**World Quant University**

**Professor: Tiberiu Stoica**

**Machine Learning**

Nikolas Lippmann Pareschi - [nikolaslippmann@gmail.com](mailto:nikolaslippmann@gmail.com)

**Unit 5 Project**

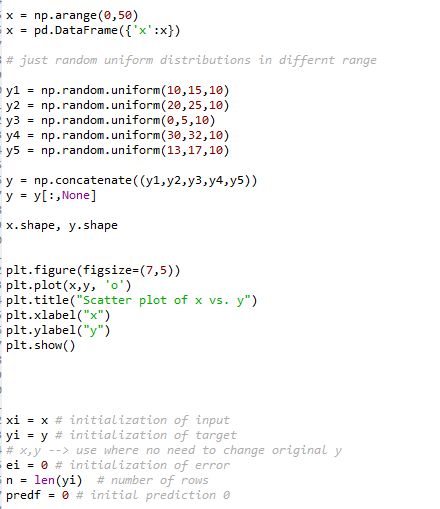
### Overview

Gradient Boosting was developed by Stanford Professor Jerome Friedman.  Gradient Boosting develops an ensemble of tree-based models by training each of the trees in the ensemble on different labels and then combining the trees. For a regression problem where the objective is to minimize MSE, each successive tree is trained on the errors left over by the collection of earlier trees.

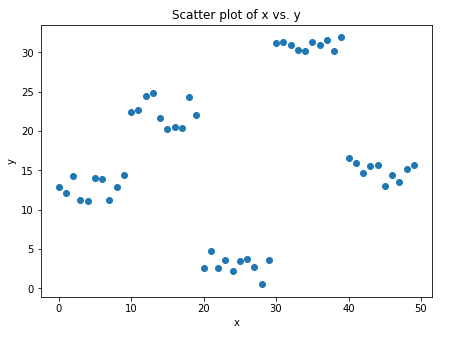
The derivation of the algorithm can be found here:  <https://en.wikipedia.org/wiki/Gradient_boosting>

The purpose of this project is to code up a Gradient boosting on a simple synthetic data set and find out  what is the optimal ensemble size which lowers the MSE of the problem.

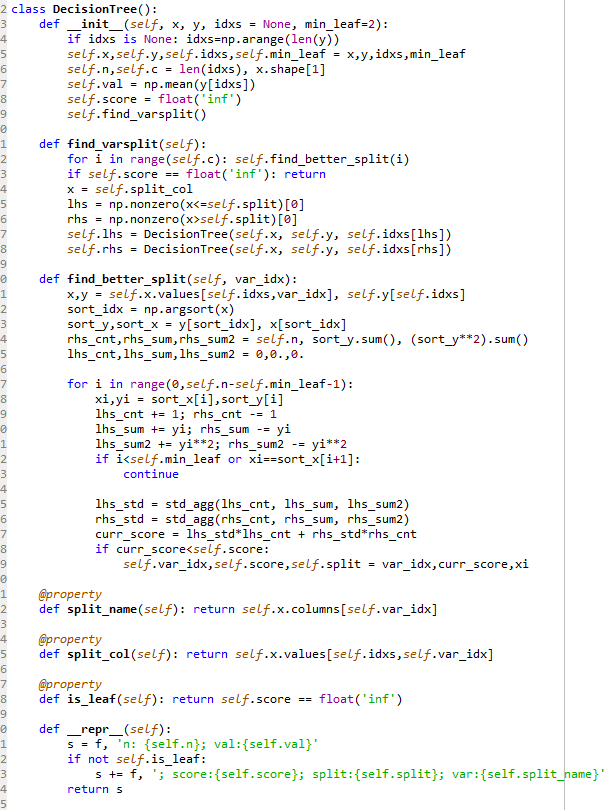
The first thing we should do is to generate random data to use the Gradient boosting technique:



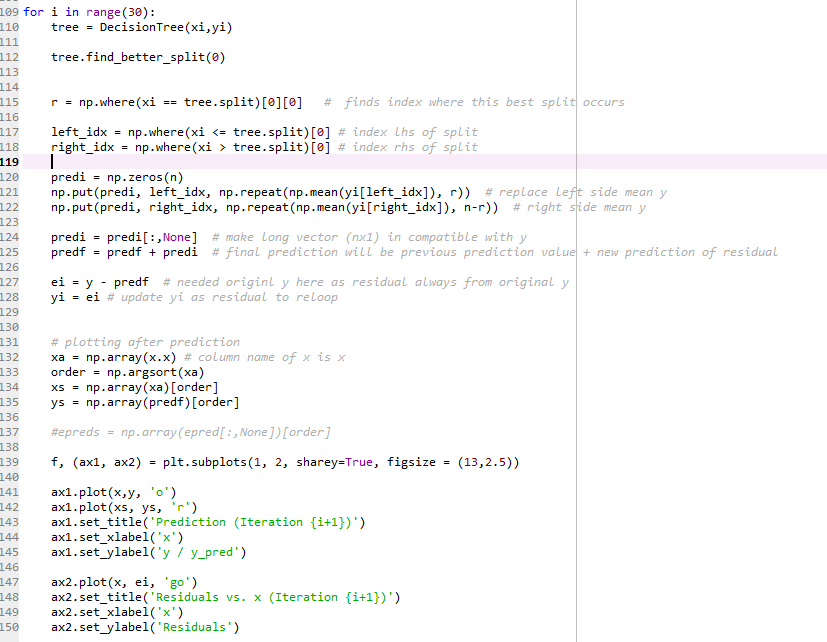
This is a graph of the random generated data:



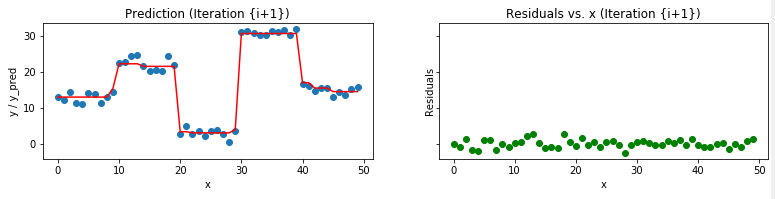
We use this class to process the tree classification of the data and to do the predictions:



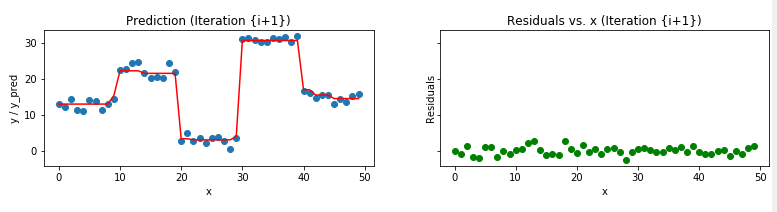
We need to do the gradient boosting technique to generate a number of trees and then minimize the objective function of the residuals. If everything is fine the residuals should have a lower magnitude after n iterations:



First iteration:



Last iteration:



We used the following tutorials to implement the gradient boost in the random data:

<http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html>

<https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d>

<https://www.analyticsvidhya.com/blog/2017/04/40-questions-test-data-scientist-machine-learning-solution-skillpower-machine-learning-datafest-2017/>

<https://nbviewer.jupyter.org/github/groverpr/Machine-Learning/blob/master/notebooks/01_Gradient_Boosting_Scratch.ipynb>

<https://gist.github.com/tonicebrian/4018084>