Lecture 20

Ensemble Methods: Boosting

Murat Unal

Johnson Graduate School of Management
Cornell University

11/09/2021

Agenda

Ensemble methods

- We can overcome the weaknesses of a single tree by combining many individual trees.
- The following methods use trees as building blocks to construct more powerful prediction models.

1. Bagging
2. Random forests
3. Boosting

Random forests

- If there's a single strong predictor in our model, then all the bagged trees will have the same predictor as an important input and all the trees will be highly correlated.
- This will prevent bagging from reducing the variance among the trees.
- Random forests provide an improvement over bagged trees by way of a small tweak that **decorrelates** the trees.
- This reduces the variance when we average the trees.

Random forests

- As in bagging, we build a number of decision trees on bootstrapped training samples.
- In order to **decorrelate** its trees, a random forest only considers a random subset of predictors when making each split (for each tree).

Random forests

- Random forests help to reduce tree correlation by injecting more randomness into the tree-growing process
- While growing a decision tree during the bagging process, random forests perform split-variable randomization where each time a split is to be performed, the search for the split variable is limited to a random subset of m_{try} of the original p features.
- Typical default values are $m_{try} = \frac{p}{3}$ for regression and $m_{try} = \sqrt{p}$ for classification, but this should be considered a tuning parameter.

Boosting

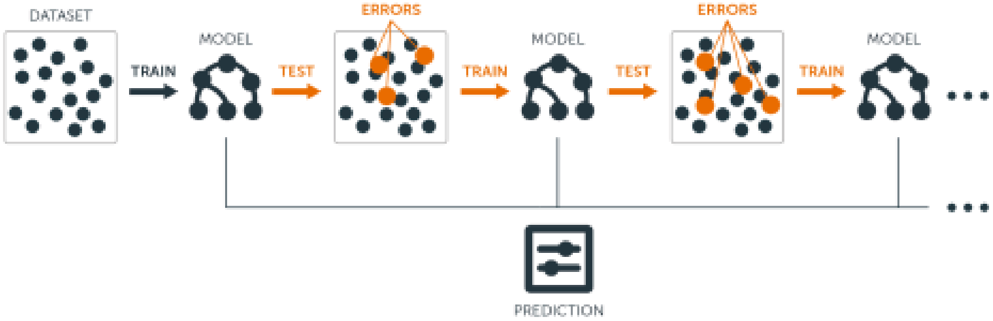
- Gradient boosting machines (GBMs) are an extremely popular machine learning algorithm that have proven successful across many domains and is one of the leading methods for winning Kaggle competitions.

Boosting

- The previous methods grow trees independent of each other.
- Boosting trains trees **sequentially** thereby allows the flow of information between two consecutive trees.
- This is accomplished by training each new tree on the residuals (mistakes) from its predecessor.

Boosting

Sequential ensemble approach

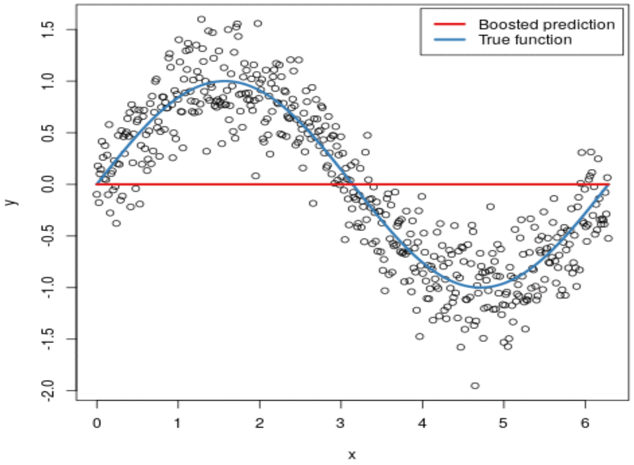


Source: HOML

Boosting

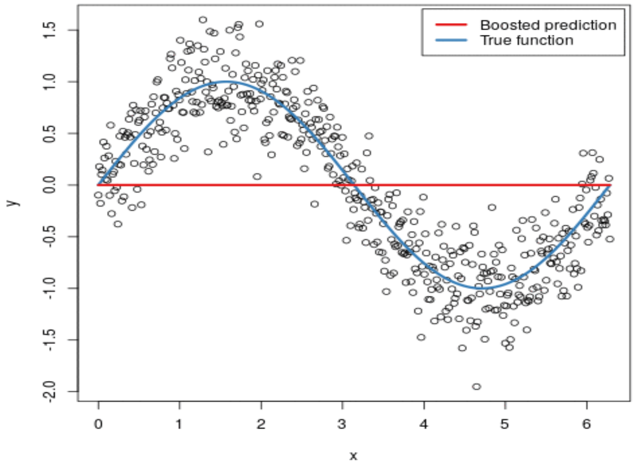
- Each of these trees can be rather small, with just a few terminal nodes.
- By fitting small trees to the residuals, we slowly improve the model in areas where it does not perform well.
- Unlike fitting a single large decision tree to the data, which leads to potential overfitting, the boosting approach instead learns slowly.

