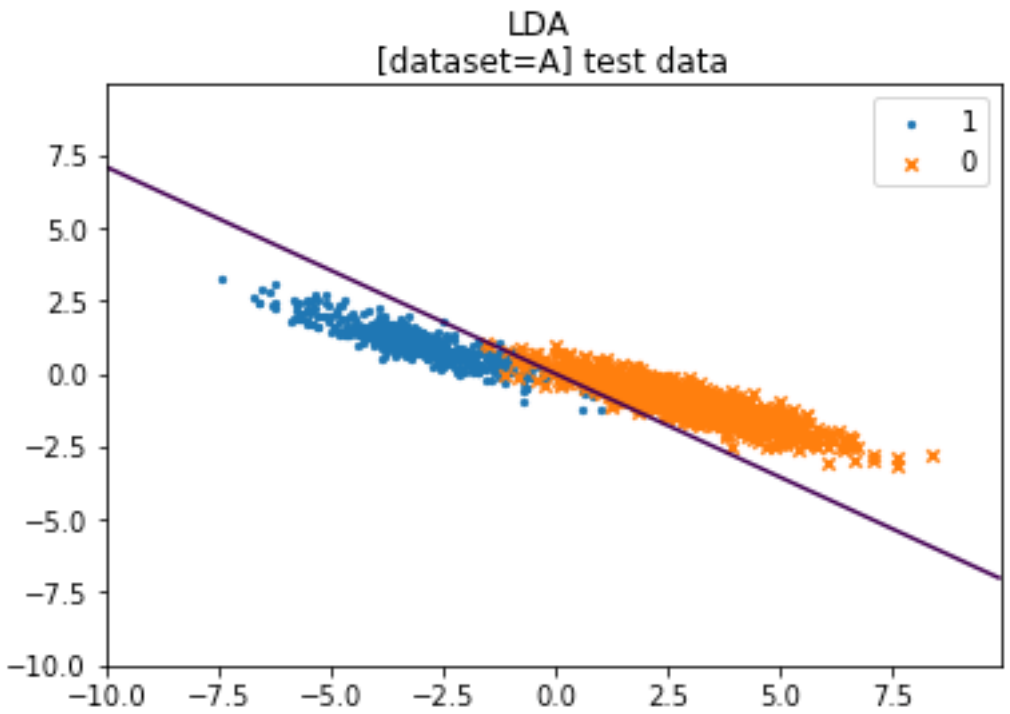
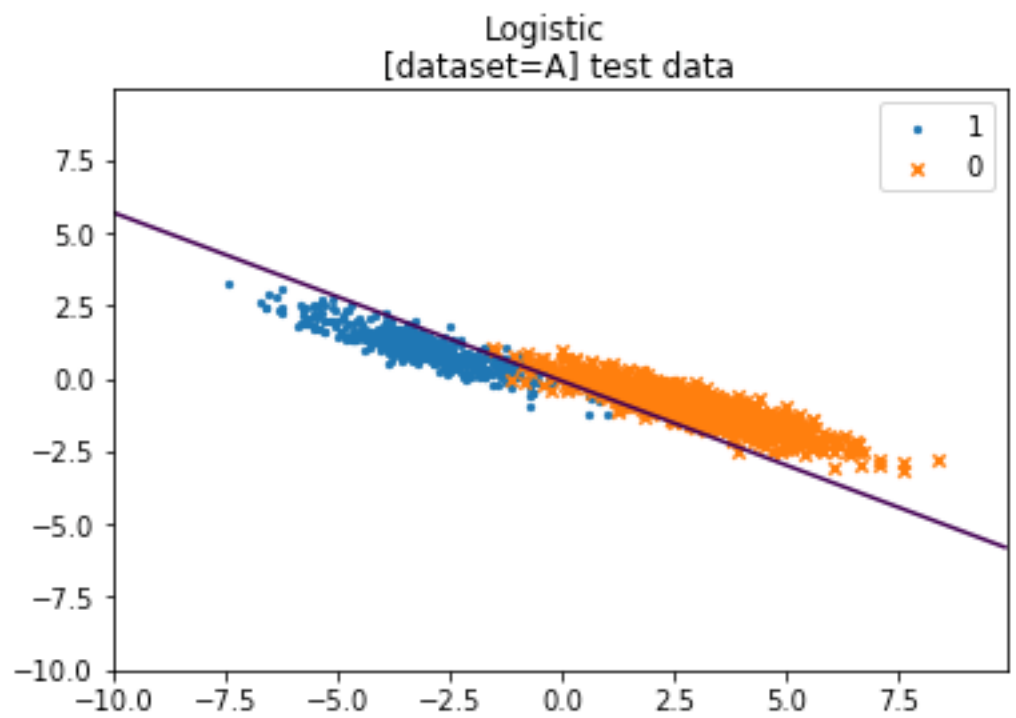
