Aston University
Birmingham

Week 5:
Practical Aspects and Ensembles

Anikó Ekárt

Module CS4730: Machine learning

Learning outcomes

In this unit we studied practical aspects, including nested cross-validation and machine learning ensembles. In this practical, we shall deepen our understanding by applying nested cross-validation for SVC with kernel functions and creating an ensemble of classifiers using bagging.

Instructions

Download the datasets `haberman.csv` [1] and `ENB2012_data.xlsx` [2]

We shall build non-linear classification models for both these datasets, re-using previous solutions created in week 4.

You are encouraged to use Python Jupyter Notebooks for your work.

Task 1

The task here is to decide which hyperparameters would be best to apply for the `haberman.csv` dataset. The hyperparameters are the kernel function, the value of C for polynomial or sigmoid, and in the case of RBF, additionally the value of γ. As the goal here is to practice nested cross-validation, feel free to choose one measure for quality, for example accuracy.

Split the data for the outer loop and prepare the outer loop (decide how many folds to use first).

For the inner loop's exhaustive search for the hyperparameters use `sklearn.model_selection.GridSearchCV`. Again, decide how many folds to use for the inner loop. Ensure you print out the best score and the best parameters for each outer fold, so that you can analyse the results.

Note. To allow re-fitting the final model selected in the inner loop on the entire training dataset for the outer loop, set the `refit` argument for `GridSearchCV` to *True*.

Discussion. *Based on the use of nested cross-validation what is the accuracy? How does the best model compare to the solution found in week 4?*

[1] Donated to UCI repository by Tjen-Sien Lim, used by Haberman, S. J. (1976). Generalized Residuals for Log-Linear Models, Proceedings of the 9th International Biometrics Conference, Boston, pp. 104-122.

[2] UCI Machine learning repository, Energy Efficiency Dataset `https://archive.ics.uci.edu/ml/datasets/Energy+efficiency`

Task 2 (optional)

If you completed task 1 and still have time, proceed with this task to create an ensemble classifier for `ENB2012_data.xlsx`. Use `sklearn.ensemble.BaggingClassifier` with a base estimator set to `SVC`, which you are already familiar with.

If you prefer, you can use the default base estimator which is decision tree, noting that we have not studied decision trees.

Discussion. *How do your results compare to those reported in the published paper? Is the ensemble performing better than a single SVC?*