Boosting in action



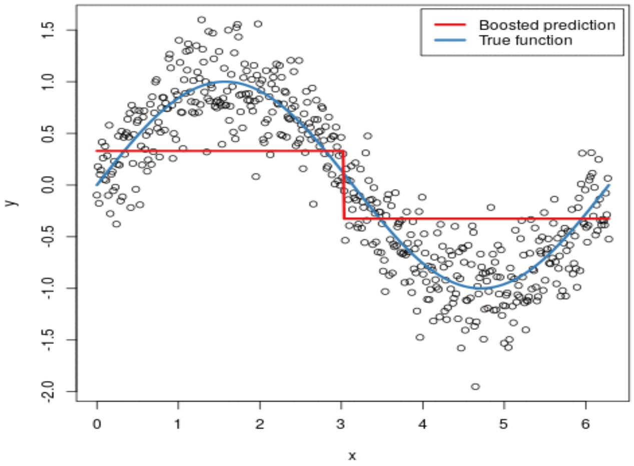
Source:HOML

Boosting in action



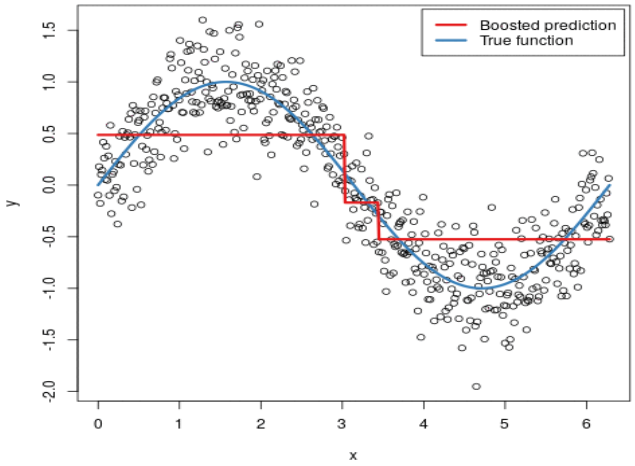
Source:HOML

Boosting in action



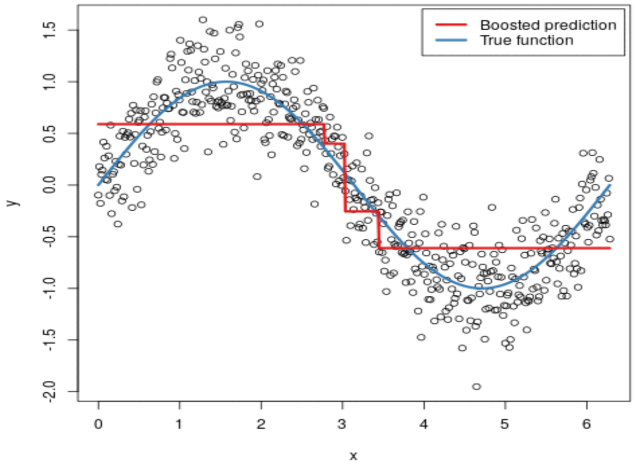
Source:HOML

Boosting in action



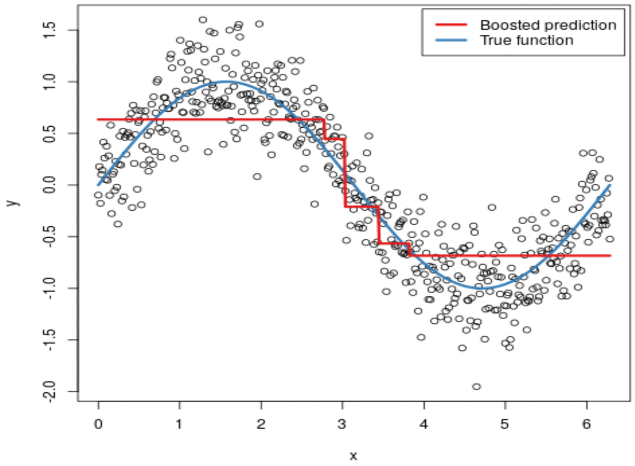
Source:HOML

Boosting in action



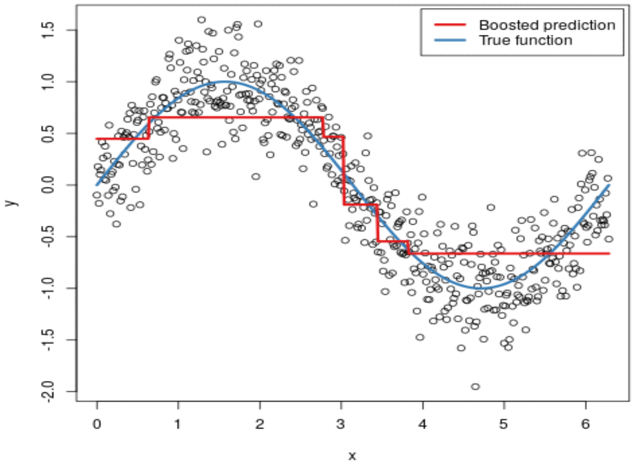
Source:HOML

Boosting in action



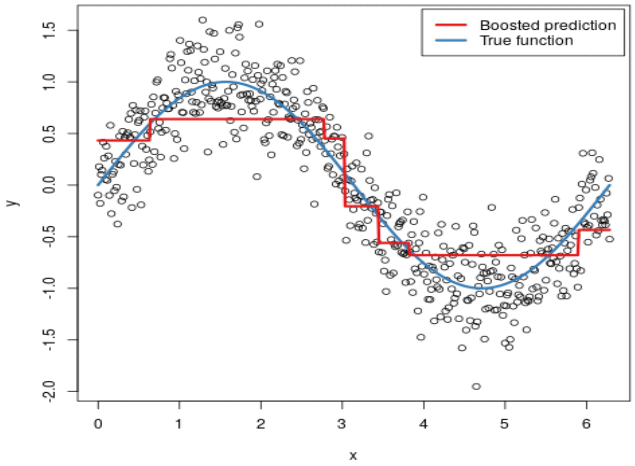
Source:HOML

Boosting in action



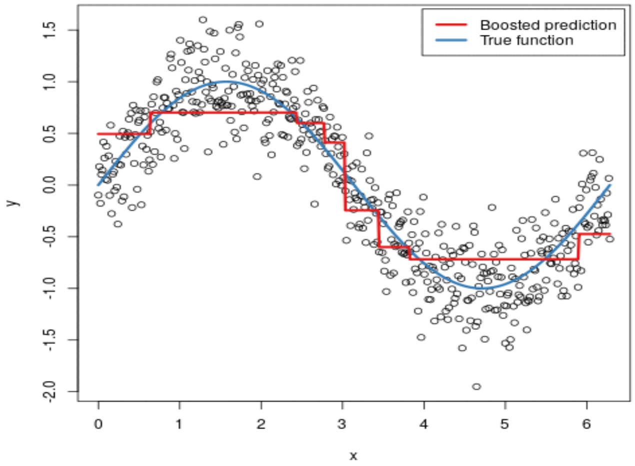
Source:HOML

Boosting in action



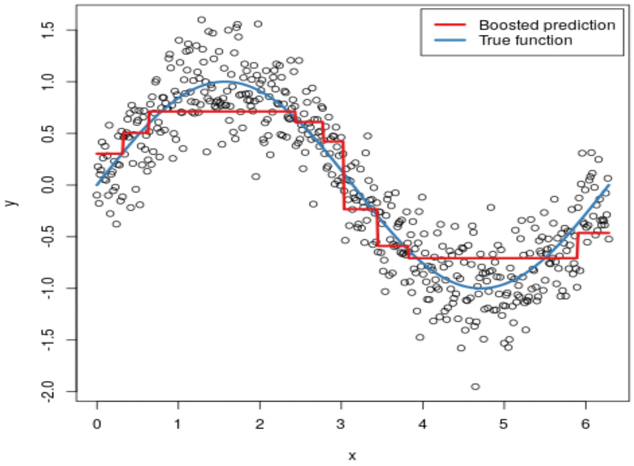
Source:HOML

Boosting in action



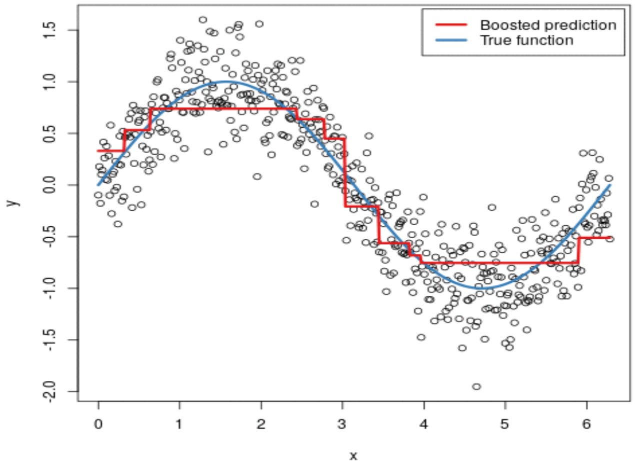
Source:HOML

Boosting in action



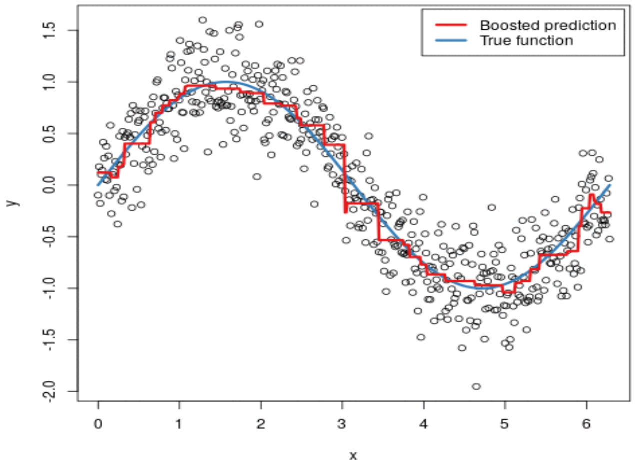
Source:HOML

Boosting in action



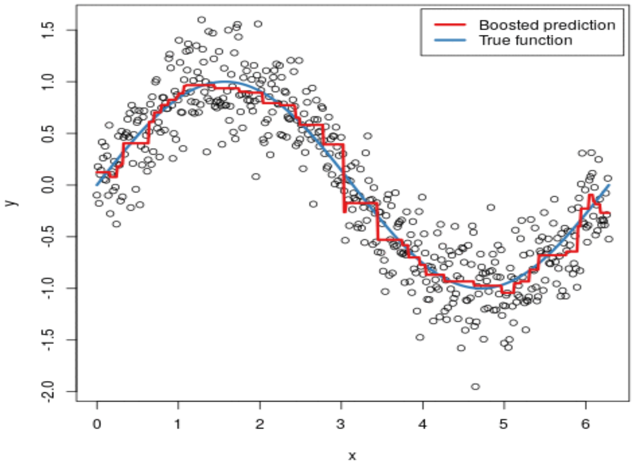
Source:HOML

Boosting in action



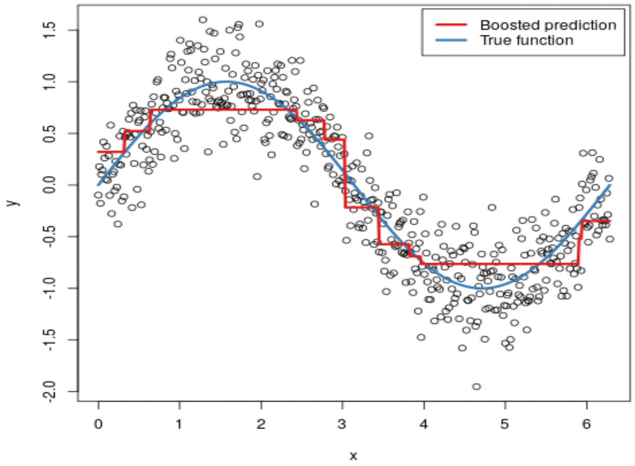
Source:HOML

Boosting in action



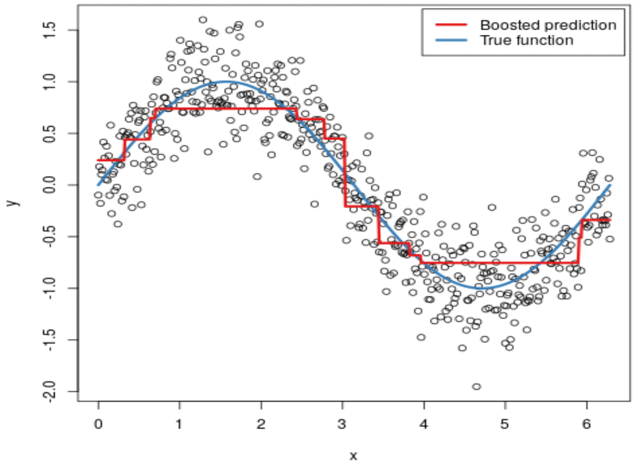
Source:HOML

Boosting in action



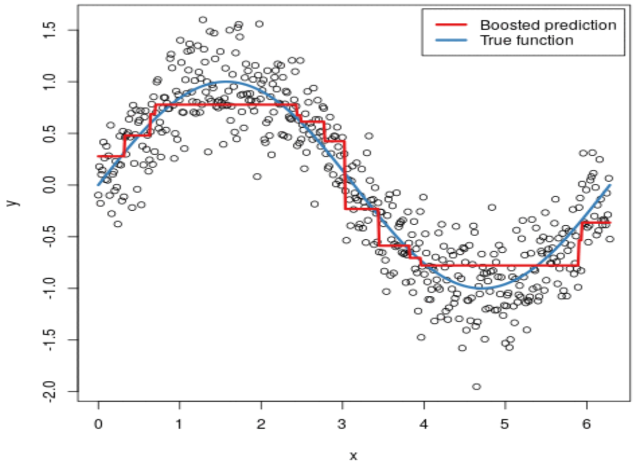
Source:HOML

Boosting in action



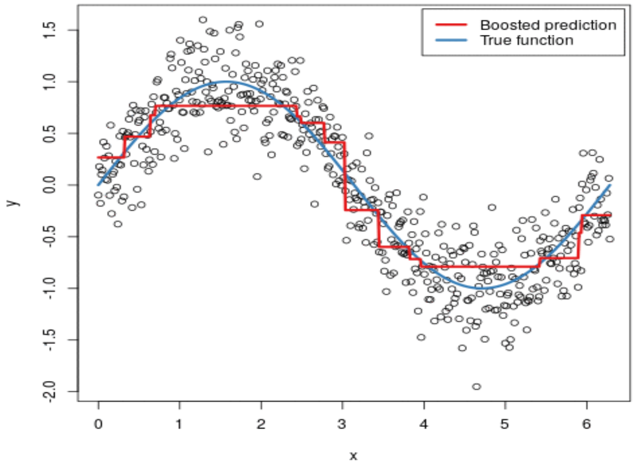
Source:HOML

Boosting in action



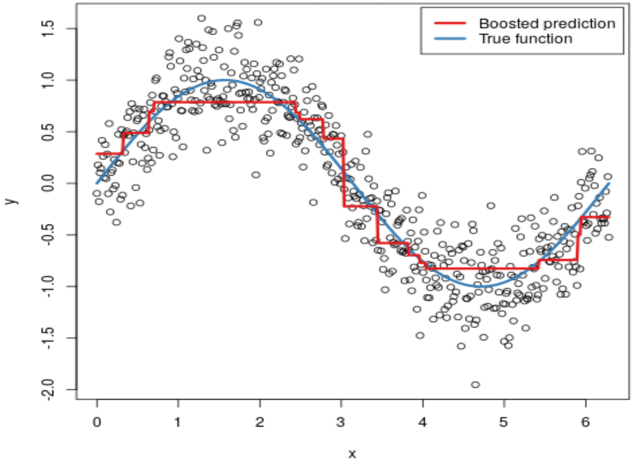
Source:HOML

Boosting in action



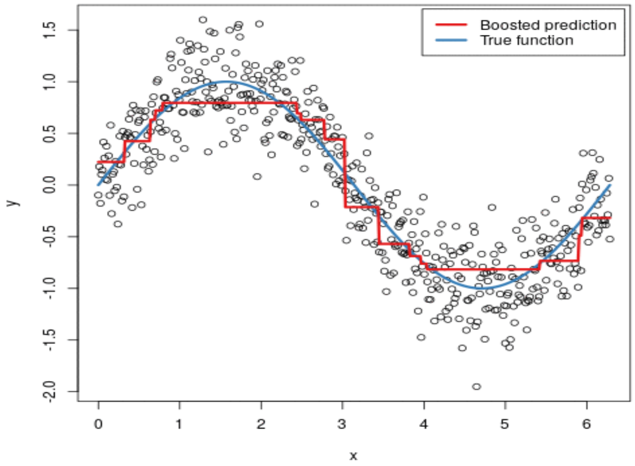
Source:HOML

Boosting in action



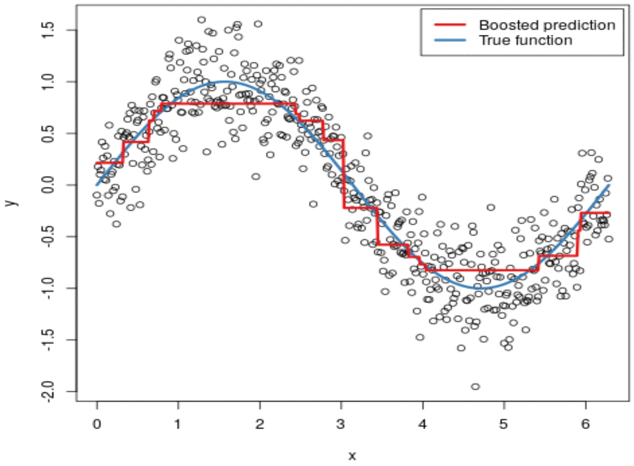
Source:HOML

Boosting in action



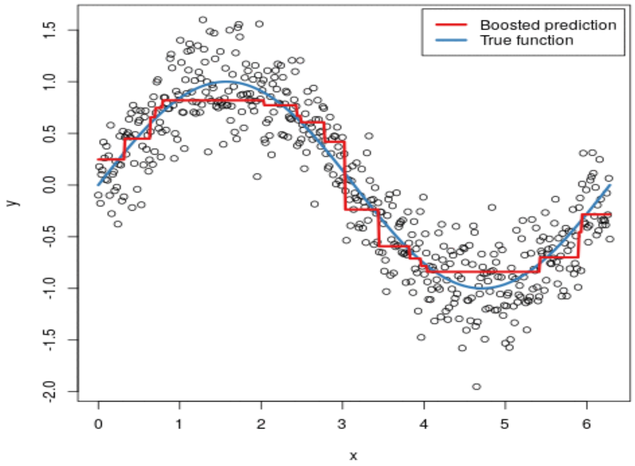
Source:HOML

Boosting in action



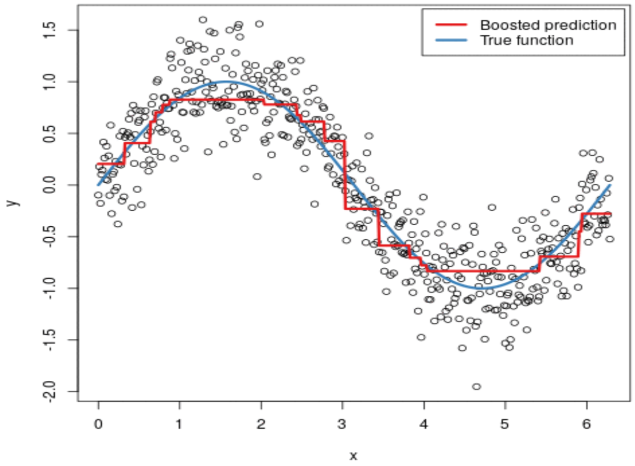
Source:HOML

Boosting in action



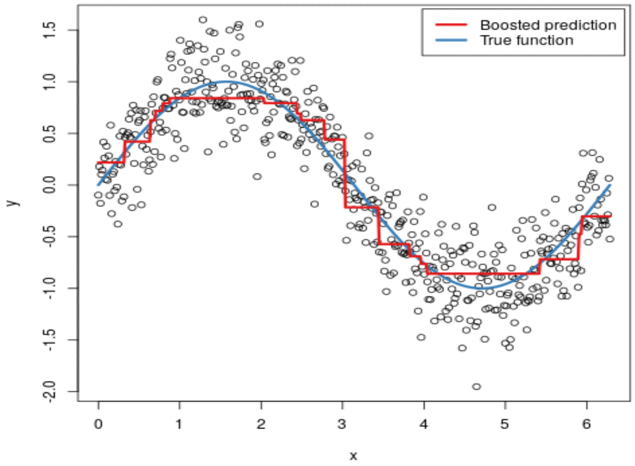