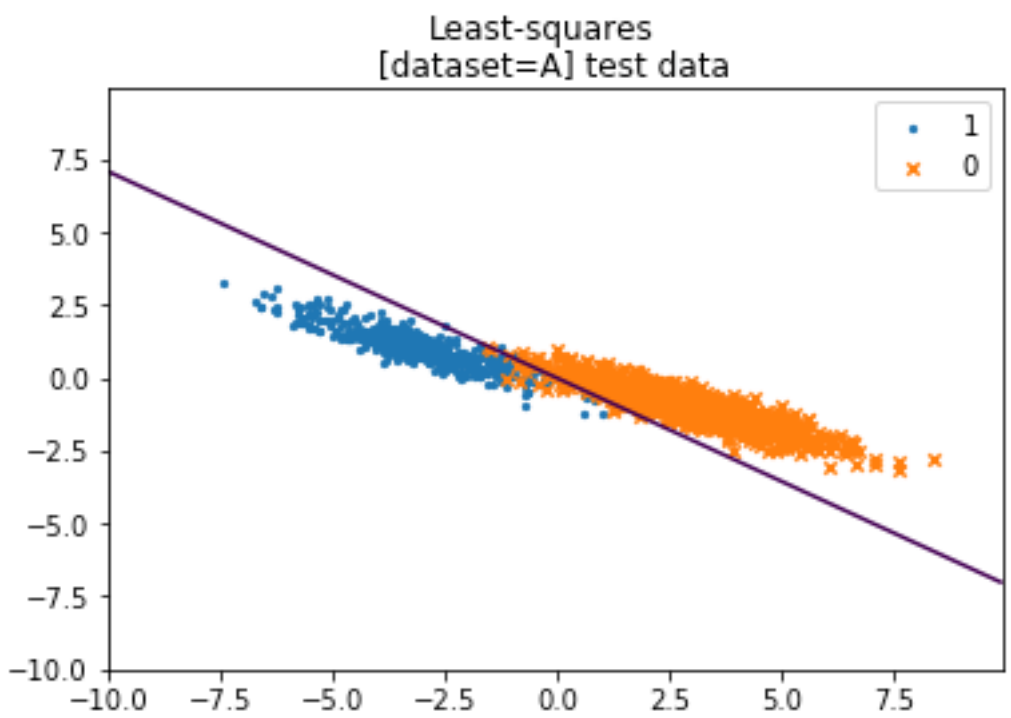
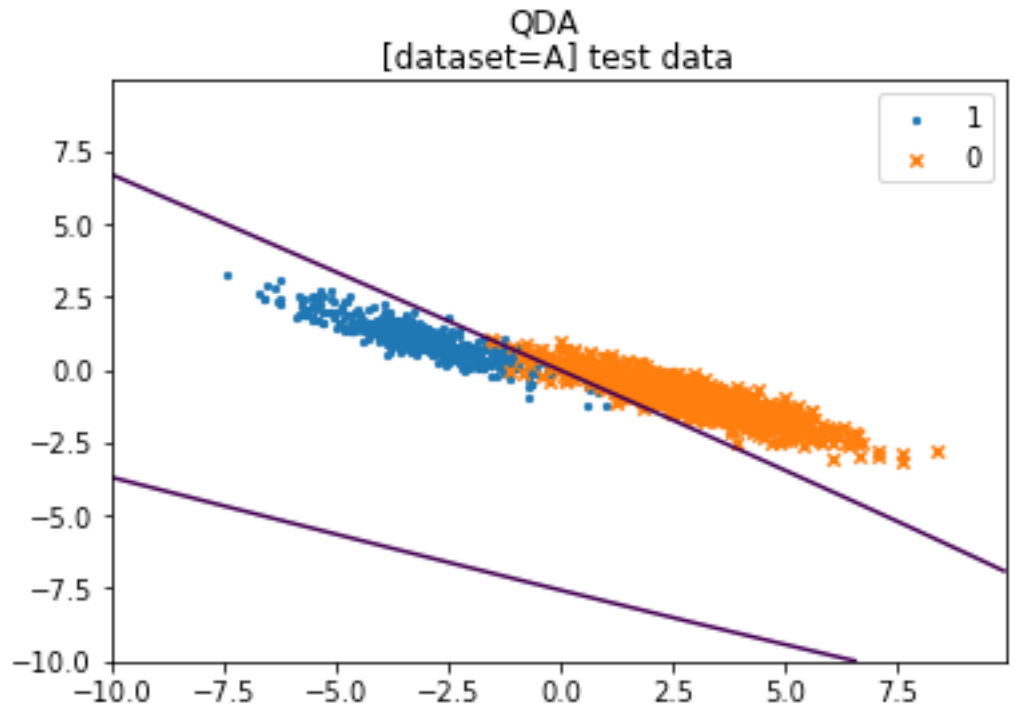
Dataset A









