Analytics with Apache Spark and MLlib tutorial notes

A. Antonio R. Marin[1, a)] and B. Joeri R. Hermans[1, b)]
CERN, IT-DB-SAS

(Dated: 31 August, 2016)

Machine learning has become a hot-topic in the recent years. Together with the rise of distributed computing and GPU's, models can be computed and evaluated faster. In this tutorial the participant will learn how distributed computing frameworks like Apache Spark benefit the analysis of a big data set. We will guide the participant through the complete analysis pipeline using Spark's MLlib (Spark's built-in Machine Learning library); starting with data preparation and feature selection, and ending with model evaluation techniques such as cross-validation.

Keywords: Apache Spark, MLlib, Analytics, Machine Learning, Distributed Computing

I. INTRODUCTION

[a)] Electronic mail: antonio.romero.marin@cern.ch

[b)] Electronic mail: joeri.hermans@cern.ch