

DSI Machine Learning Workshop

Introduction to Practical Cross Validation

Mandla Gwetu (gwetum@ukzn.ac.za)

30/01/2023

Objectives

This workshop will aim to achieve the following objectives:

- Introduce participants to the Google Colab environment.
- Introduce participants to Python and notebooks.
- Explain the various ways of running Python.
- Recap on supervised machine learning definition, metrics and the KNN classifier.
- Define cross validation, list its variants, explain its need and describe its shortcomings.
- Introduce participants to the scikit-learn library.
- Demonstrate the KNN classifier using scikit-learn and the KNN classifier on colab.
- Demonstrate cross validation via test/training splits and k-folds using scikit-learn.
- Explain the class imbalance problem in supervised machine learning.
- Explain stratified cross validation.
- Demonstrate the use of stratified cross validation using scikit-learn on colab.
- Explore the full cycle of using cross validation and the KNN classifier.

Requirements

To fully engage in this workshop, delegates will require the following resources/attributes:

- A modern computer with reliable access to the internet via a browser.
- A Google account.
- A basic understanding of machine learning.
- A desire to learn new technologies.
- We will make extensive use of the following sources:
 - <https://www.kdnuggets.com/2022/07/knearest-neighbors-scikitlearn.html>
 - <https://machinelearningmastery.com/cross-validation-for-imbalanced-classification/>

Outline

- Session 1 [9:00-10:30 am] - Introduction to Python, Google Colab, Cross Validation; Types of Cross Validation.
- Session 2 [10:45-12:00 pm] – Revision of machine learning concepts and practical Dataset partitioning.
- Session 3 [1:00-2:15 pm] - Dealing with imbalanced datasets.
- Session 4 [2:30-3:45 pm] - Practical evaluation of model performance through cross validation

Task1 - Introduction

To get started with this workshop go to the following address using your browser:

<https://colab.research.google.com/>

Click **New notebook** at the bottom.

To test your new notebook, perform the following tasks.

- Provide an appropriate title for the notebook – the default title is untitled; double click it to edit it.
- Create a cell with text that describes the objectives of this practical.
- Create a cell with code to display hello word.

Task2 – Simple Cross Validation Demo

- Upload the given notebook: DSI-Practical-ML-Python-CV.ipynb onto Google Colab.
- Run each cell and observe the output.
- Write descriptive text block above each code block to explain the given example.

Task3 – Stratified Cross Validation and KNN Assignment

Create a notebook to explain and demonstrate stratified cross validation by combining the concepts covered in the above mention two sources (kdnuggets and machinelearningmastery).