Володин Вадим.

Распределенный алгоритм градиентного спуска. Java, Spark

Отчет.

Производился запуск алгоритма с 10 шагами на данных testdata/learn.csv.

Время работы алгоритма на одном потоке - 50,798 сек.

Время работы алгоритма на четырех потоках - 30,321 сек.

Таким образом, присутствует горизонтальная масштабируемость.

Алгоритм проверен на небольших тестовых данных, на которых несложно визуализировать ответ. Алгоритм выдает близкий к задуманному ответ.

Каждую новую итерацию алгоритм уменьшает rmse. Так, за 100 итераций на данных testdata/learn2.csv было достигнуто rmse 30.41163324163568 при погрешности между итерациями 0.10570380898558795.

Коэффициент при производной высчитывается следующим образом. Изначально он равен единице. Затем пытаемся сделать шаг. Если rmse увеличилось, то шаг был сделан в сторону от локального минимума, и необходимо уменьшить коэффициент. Иначе делаем шаг и переходим к следующему. На testdata/learn.csv начальный коэффициент достигает 1e-6, то есть осуществляется 10 холостых итераций.

https://github.com/PolyProgrammist/internship\_task