Misclassification error

MODEL	TRAIN	TEST
LDA	1.33%	2.00%
Logistic	0.00%	3.40%
Least-squares	1.33%	2.07%
QDA	1.33%	2.07%

Comments

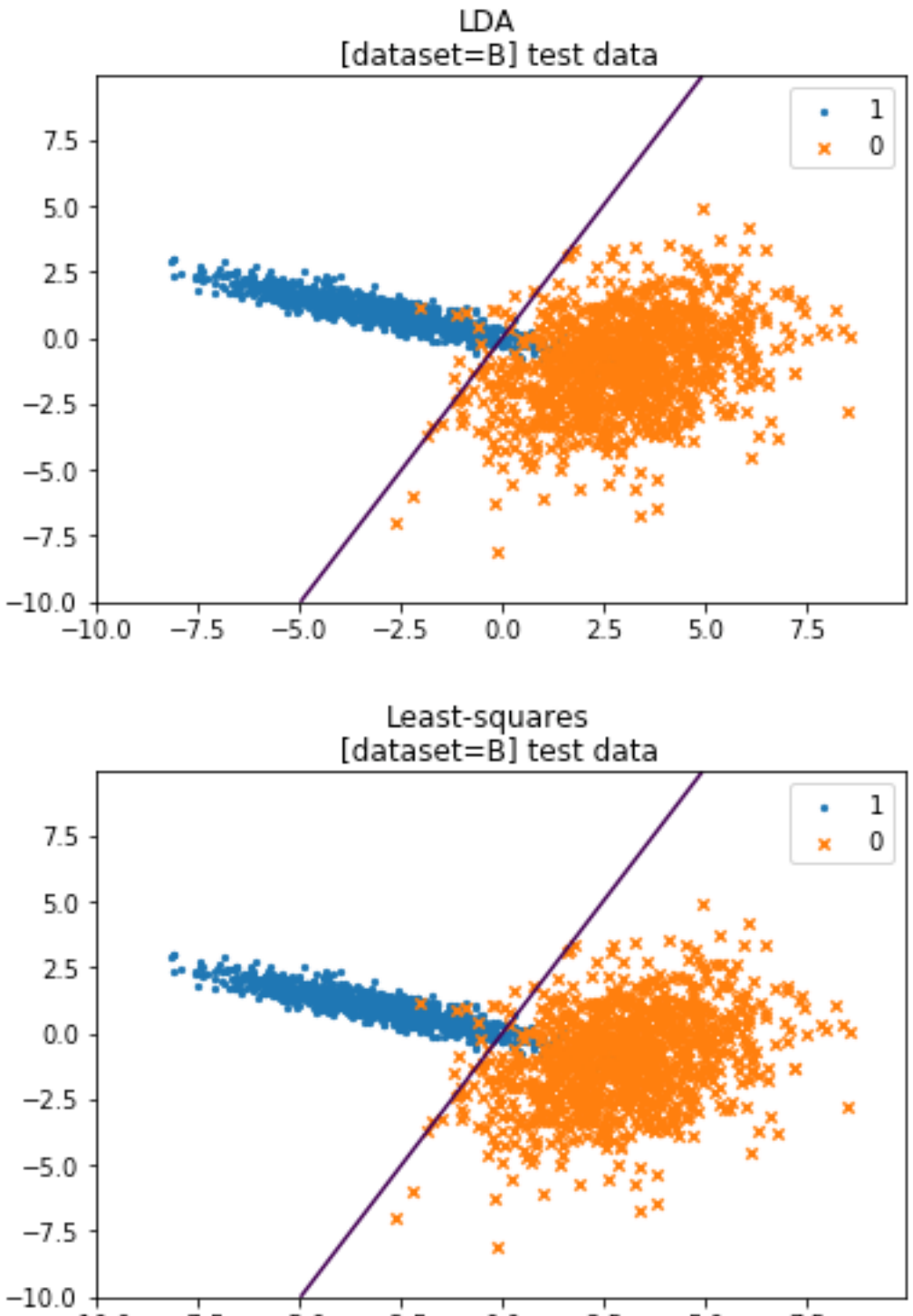
- The data is linearly separable in the training set as the **logistic regression** shows 0% misclassification error. The error rates are higher in the test set due to generalization error and the fact that data isn't separable in test set.

