Распределенный градиентный спуск. Java, Spark

Володин Вадим, 2018

Свойства алгоритма

Время работы алгоритма на одном потоке - 50,798 сек.
Время работы алгоритма на четырех потоках - 30,321 сек.

Свойства алгоритма

rmse уменьшается с каждой итерацией

rmse = sum[(y - h(x)) ^ 2]

Свойства алгоритма

Коэффициент a при производной изначально равен единице. Он уменьшается таким образом, чтобы rmse на новой итерации было меньше, чем на предыдущей.

Использованные технологии

Java

Apache Spark

Junit

Дополнительные инструменты

Средство сборки проекта

- Maven

Средство контроля версий

- Git

Тестовые данные

testdata/

learn.csv - данные с одного домашнего задания по линейной регрессии

learn2.csv - данные, сгенерированные вручную, для проверки правдоподобия ответа

learn3.csv - данные, сгенерированные вручную, для проверки правдоподобия ответа

Структура проекта.

Sample - Один сэмпл. X_i (row), y_i (val)

CalcRmse - Лямбда-выражение, позволяющее аккумулировать rmse.

CalcRmse - Лямбда-выражение, позволяющее аккумулировать градиент.

VectorMethods - статическиий метод dotProduct, считает скалярное произведение двух векторов

StringToSampleRdd - позволяет перевести JavaRDD<String> в формат JavaRDD<Sample>

GradientDescent - метод градиентного спуска (getcoeff). Возможна инициализация от максимального числа шагов, и от погрешности между итерациями

Спасибо за внимание

https://github.com/PolyProgrammist/internship_task