FACULDADE
ENGENHARIA

Departamento de
Informática

Black Box to Mining of Massive Datasets

Proposta de Projeto

Orientador: Sebastião Pais(sebastiao@di.ubi.pt)

1 Context

Machine learning and data mining play crucial roles in handling massive datasets. With the ever-increasing volume of data generated, traditional methods often struggle to extract valuable insights and patterns effectively. However, leveraging machine learning algorithms and data mining techniques makes it possible to navigate and uncover meaningful information from these vast datasets.

Machine learning algorithms can automatically learn from data and make predictions or take actions without explicit programming. They can analyze patterns, identify trends, and make data-driven decisions. In the context of massive datasets, machine learning algorithms can handle large volumes of information, effectively scaling to process and extract knowledge from diverse data sources.

Data mining, on the other hand, focuses on discovering hidden patterns, relationships, and structures within data. It involves various techniques such as clustering, classification, association rule mining, and anomaly detection. These methods can sift through massive datasets, identify valuable insights, and extract practical knowledge to drive informed decision-making and facilitate future predictions.

When applied together, machine learning and data mining enable researchers and practitioners to tackle the challenges posed by massive datasets. They provide the means to analyze, process, and extract meaningful information from vast amounts of data, leading to valuable discoveries, improved decision-making, and enhanced understanding of complex phenomena. With advancements in computational power and algorithmic techniques, machine learning and data mining continue to revolutionize our ability to leverage massive datasets for insights and innovation.

2 Keywords

Big Data, Machine Learning

Goals

The main objective of this project is to propose the development of an abstract "black box"package in Python that integrates Spark machine learning capabilities. This package aims to provide a simplified and user-friendly interface for utilizing Spark's powerful machine-learning algorithms and tools.

By encapsulating the complexities of Spark's machine learning framework within a user-friendly package, this project aims to enable researchers and practitioners to leverage Spark's distributed computing capabilities without delving into the underlying technology's intricacies. The proposed package will offer a high-level abstraction that hides the complexities of distributed data processing and allows users to focus on their specific machine-learning tasks.

This abstract "black box"package will provide a set of intuitive and easy-to-use APIs for tasks such as data preprocessing, feature engineering, model training, and evaluation. Users can seamlessly leverage Spark's distributed processing capabilities, allowing them to scale their machine-learning workflows to handle large-scale datasets and benefit from parallelized computations.

The development of this package will be based on Python, a popular programming language among data scientists and machine learning practitioners. Python's rich ecosystem of libraries and its ease of use make it an ideal choice for creating a user-friendly and accessible package for Spark machine learning.

Overall, this project aims to bridge the gap between the power of Spark's machine learning capabilities and the ease of use for researchers and practitioners by proposing the development of an abstract "black box"package in Python. This package will empower users to leverage Spark's distributed computing capabilities for efficient and scalable machine learning workflows by providing a simplified interface.

Workplan

T1 Review the State-of-the-art;

T2 R&D in context of the "Black Box";

T3 Implementation;

T4 Testing and evaluation;

T5 The writing of the report.

Academic Prerequisites

Interest about Artificial Intelligence, Big Data, Machine Learning.

Assessment elements to deliver.

Source code and documentation of all code development; Report.

Contacts

Sebastião Pais (sebastiao@di.ubi.pt)