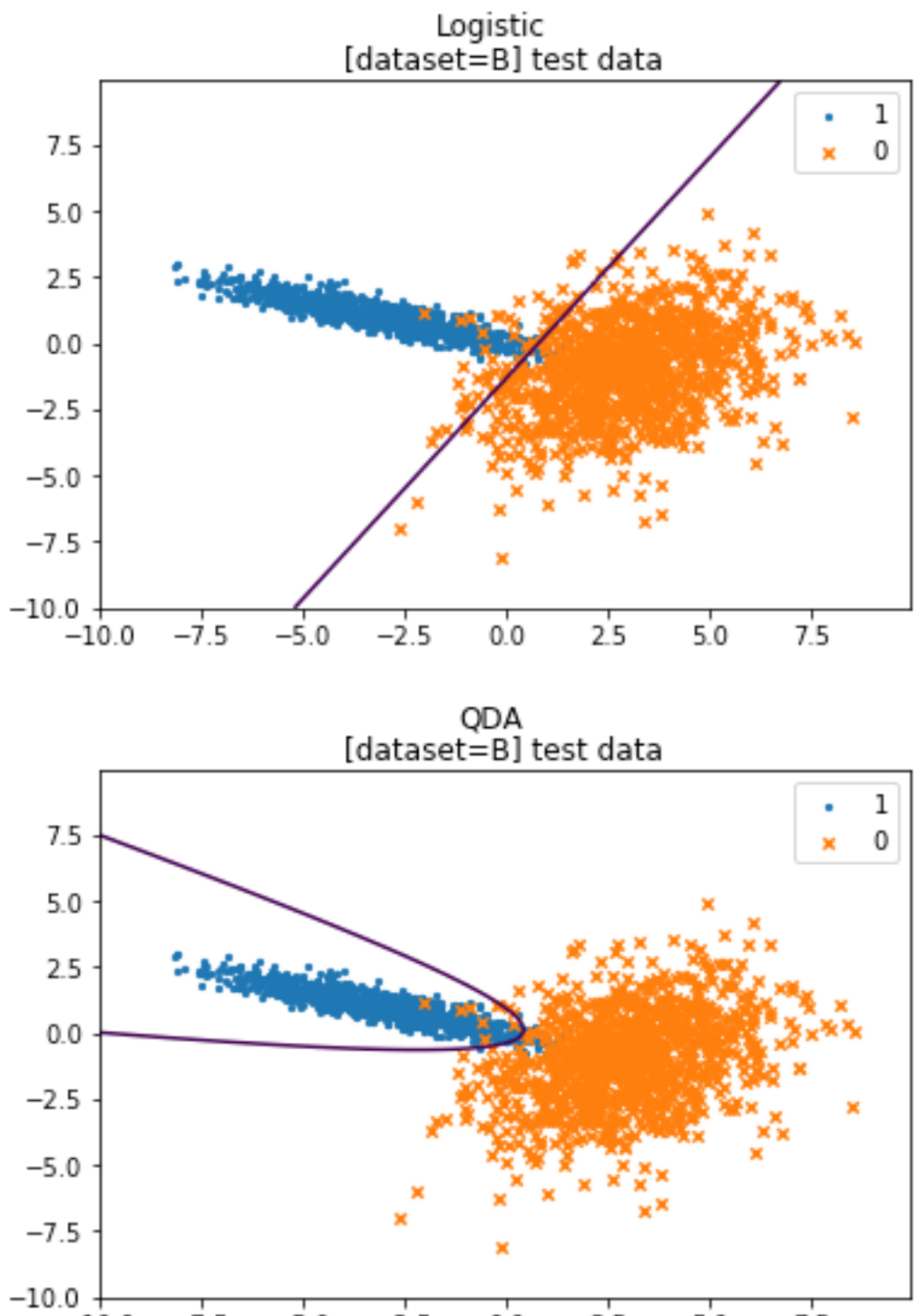
- Similarly, **LDA** and **Least-squares** model yield good results as their decision boundary is also linear. In particular, the covariance of normal distributions X|Y=0 and X|Y=1 seem very similar, which is an assumption used by LDA model.

