A detailed Machine Learning code walk-through - best practices

Dr Geoff Spence and Dr Vince Hall

Machine learning is now widely applied in medical devices and many other applications. In this code walk-through [1], we will set out the steps required for applying machine learning to a breast cancer example [2],[3]. This requires the classification of breast cancer as malignant or benign. This work-flow is directly applicable to other machine learning problems and we set out best practices.

At the end of the code walk-through you will learn to:

- Understand 6 work-flow stages in a machine learning project
- Why formulating the goals of the project and defining expected final outcome is important
- Use descriptive statistics and visualization to better understand the data
- How to prepare the data and better expose the structure of the prediction problem to modelling algorithms.
- How to evaluate several candidate machine learning classification algorithms including dealing with class imbalance
- Learn how to apply pipelines for avoiding data leakage including
    - Preparing the Features
    - Automatic feature selection
- How to improve performance and why hyperparameter selection is important
- Understand why a performance metric is important
- How to prepare the final model, make predictions on the validation data and present the results
- Apply PANDAs for reading and analysing data
- Apply Scikit learn for machine learning in python
- Apply JUPYTER notebook

[1] The machine learning WorkFlow is based on machine learning mastery (https://machinelearningmastery.com/) by Jason BrownLee.

[2] The dataset used is publicly available and was created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA. http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29

[3] Excellent Tutorial here https://towardsdatascience.com/building-a-simple-machine-learning-model# -on-breast-cancer-data-eca4b3b99fa3