Source:HOML

Boosting in action



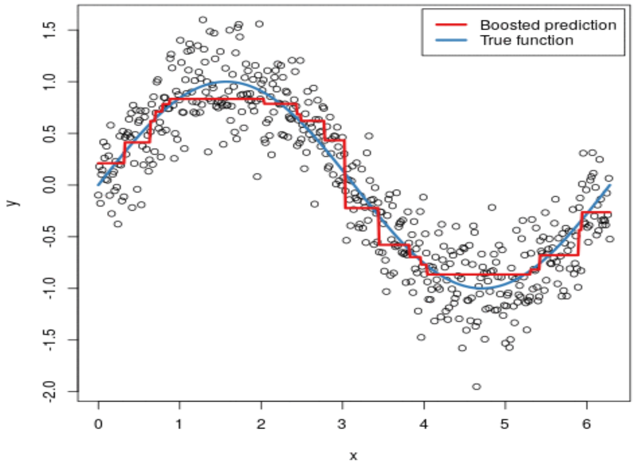
Source:HOML

Boosting in action



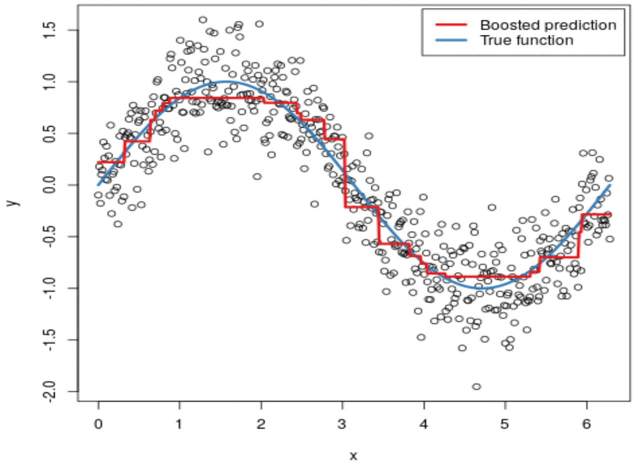
Source:HOML

Boosting in action



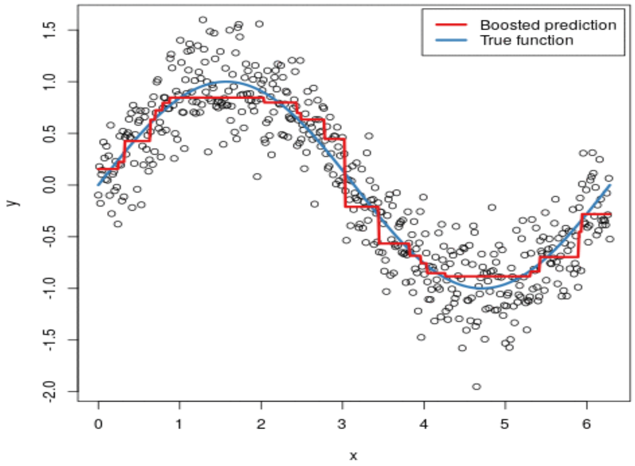
Source:HOML

Boosting in action



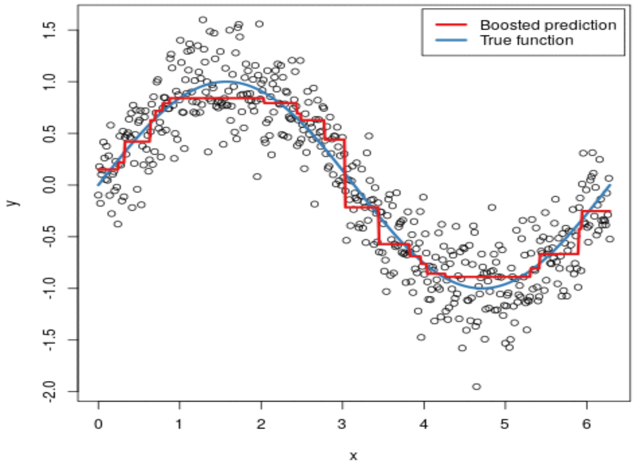
Source:HOML

Boosting in action



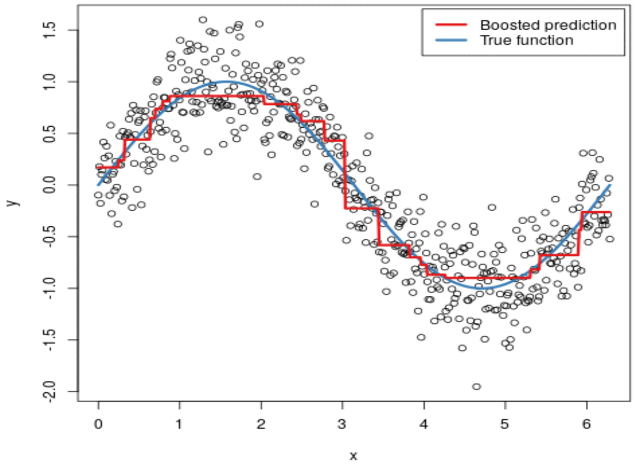
Source:HOML

Boosting in action



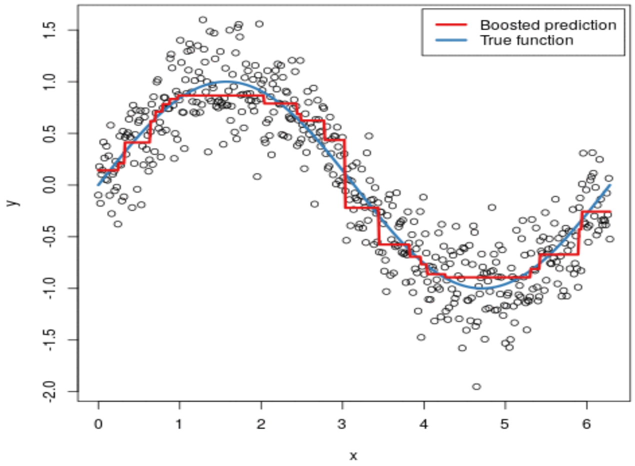
Source:HOML

Boosting in action



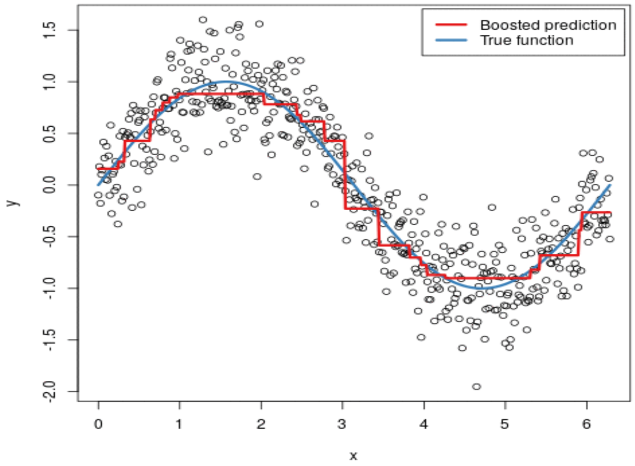
Source:HOML

Boosting in action



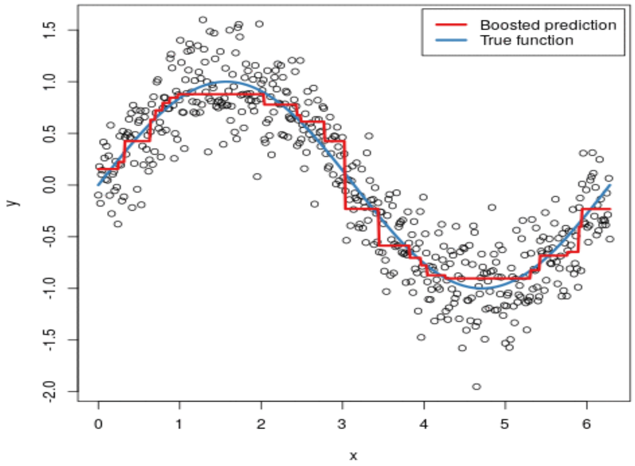
Source:HOML

Boosting in action



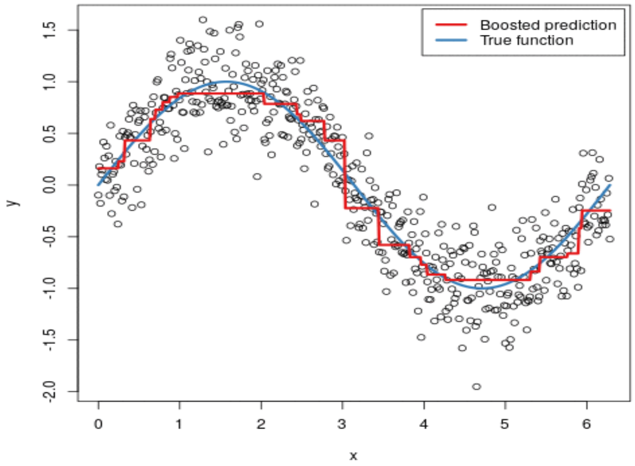
Source:HOML

Boosting in action



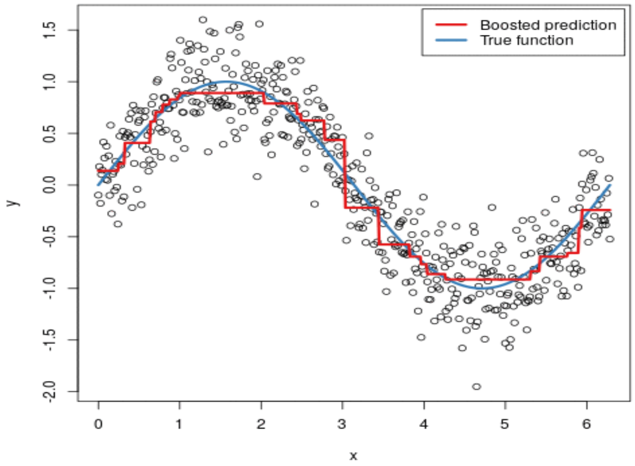
Source:HOML

Boosting in action



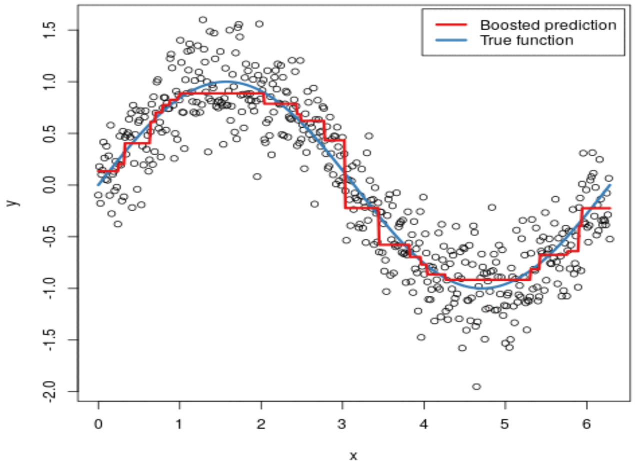
Source:HOML

Boosting in action



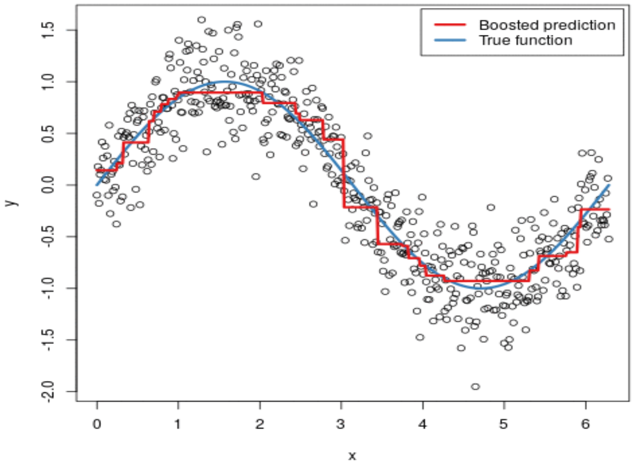
Source:HOML

Boosting in action



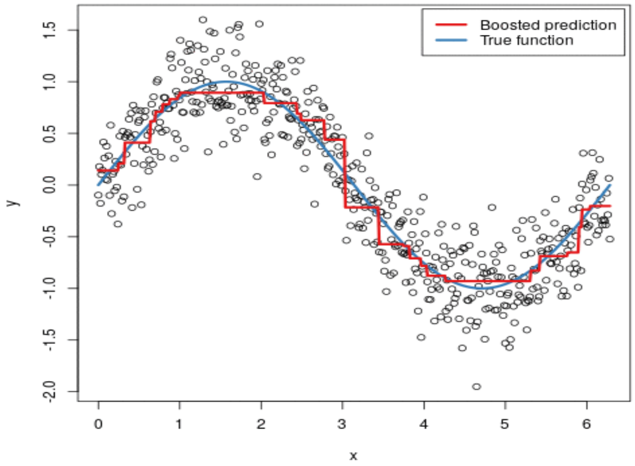
Source:HOML

Boosting in action



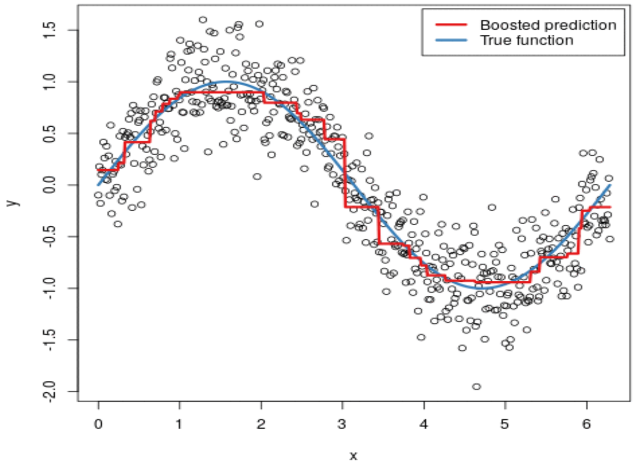
Source:HOML

Boosting in action



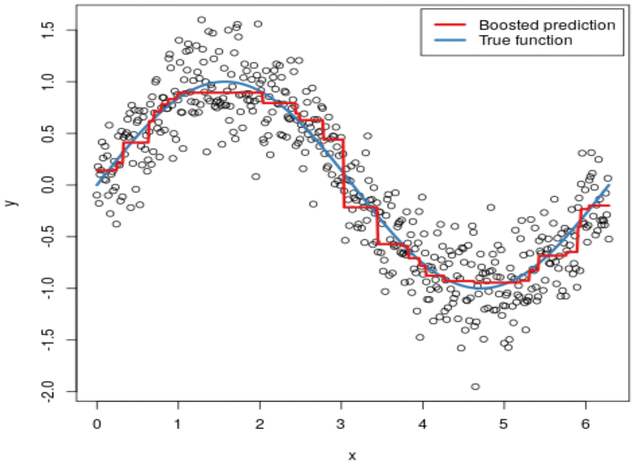
Source:HOML

Boosting in action



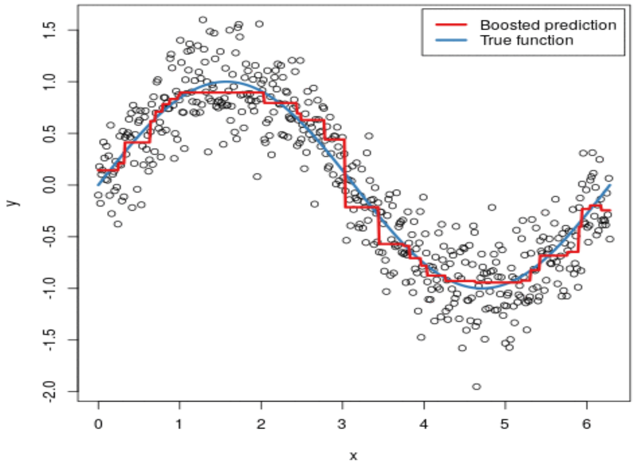
Source:HOML

Boosting in action



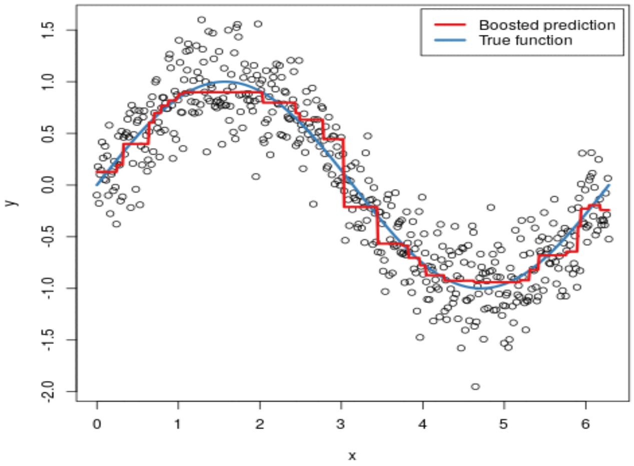
Source:HOML

Boosting in action



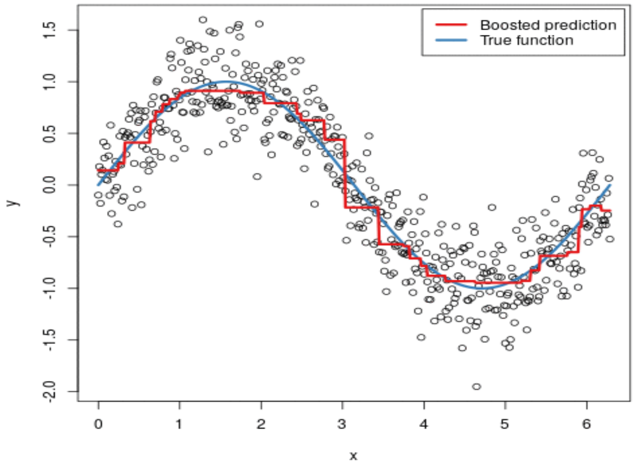
Source:HOML

Boosting in action



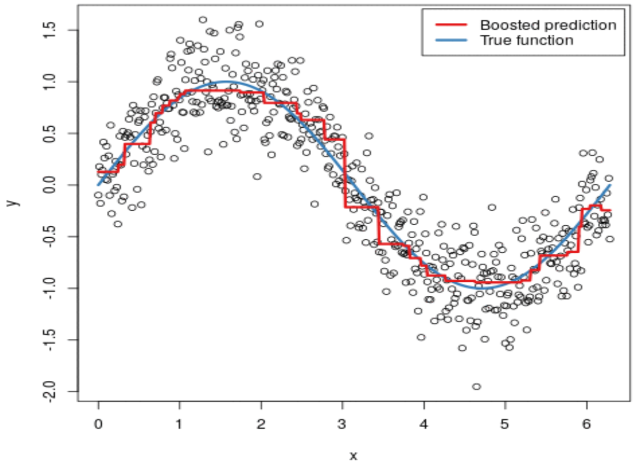
Source:HOML

Boosting in action



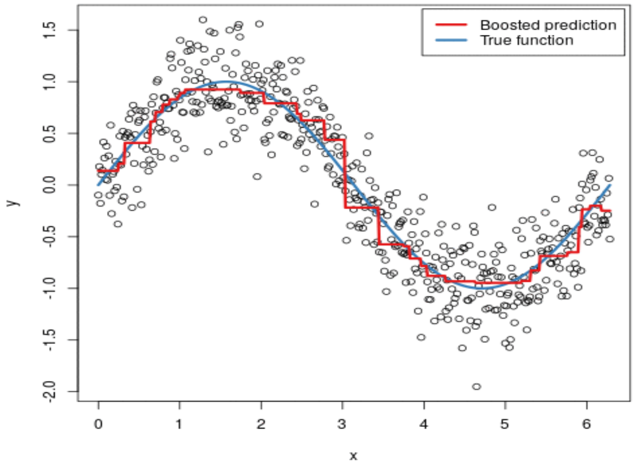
Source:HOML

Boosting in action



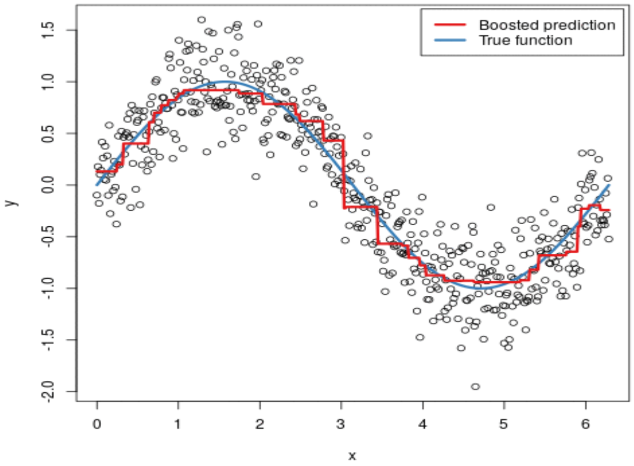