- **QDA** does not yield better results because data are almost linearly separable.

- **Logistic regression** : we can notice that the model has slightly overfitted, the test result is the worst among the different models. This may be because of unstable estimators: in the optimization process we invert the Hessian of the loss function, which is propotionial to the empirical covariance matrix. But its columns are highly colinear (correlation coefficient of -0.92 between the features x1 and x2) thus bringing numerical instabilities. It can be noticed when decreasing the tolerance stopping criteria of the optimization: when set to 10^{-6} the loss function has an infinite value towards the end of the optimization.

Dataset B





Misclassification error

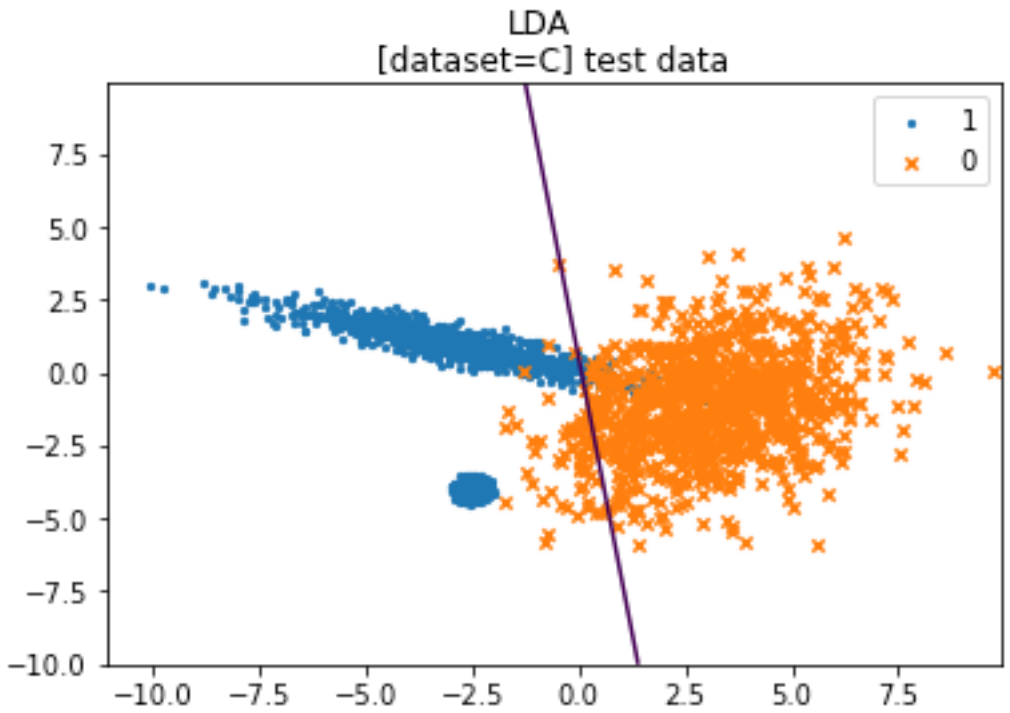
MODEL	TRAIN	TEST
LDA	3.00%	4.15%
Logistic	2.00%	4.30%
Least-squares	3.00%	4.15%
QDA	2.00%	2.80%

