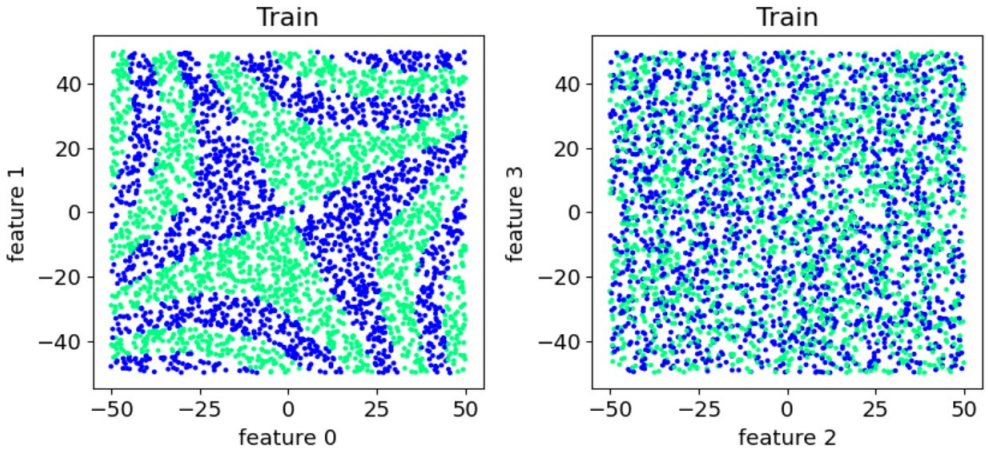
LCPB 24-25 Exercise 4, XGBoost (XGB)

Study the data in the file **x_XGB_25.dat** (N=5000 samples) with labels **y_XGB_25.dat**.

The dataset should be split into N' training samples and N'' validation samples, with N'+N''=N.



1. Model complexity, parameters' and regularization

Try different parameters (λ, γ, n_estimators, …). Which is the simplest yet effective XGB model that keeps a good validation accuracy? Is regularization useful for this analysis?

2. Dimensionality reduction

Consider reduced data samples with L'<L features. For example, feature 0,1, and 3 out of the L=4 features.

Check if the exclusion of the least important feature(s) from training data leads to better accuracy.

3. XGBoost vs NN

Compare the validation accuracy of XGB with that of a simple feed-forward neural network (FFNN)

- By varying the number of data samples N' in the training set (i.e., reducing the fraction N'/N of the data set used for training)
- With cross-validation for all cases.

Is the FFNN or the XGB performing significantly better at low N'?