Министерство науки и высшего образования Российской Федерации

Московский физико-технический институт

Промышленное машинное обучение

Осень

2022

Лабораторная работа №4

КОМПОЗИЦИЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ НА SPARK

Цель работы:

Получить навыки организации взаимодействия различных моделей машинного обучения и реализации единого жизненного цикла эксплуатации (pipeline).

Задачи работы:

1. Использовать модель кластеризации, реализованную в лабораторной работе №3, для генерации размеченного набора данных, полученного также в лаб. раб. №3, по кластерам (классам).
2. Провести анализ сгенерированных данных, разделить на обучающую и тестовую выборки.
3. Реализовать модель классификации сгенерированных данных с любым изученным алгоритмом машинного обучения.
4. Реализовать модель **линейной** регрессии для каждого вектора признаков, соответствующему метке класса, полученной на выходе модели классификации.
5. Вывести метрики трёх моделей, то есть построить матрицы ошибок, графики регрессии и т.д.
6. Использовать API Apache Spark/PySpark.
7. Рекомендуется использовать Docker контейнеры.

Основное задание:

Каждая модель композиции запускается по требованию или по расписанию. Модели могут быть либо в разных Spark приложениях, либо в одном.

Дополнительное задание:

Организовать единую точку входа в pipeline и его эксплуатацию с помощью таких сторонних средств, как ML Flow, Apache Airflow. Допускается формирование Spark pipeline.