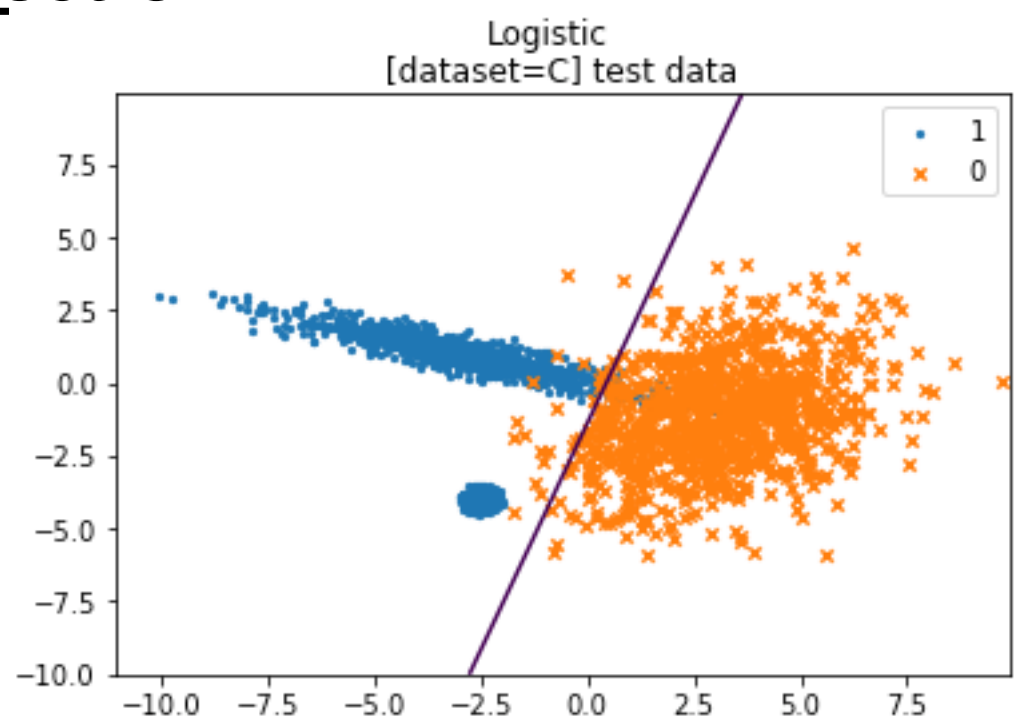
Comments

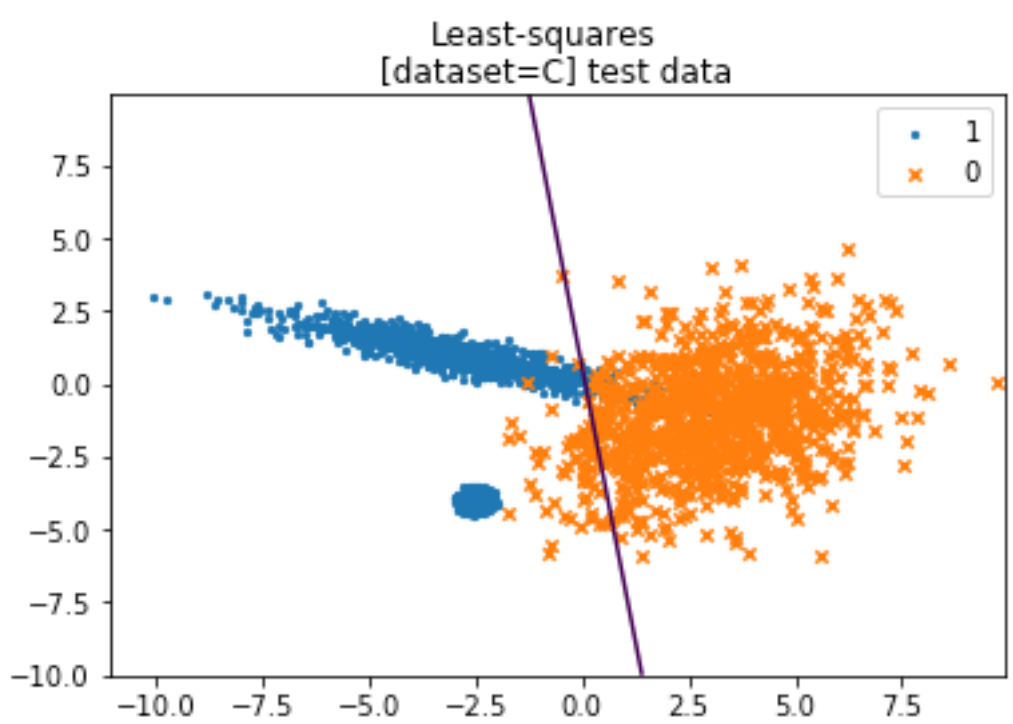
- Compared to the dataset A, X|y=0 distribution has been changed and its covariance structure is quite different from X|y=1 one. Hence **QDA**, which does not assume identical covariances between X|y=0 and X|y=1, performs better than **LDA**.

