

HW3: Spark ML + Breeze



Описание работы и критериев оценивания

ДЗ предлагается оформить в виде Scala проекта и выложить в свой репозиторий

Бонусы и штрафы:

- **100%** за плагиат
- **30%** за посылку решения в течение недели после deadline

Блок 1. Линейная регрессия на Spark

Используя Spark ML и Breeze разработать модуль распределенного обучения алгоритма линейной регрессии. Проверку корректности работы предлагается сделать с помощью синтетического датасета с известными параметрами с добавлением шума (по вкусу).

Основные требования (40 баллов):

- Размер тренировочного датасета не меньше 100000 сэмплов
- Параметры скрытой модели для генерации тренировочного датасета (1.5, 0.3, -0.7)
- Приложить в репозиторий инструкцию по запуску и логи обучения с результатами

Покрытие тестами проверяющими работоспособность компонентов решения (20 баллов)

Реализовать обучение минибатчами с матричной арифметикой (20 баллов)

Проверить обучить модель на датасете <https://www.kaggle.com/datasets/mirichoi0218/insurance>

Привести метрики, сделать выводы о качестве (20 баллов)