

LAB 1
INTRODUCTION TO PYTHON AND SCIKIT-LEARN
ADVANCED MACHINE LEARNING
DATA 442/642

Note: This lab has been generated using material from Chapter 3 in "Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow". The following link (<https://github.com/ageron/handson-ml>) contains the extended Jupyter notebook as well as more tasks so you can better familiarize yourself with Python and Scikit-learn.

1. MNIST

MNIST dataset, is a set of 70,000 small images of digits handwritten by high school students and employees of the US Census Bureau. Each image is labeled with the digit it represents. For the first task download and display some of the digits in the dataset.

2. Binary Classifier

2.1 Identify one digit for example, the number 5. This "5-detector" will be an example of a binary classifier, capable of distinguishing between just two classes, 5 and not-5. For this task pick the Stochastic gradient descent classifier from the Scikit-Learn's `SGDClassifier` class.

2.2 Evaluate the performance of your classifier by

- (a) Measuring accuracy using cross-validation.
- (b) The use of the confusion matrix.
- (c) Understanding the precision/recall trade-off.
- (d) The use of the ROC curve.

2.3 Compare the ROC curve generated by the `RandomForestClassifier` with the ROC curve generated by the `SGDClassifier`