Fitting lines to data using Spark:

Uses a technique called Stochastic Gradient Descent (SGD). Friendly to multi-dimensional data as it looks for contours in higher dimensions.

We train the model and make predictions using LabeledPoint objects. Here “label” is the value we are trying to predict – usually our Y axis – and the “feature” is X axis or other axes.

So we train it with a bunch of known x,y points.

And predict new Y values for given X’s using the line the model created.

SGD doesn’t handle “feature scaling” well. It assumes data is similar to a normal distribution.

Here we are trying to predict revenue based on page speed using a linear model (e-commerce).