Source:HOML

Boosting in action



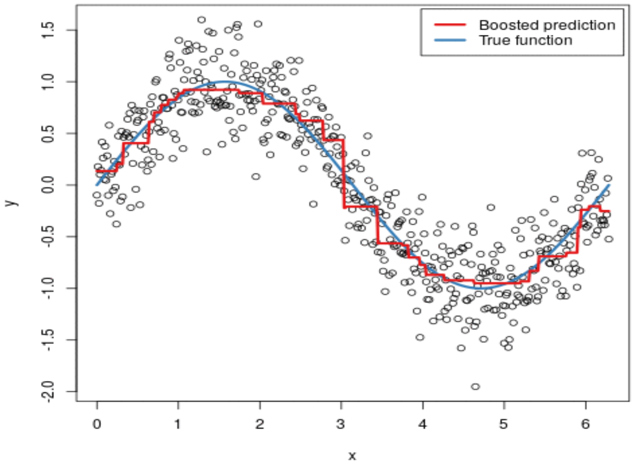
Source:HOML

Boosting in action



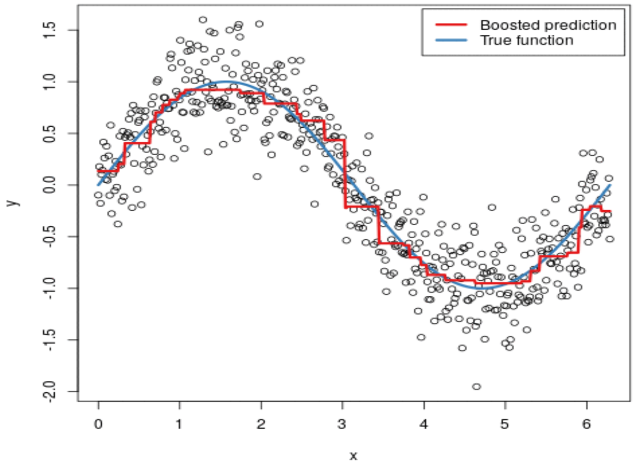
Source:HOML

Boosting in action



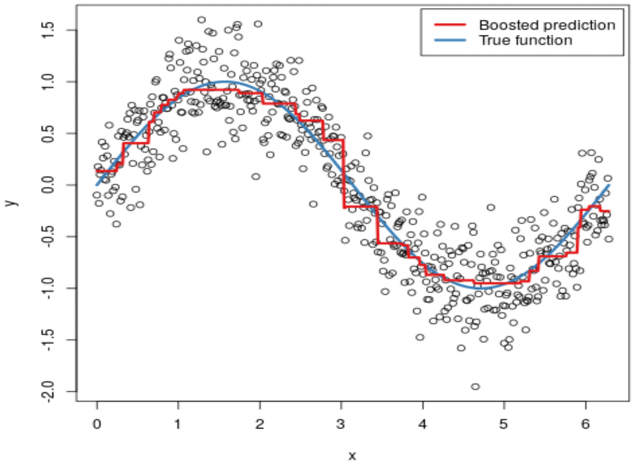
Source:HOML

Boosting in action



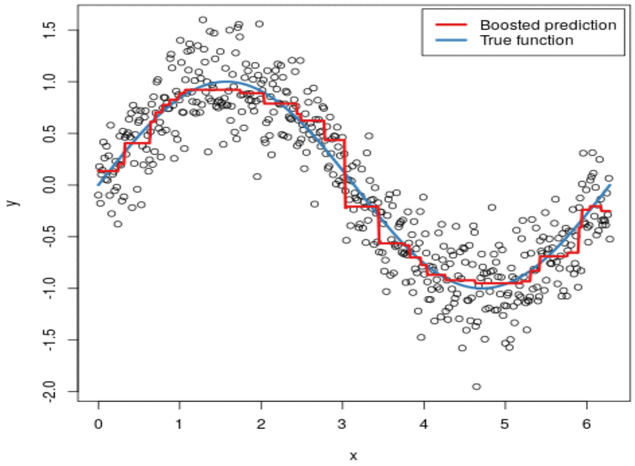
Source:HOML

Boosting in action



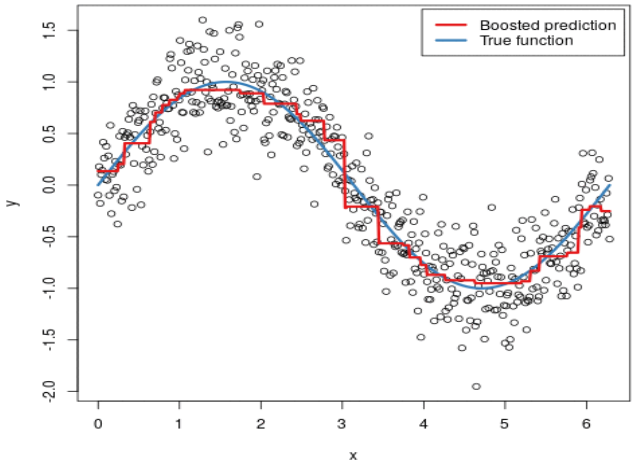
Source:HOML

Boosting in action



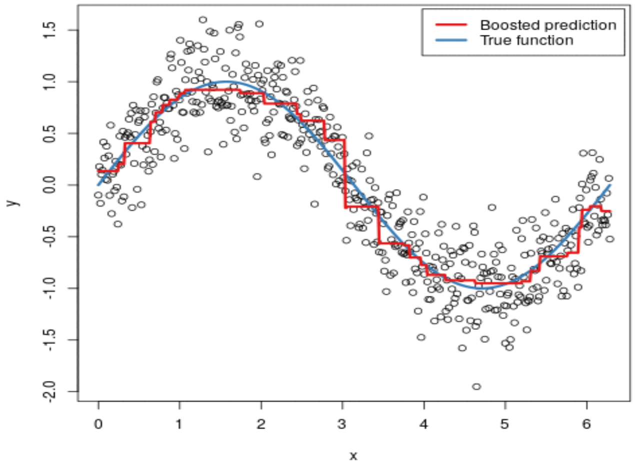
Source:HOML

Boosting in action



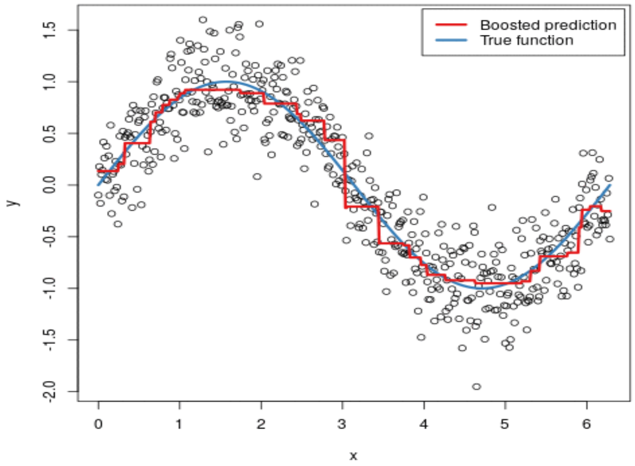
Source:HOML

Boosting in action



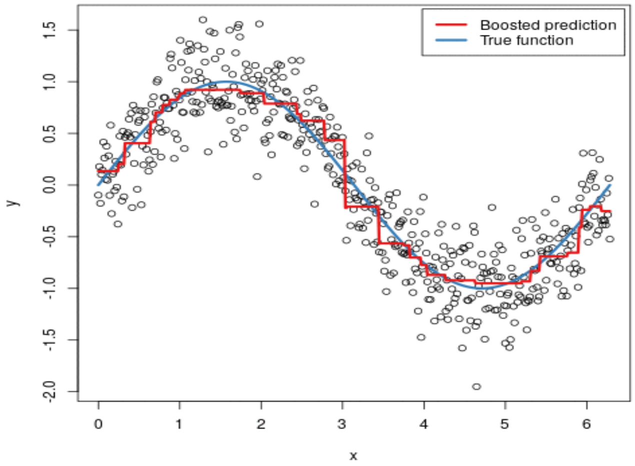
Source:HOML

Boosting in action



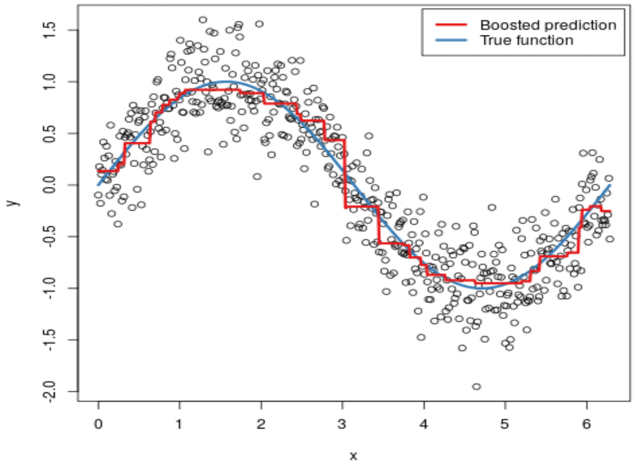
Source:HOML

Boosting in action



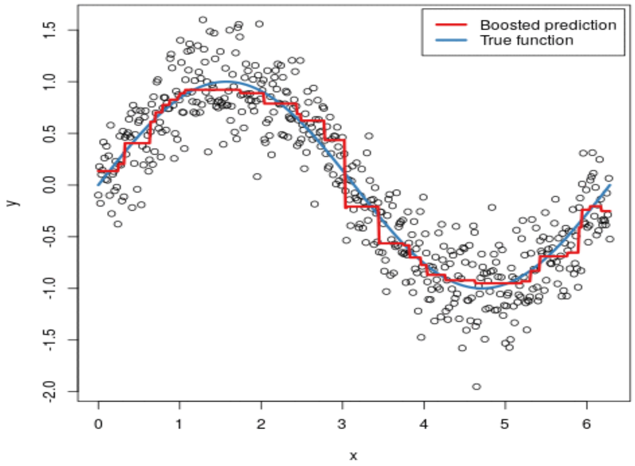
Source:HOML

Boosting in action



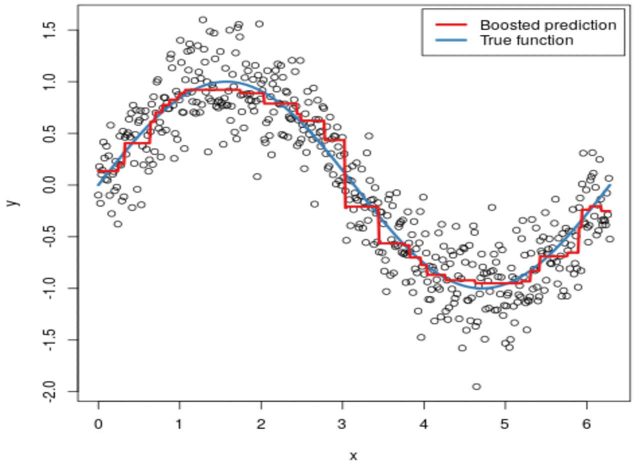
Source:HOML

Boosting in action



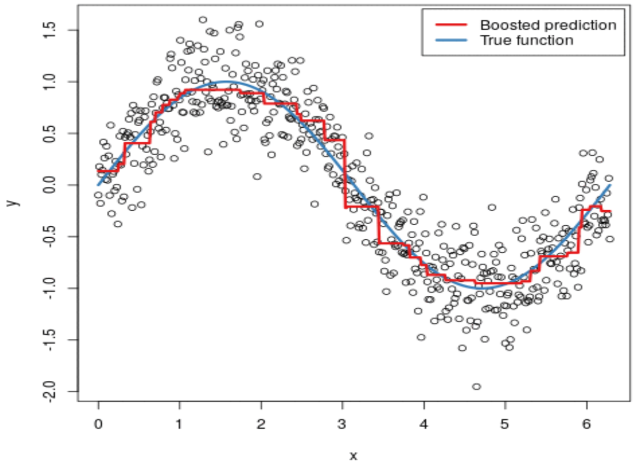
Source:HOML

Boosting in action



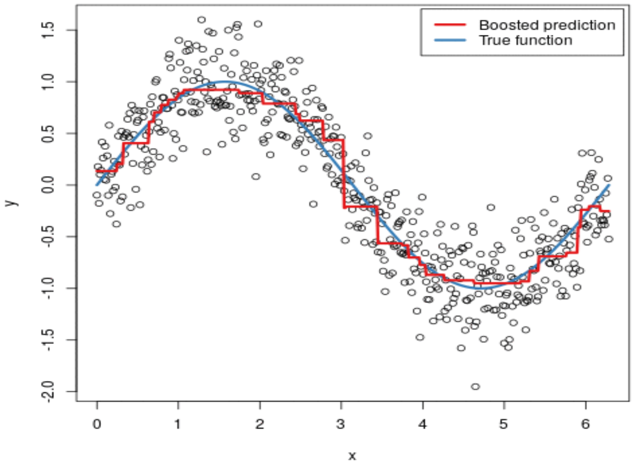
Source:HOML

Boosting in action



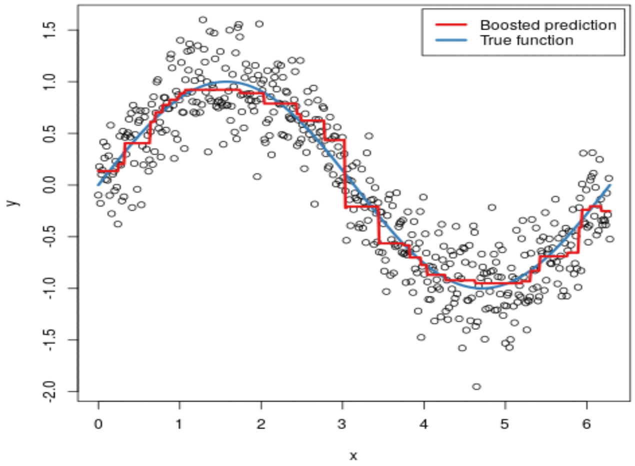
Source:HOML

Boosting in action



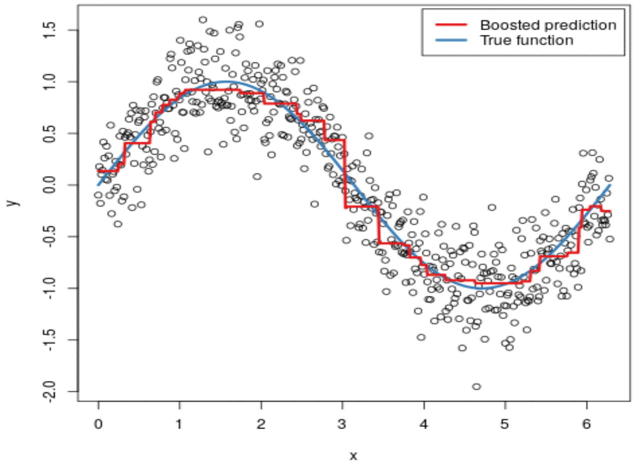
Source:HOML

Boosting in action



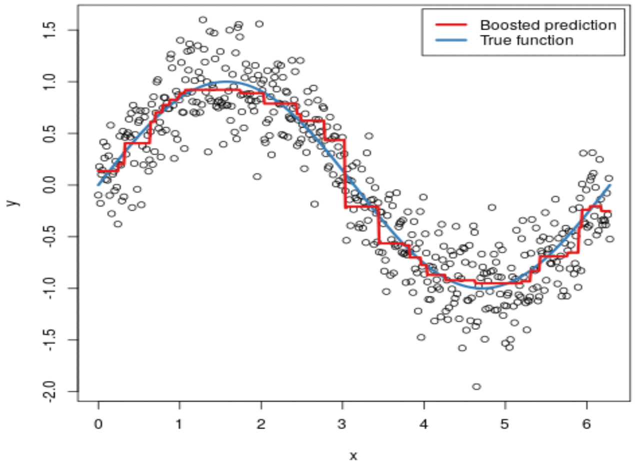
Source:HOML

Boosting in action



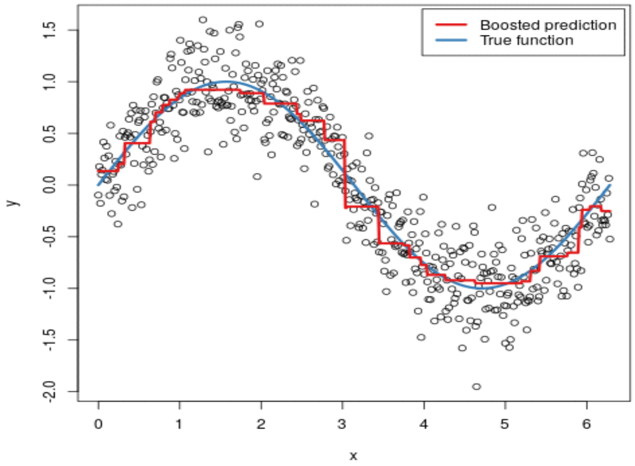
Source:HOML

Boosting in action



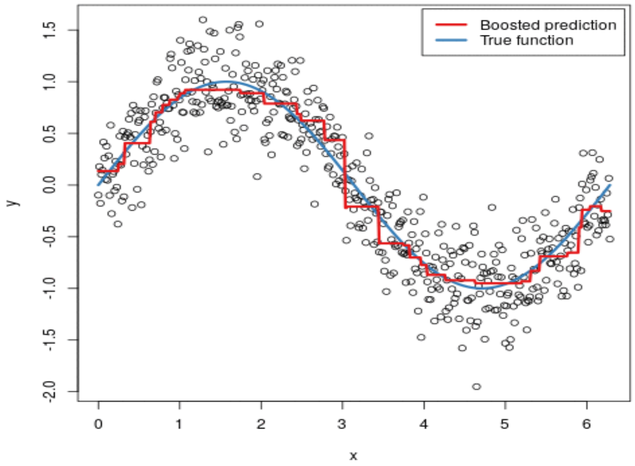
Source:HOML

Boosting in action



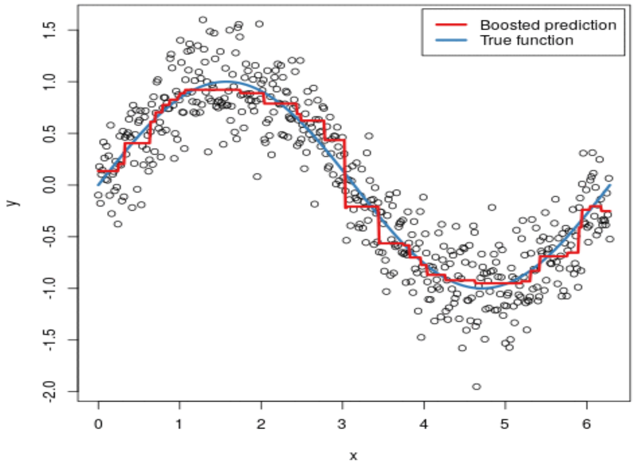
Source:HOML

Boosting in action



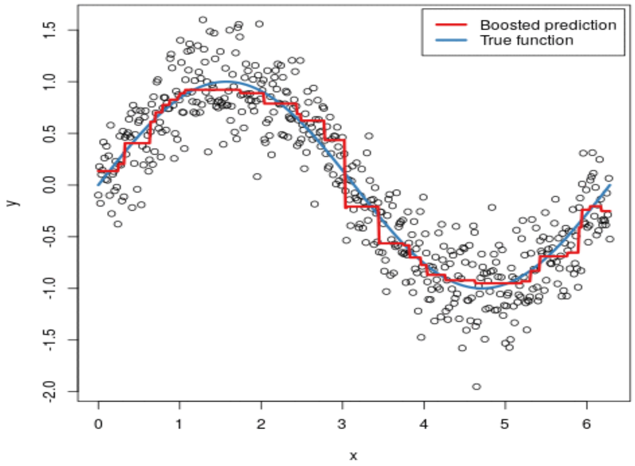
Source:HOML

Boosting in action



