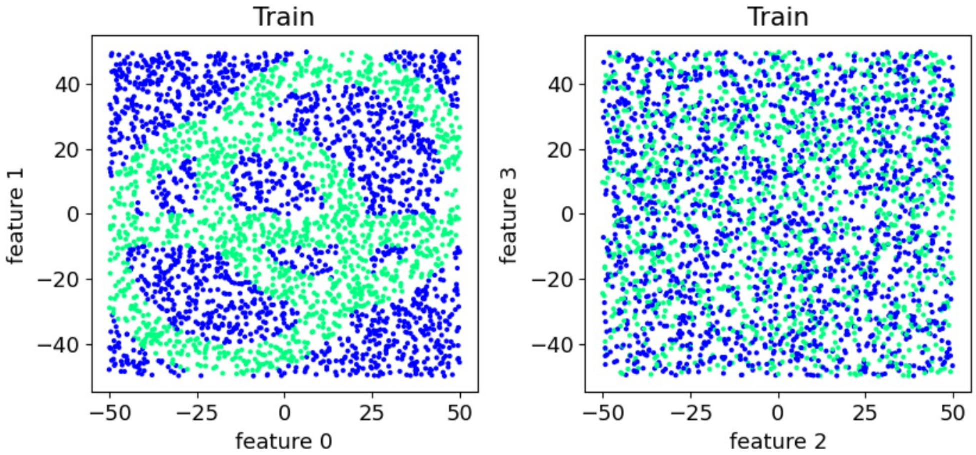
LCPB 23-24 Exercise 3, XGBoost

Study the data in the file **x_XGB_24.dat** (N=2000 samples) with labels **y_XGB_24.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 XGBoost 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 XGBoost with that of a simple feed-forward neural network (NN)

- 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 NN or the XGB performing significantly better at low N'?

Cross-validation collects the statistics from multiple realizations of training and validation, each performed for a different selection of the training set. For example, one can leave out a given block [0,1,2,…, N''-1] of data samples for validation and train on the remaining samples. The procedure is iterated for the next block [N'',…, 2N''-1], etc., so that, in total, N/N'' independent training and validations are performed with the same full dataset. Another possibility is randomly picking the validation samples. As a result of cross-validation, one gets an error estimate for the accuracy of the model.