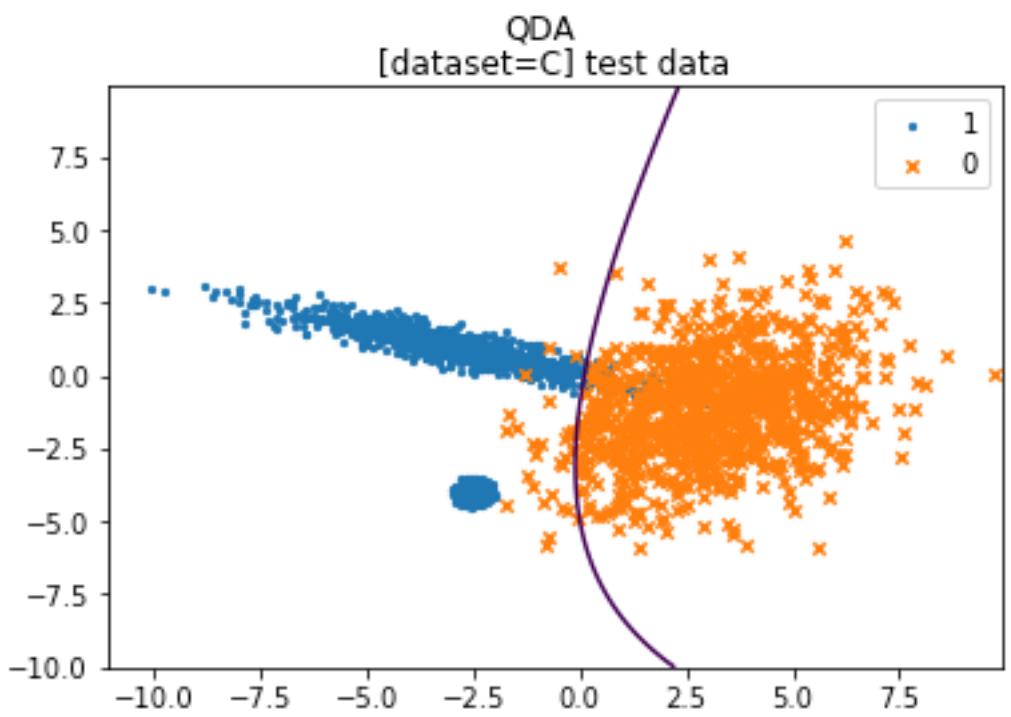
- As the data is not linearly separable, **Logistic** and **Least-squares** models perform less well. If the underlying distribution were to be known, one should compute the Bayes error rate as a benchmark for our similarly-performing results.

Dataset C









Misclassification error

MODEL	TRAIN	TEST
LDA	5.50%	4.23%
Logistic	4.00%	2.27%
Least-squares	5.50%	4.23%
QDA	4.75%	3.13%

Comments

- Compared to the dataset B, a concentration of points of the class 1 were added on the bottom left part. They do not follow the same distribution as the previous points of class 1. They can be seen as outliers with important weights and allow us to test the robustness of our models. Linear classifiers sure cannot capture this.

- **LDA** and **QDA** uses the assumption of gaussian distribution of X|Y and **Logistic regression** does not assume any distribution on X and is thus more robust to the outliers.

- **Least-squares** is sensitive to outliers as loss function used is the square of residuals. The heavy penalization makes it less robust to outlier. Logistic regression used is logistic loss which behaves linearly for highly negative values.

- This explains why compared to dataset B, after adding outliers, the slope of LDA and Least-Squares strongly shifted, and not so much for Logistic Regression.