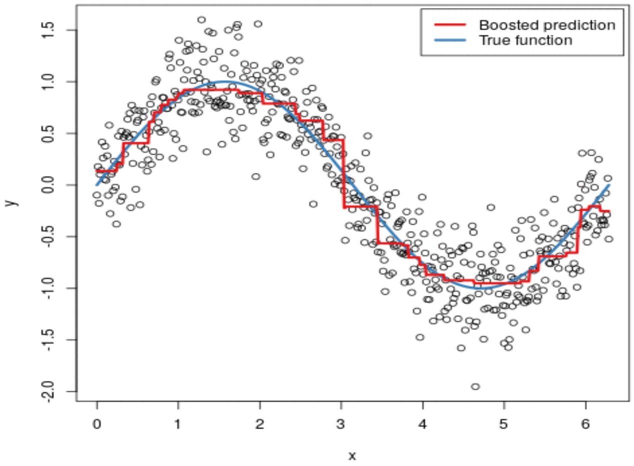
Source:HOML

Boosting in action



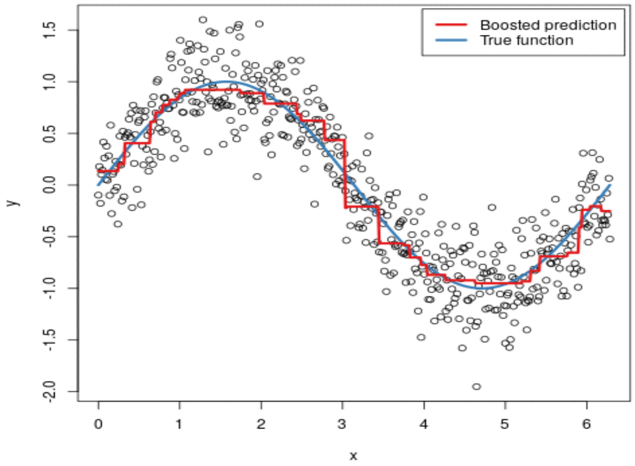
Source:HOML

Boosting in action



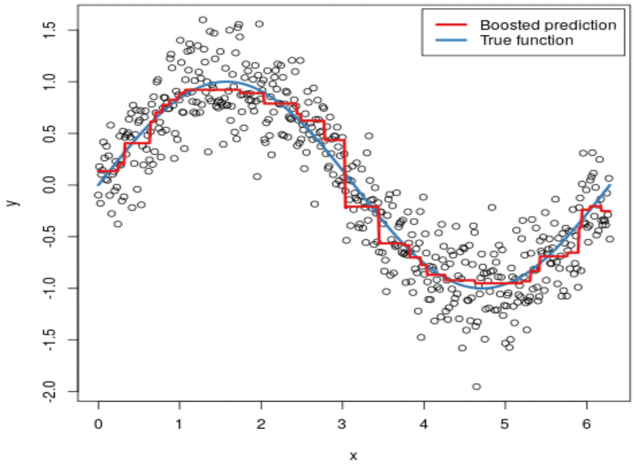
Source:HOML

Boosting in action



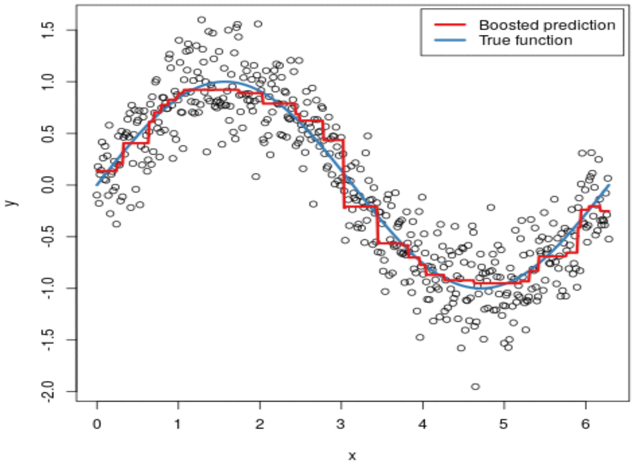
Source:HOML

Boosting in action



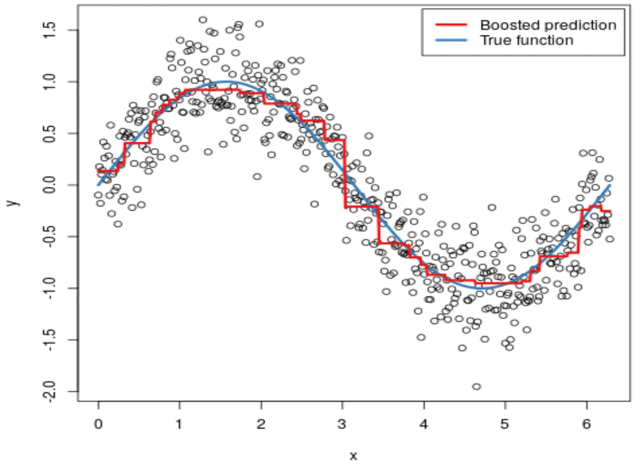
Source:HOML

Boosting in action



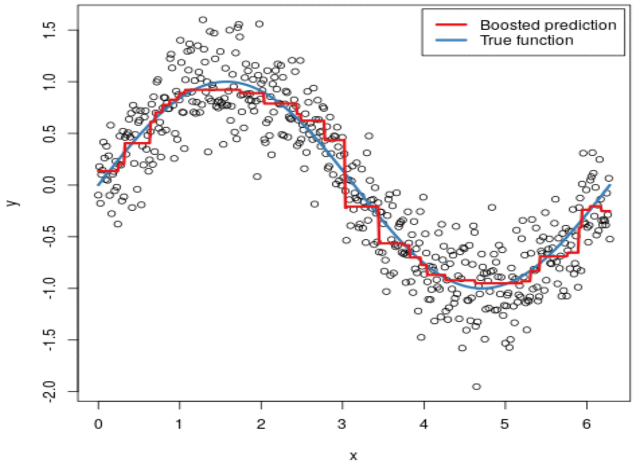
Source:HOML

Boosting in action



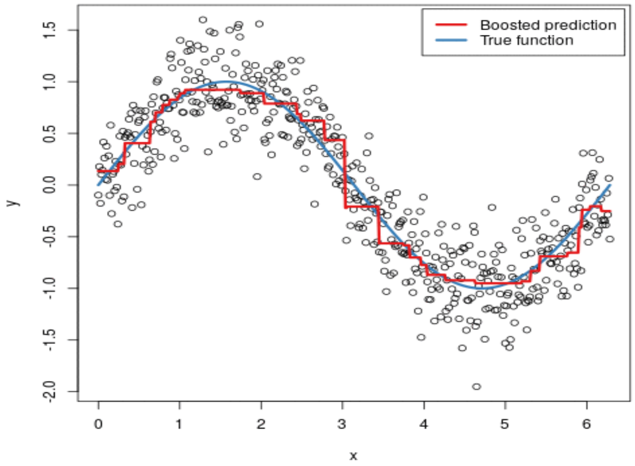
Source:HOML

Boosting in action



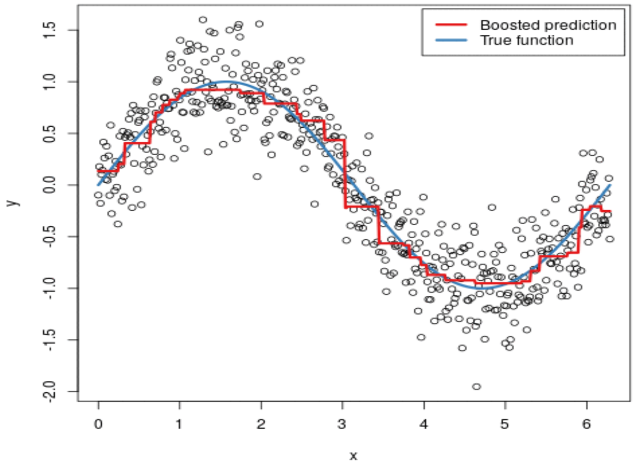
Source:HOML

Boosting in action



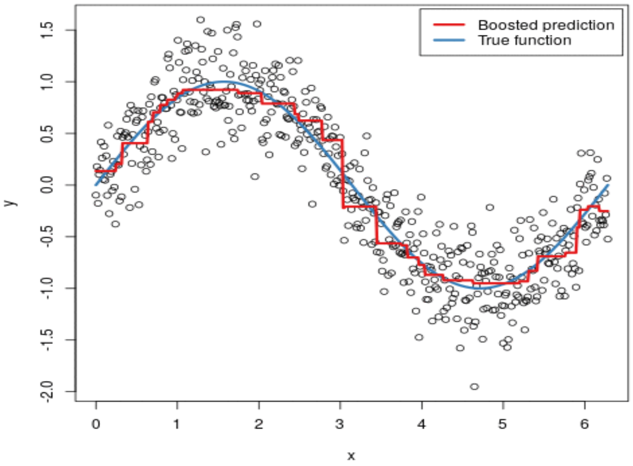
Source:HOML

Boosting in action



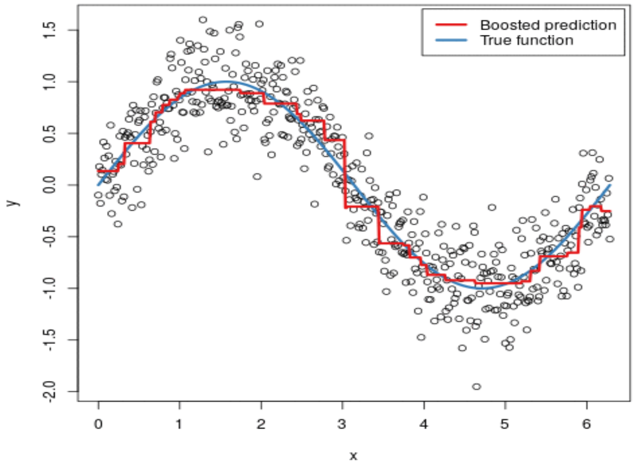
Source:HOML

Boosting in action



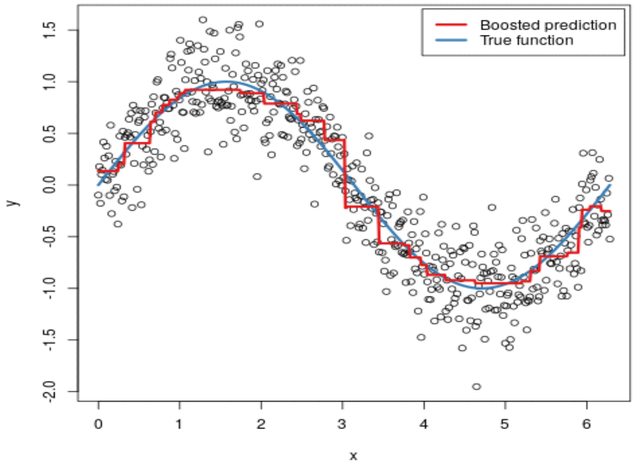
Source:HOML

Boosting in action



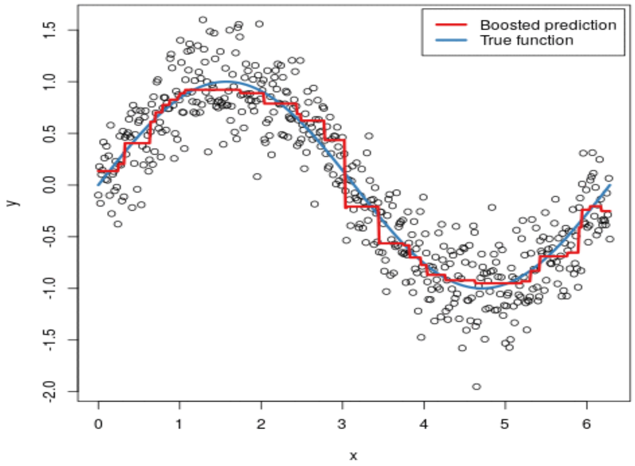
Source:HOML

Boosting in action



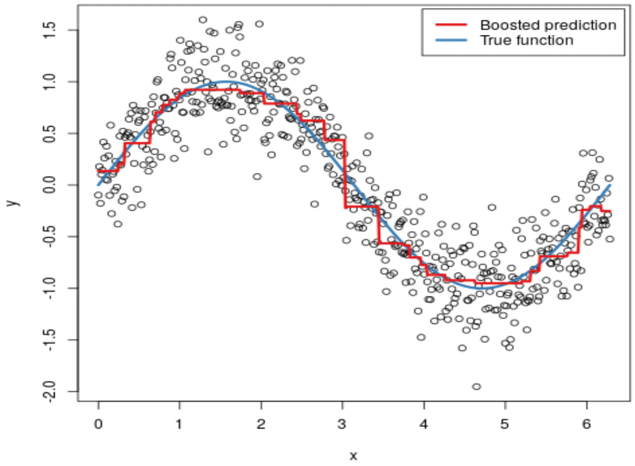
Source:HOML

Boosting in action



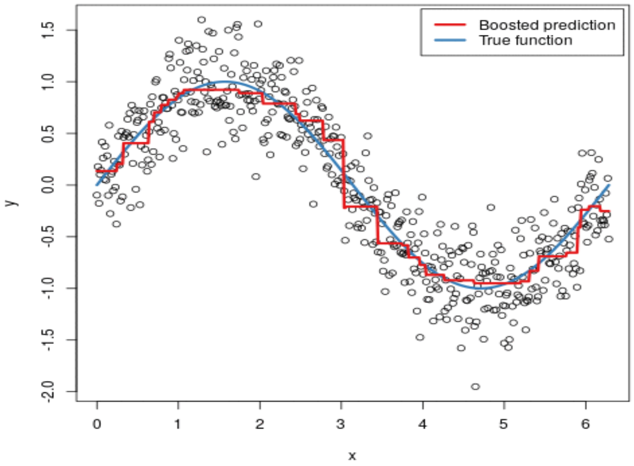
Source:HOML

Boosting in action



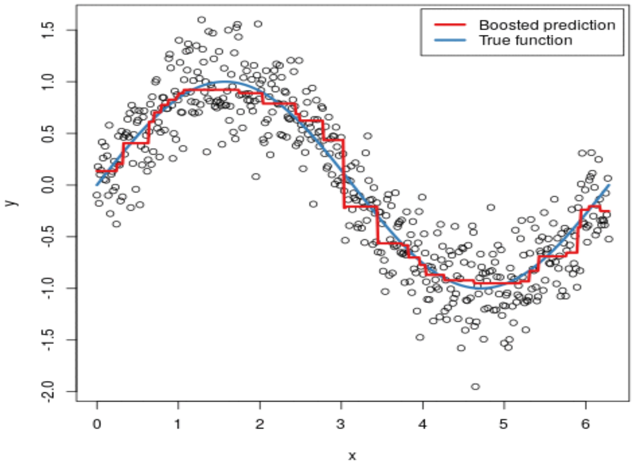
Source:HOML

Boosting in action



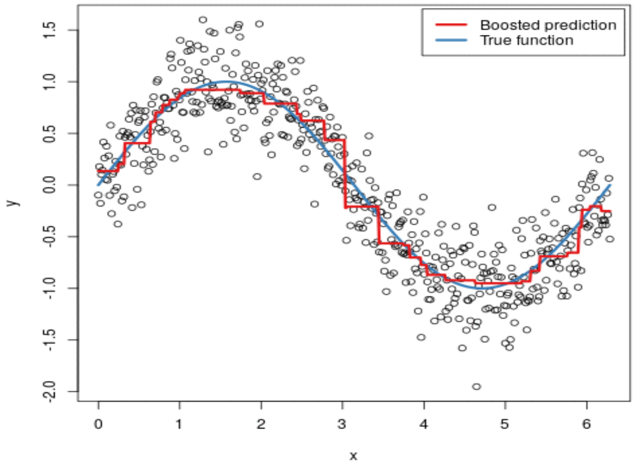
Source:HOML

Boosting in action



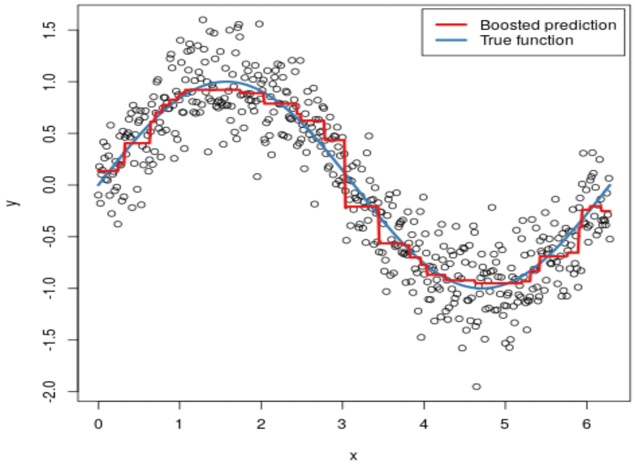
Source:HOML

Boosting in action



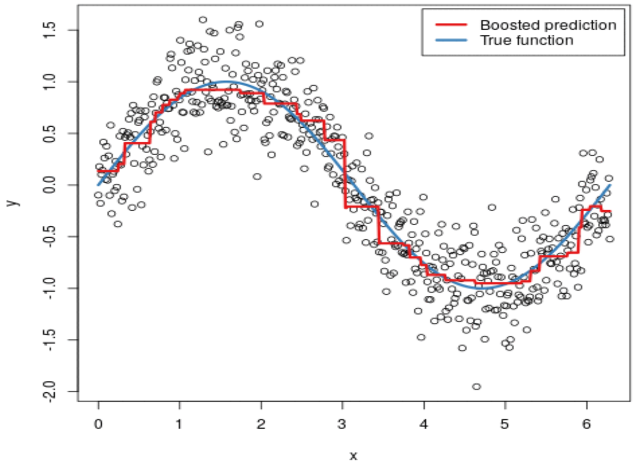
Source:HOML

Boosting in action



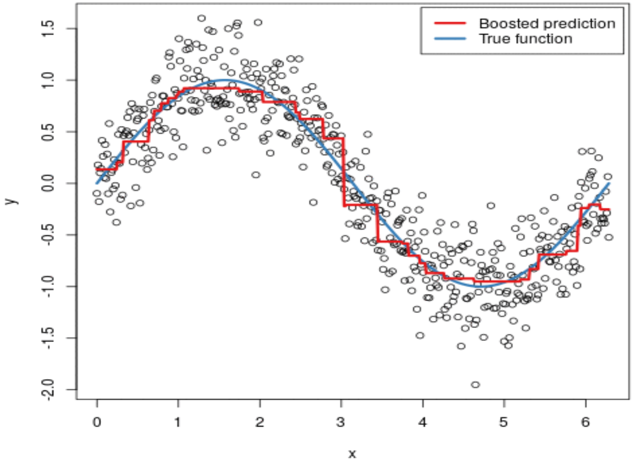
Source:HOML

Boosting in action



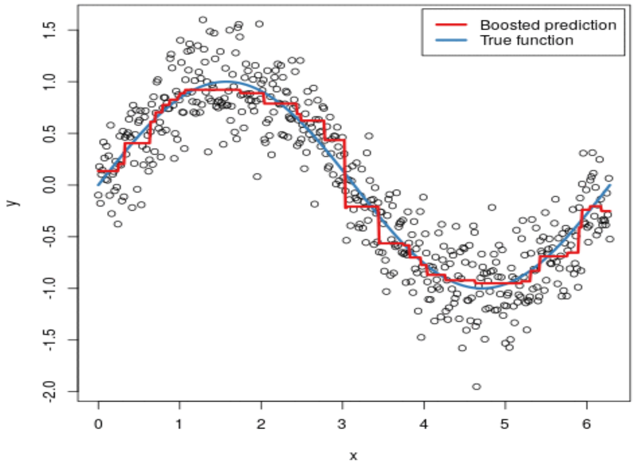
Source:HOML

Boosting in action



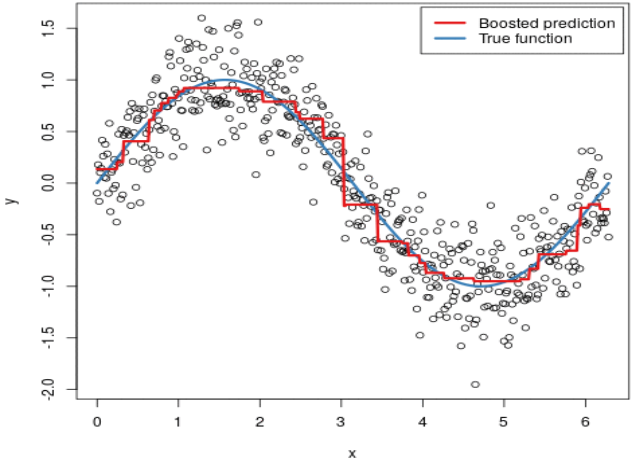
Source:HOML

Boosting in action



