Introduction to Machine Learning methods
January 19-20, 2023

Day 2: Implementing Machine Learning methods in Python
Carlotta Montorsi and Thiago Quaresma Brant Chaves

This second session of the course will cover the basics of Machine learning model building, training, and testing with Python. We will cover three main steps common in any ML pipeline: data preparation, model fitting, and model out-of-sample generalization. Lastly, we will understand how to interpret black box predictions with SHAP methods.

Prerequisites

This Machine Learning Course does not presume or require any prior knowledge in machine learning. (Day 1 of the course will cover the basic theoretical principles for methods presented here.) However, to understand the concepts presented and complete the exercises, I recommend that students meet the following prerequisites:

i) You must be comfortable with variables, linear equations, functions, histograms, and statistical means.

ii) Ideally, you should have some experience in programming in Python because the programming exercises are in Python. However, programmers without Python experience can usually complete the programming exercises anyway.

If you are not familiar with Numpy or Pandas libraries do this (very basics) exercises or look at suggested you tube tutorials

i) Numpy
ii) Pandas (1)
iii) Pandas (2)
iv) Numpy tutorials: https://www.youtube.com/watch?v=QUT1VHiLmmI
v) Pandas tutorials: https://youtu.be/_Eb0utIRdkw

During the course, we will be using Google Colab to execute Python scripts. Please ensure that you can logon to https://colab.research.google.com/ before the course starts. For accessing the Colab environment, you will need a google account.

Since we are going to use Colab environment, attendees do not need to have Python installed on their local machines.

Google Colab comes pre-installed with many useful libraries:

vi) Numpy
vii) Scikit-Learn
viii) Tensorflow
ix) XGboost

x) Matplotlib

You can install other libraries by running the following code in a Colab cell:

!pip install <library>

These additional libraries require manual re-installation for each new Colab session.

To interact with me and other colleagues during the course I set up an etherpad notebook. Just click on the link and we will start interacting!

We are going to work on three main dataset:

-California Housing dataset. This dataset comes from the google storage databases.

-Titanic Dataset. You may want to download train/test from Kaggle. We can then easily import the training dataset into our Colab Environment

-Fashion-Mnist Dataset. We are going to use this dataset for the image recognition section. There is no need to download the files

Outline

1) Remainder: Model framing
   a) What you want to predict? Not all problems need ML techniques.
   b) Which outcome do you use? Do you have a clear measurable outcome (e.g. spam/not spam) or you need a proxy outcome (e.g. well-being)?
   c) Which type of data are you working with?
   d) Overview of the libraries available in python for ML and what they are used for

2) Predictor set
   a) Describe the dataset
   b) Detect anomalies in your data
   c) Preparing the data
      i) Standardization and scaling
      ii) Treating categorical variables
      iii) Identifying outliers
      iv) Dealing with missing variables
      v) Treating collinear variable
      vi) Near zero variance variables
      vii) Exercises with Titanic datasets

3) Classification and regression
   a) What is the difference?

4) Loss functions and predictions evaluation metrics
   a) What is a Loss function? How to decide your Loss function?
   b) Model evaluation metrics for regression problems

c) Model evaluation metrics for classification problems

5) Generalizing out of sample
   a) Cross validation
   b) Train-Test approach
   c) K fold cross validation
   d) Stratified K-fold cross validation
   e) Models' hyper-parameters tuning: randomized search and grid search

CLASSIFICATION MODELS

6) Logistic model
   a) Fit the model
   b) Find the appropriate level of flexibility
   c) Evaluate the model

7) Regularized Regression: lasso, ridge, elastic net
   a) Exploring hyper-parameters
   b) Fit the model
   c) Evaluate model

8) Trees/Forest
   a) Exploring hyper-parameters
   b) Fit the model
   c) Evaluate model fit

9) Gradient Boosting
   a) Exploring Gradient Boosting hyper-parameters
   b) Fit the model
   c) Evaluate model
   d) Visualizing the validation curve/learning curve

10) Shapley values and SHAP library
   a) SHAP force plot/ SHAP summary plots
   b) Interpreting Gradient Boosting model
   c) Exercises with other models

11) Super Learner
   a) Build a Super Learner
   b) Evaluate the model

12) Neural Network
   a) Exploring Neural Network structures and hyper-parameters
   b) Fit the model
   c) Evaluate model fit
   d) Visualizing the validation curve/learning curve

13) Image recognition with Convolutional Neural Network
    a) Data Exploration
    b) Model Building
    c) Model Evaluation
    d) Prediction on new data

14) Unsupervised learning – Clusters and dimensionality reduction
    a) PCA – Linear Algebra
    b) TSNE – Nonparametric
    c) UMAP – Manifold approximation