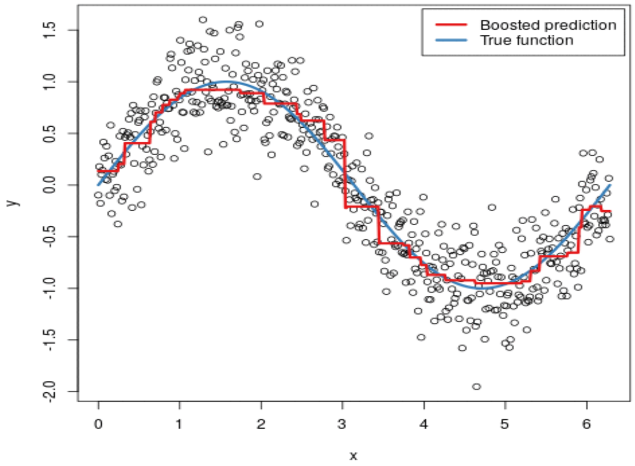
Source:HOML

Boosting in action



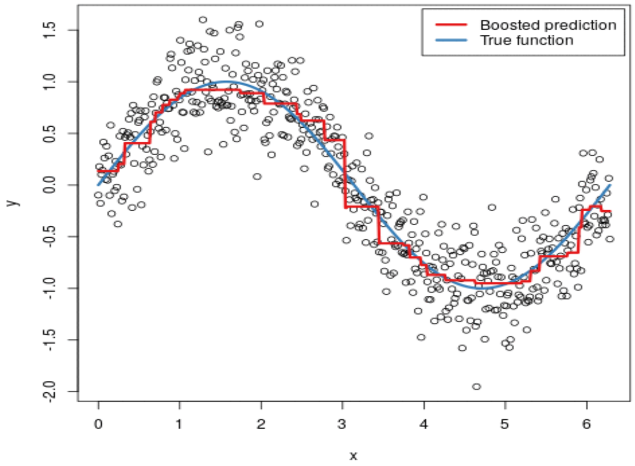
Source:HOML

Boosting in action



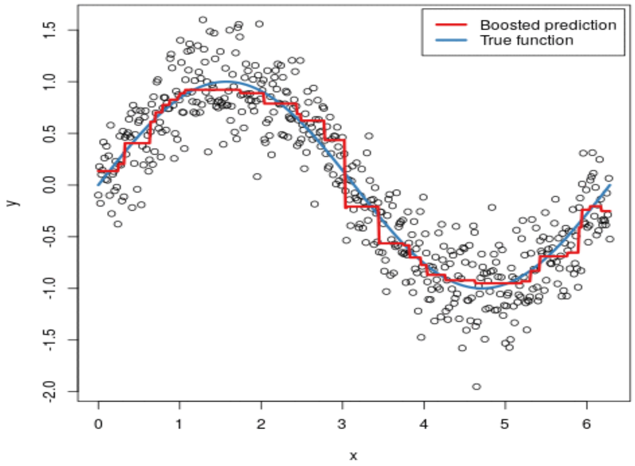
Source:HOML

Boosting in action



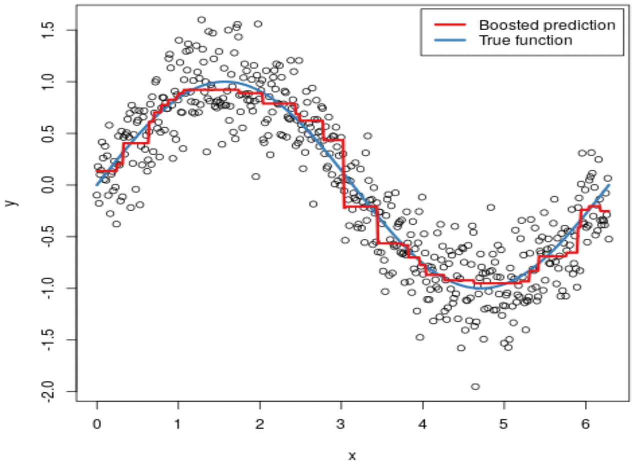
Source:HOML

Boosting in action



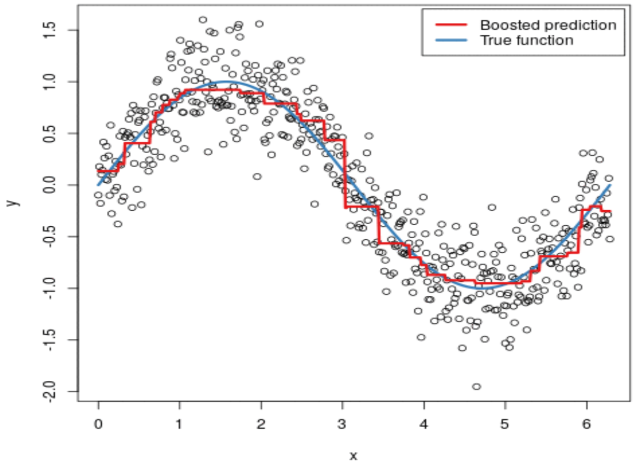
Source:HOML

Boosting in action



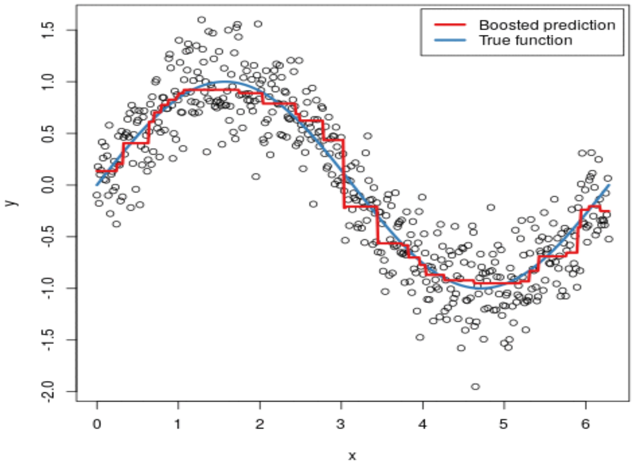
Source:HOML

Boosting in action



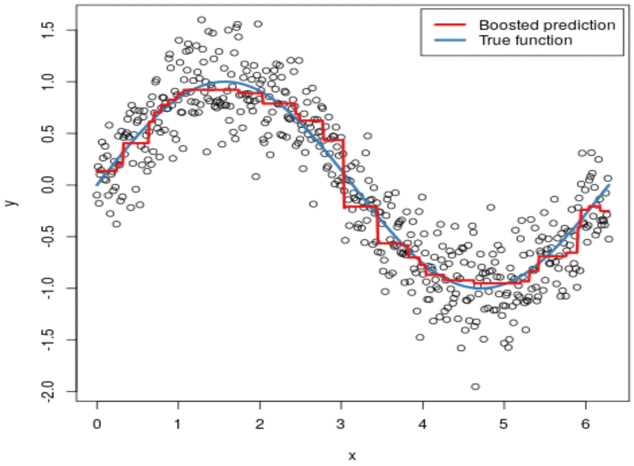
Source:HOML

Boosting in action



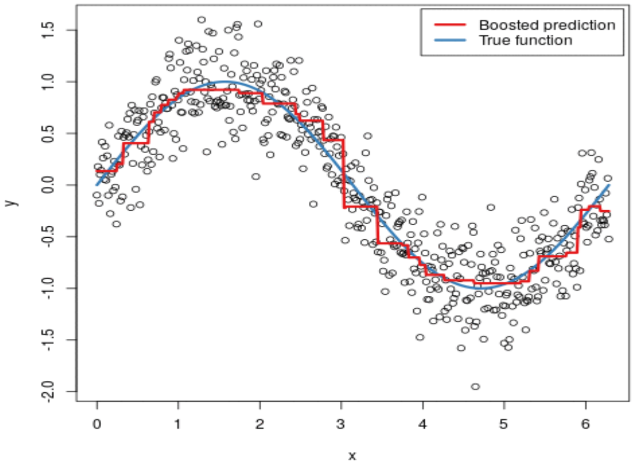
Source:HOML

Boosting in action



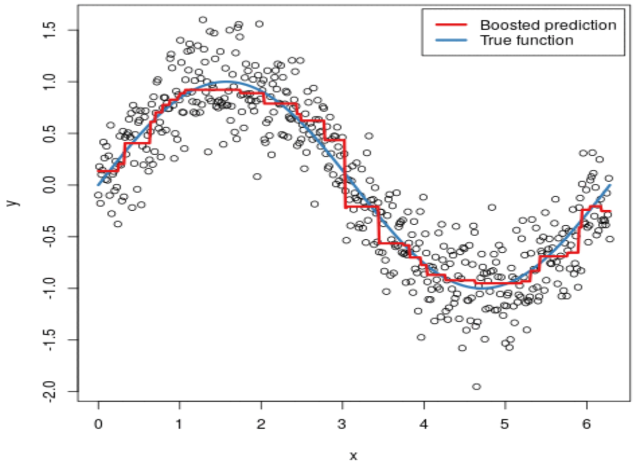
Source:HOML

Boosting in action



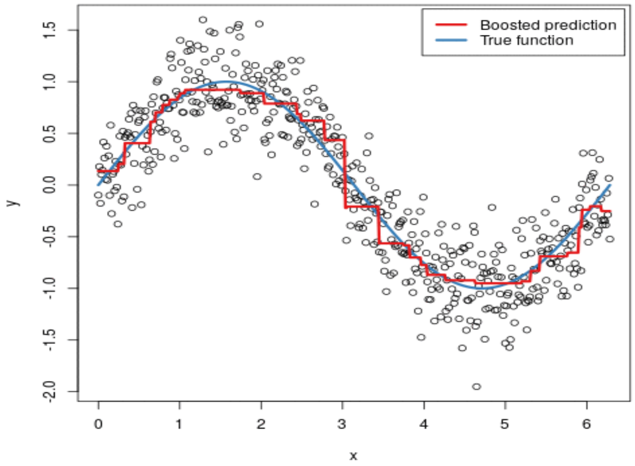
Source:HOML

Boosting in action



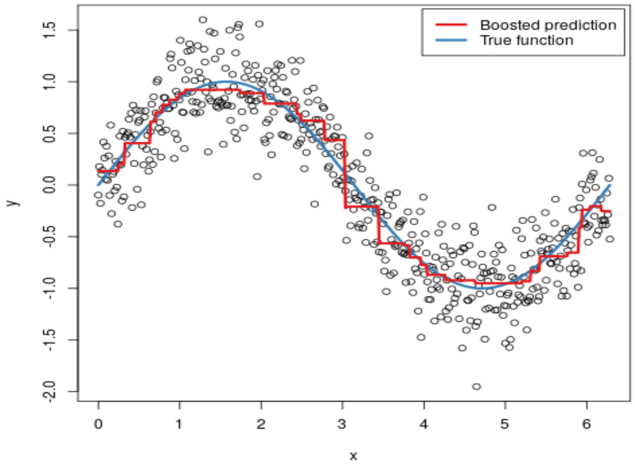
Source:HOML

Boosting in action



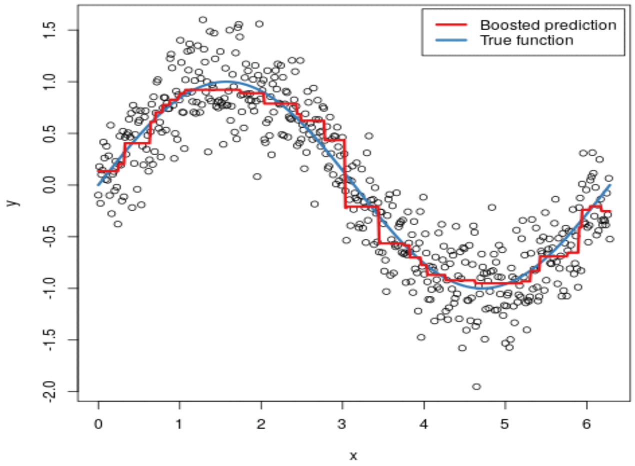
Source:HOML

Boosting in action



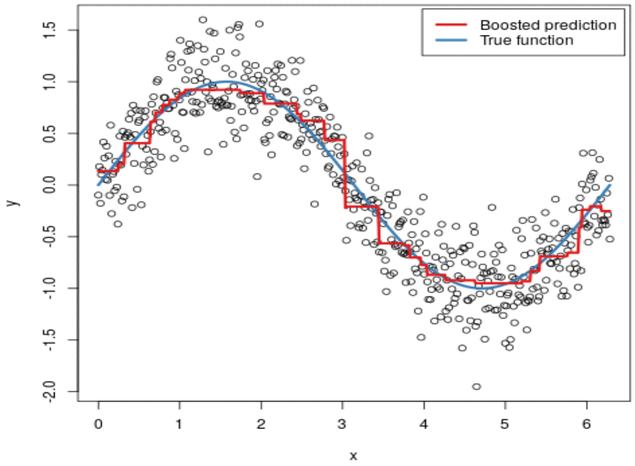
Source:HOML

Boosting in action



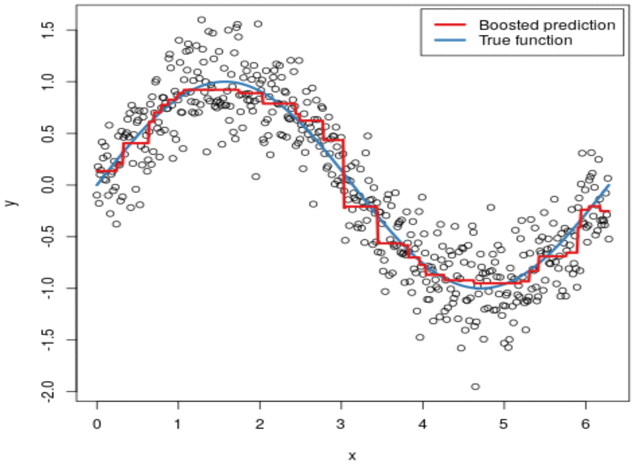
Source:HOML

Boosting in action



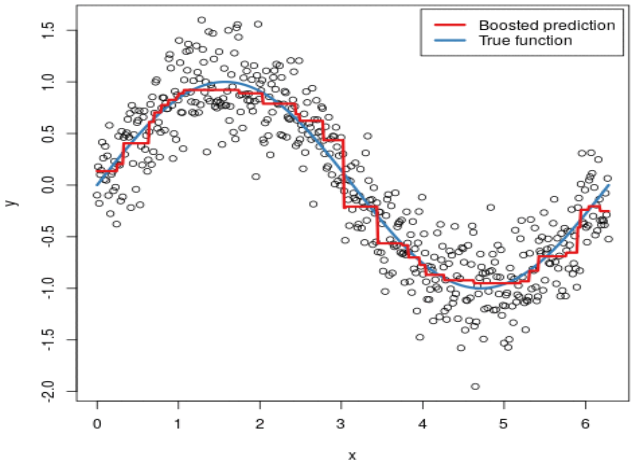
Source:HOML

Boosting in action



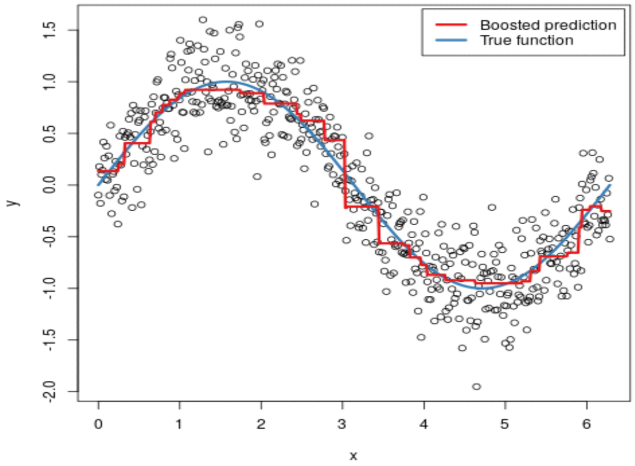
Source:HOML

Boosting in action



Source:HOML

Boosting in action



Source:HOML

Boosting

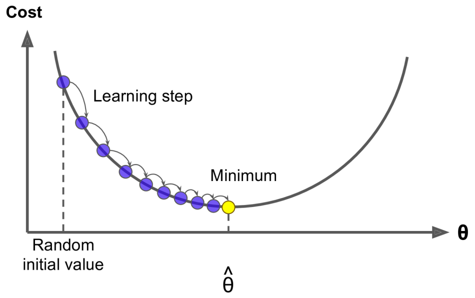
- Many algorithms, including decision trees, focus on minimizing the residuals and, therefore, emphasize the MSE loss function.
- However, often we wish to focus on other loss functions such as mean absolute error (MAE) or to be able to apply the method to a classification problem with a loss function such as deviance.
- The name gradient boosting machines come from the fact that this procedure can be generalized to loss functions other than MSE.

Gradient Descent

- Gradient boosting is considered a **gradient descent algorithm**.
- Gradient descent is a very generic optimization algorithm capable of finding optimal solutions to a wide range of problems.
- The general idea of gradient descent is to tweak parameters iteratively in order to minimize a cost function.

Gradient Descent

- Gradient descent measures the local gradient of the loss (cost) function for a given set of parameters (Θ) and takes steps in the direction of the descending gradient.
- Once the gradient is zero, we have reached the minimum.



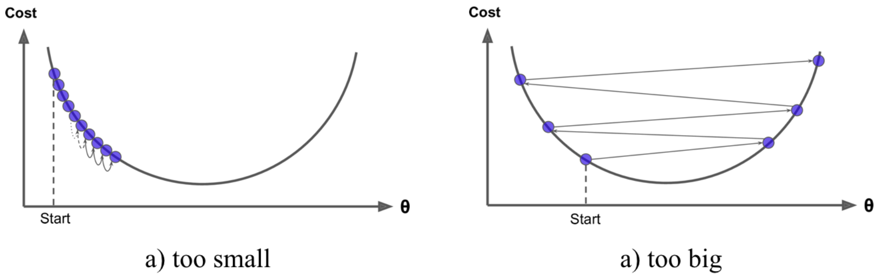
Source: HOML

Gradient Descent

- Gradient descent can be performed on any loss function that is differentiable.
- Consequently, this allows GBMs to optimize different loss functions as desired
- An important parameter in gradient descent is the size of the steps which is determined by the **learning rate**.

Gradient Descent

- If the learning rate is too small, then the algorithm will take many iterations to find the minimum. On the other hand, if the learning rate is too high, you might jump cross the minimum and end up further away than when you started.



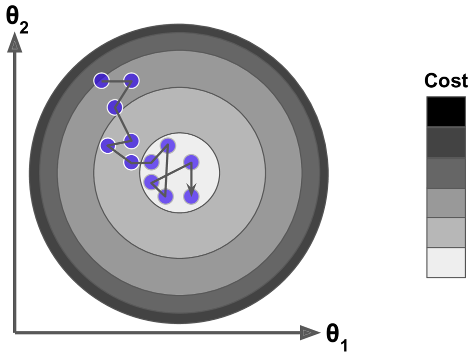
Source: HOML

Gradient Descent

- Moreover, not all cost functions are convex (bowl shaped).
- There may be local minimas, plateaus, and other irregular terrain of the loss function that makes finding the global minimum difficult.
- **Stochastic gradient descent** can help us address this problem by sampling a fraction of the training observations (typically without replacement) and growing the next tree using that subsample.

Gradient Descent

- **Stochastic gradient descent** will often find a near-optimal solution by jumping out of local minimas and off plateaus.



Source: HOML

Tuning

- Part of the beauty and challenges of GBM is that they offer several tuning parameters.
- The beauty in this is GBMs are highly flexible.
- The challenge is that they can be time consuming to tune and find the optimal combination of hyperparameters.

Tuning

- The most common hyperparameters that you will find in most GBM implementations include:
  1. B: the number of trees to fit.
  2. λ: shrinkage parameter - controls the learning rate. Typical values are 0.01 or 0.001, and the right choice can depend on the problem.
  3. $interaction.depth$: number of splits in each tree - tree's complexity. Often $interaction.depth = 1$ works well, in which case each tree is a stump, consisting of a single split.
  4. $Subsampling$: Controls whether or not you use a fraction of the available training observations. Using less than 100% of the training observations means you are implementing stochastic gradient descent.

References

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2017)
An Introduction to Statistical Learning
Springer.
https://www.statlearning.com/

Ed Rubin (2020)
Economics 524 (424): Prediction and Machine-Learning in Econometrics
Univ, of Oregon.