Gradient Boosting

Jan-Philipp Kolb

21 Mai, 2019

Gradient Boosting Machines (GBMs)

- The following slides are based on UC Business Analytics R Programming Guide on **gbm regression**
- GBMs are extremely popular, successful across many domains and one of the leading methods for winning **Kaggle competitions**.
- GBMs build an ensemble of flat and weak successive trees with each tree learning and improving on the previous.
- When combined, these many weak successive trees produce a powerful "committee" that are often hard to beat with other algorithms.

```
library(rsample)      # data splitting
library(gbm)          # basic implementation
library(xgboost)      # a faster implementation of gbm
library(caret)        # an aggregator package for machine learni
library(h2o)          # a java-based platform
library(pdp)          # model visualization
library(ggplot2)      # model visualization
library(lime)         # model visualization
```

The dataset

```
set.seed(123)
ames_split <- initial_split(AmesHousing::make_ames(), prop = .7)
ames_train <- training(ames_split)
ames_test  <- testing(ames_split)
```

Advantages

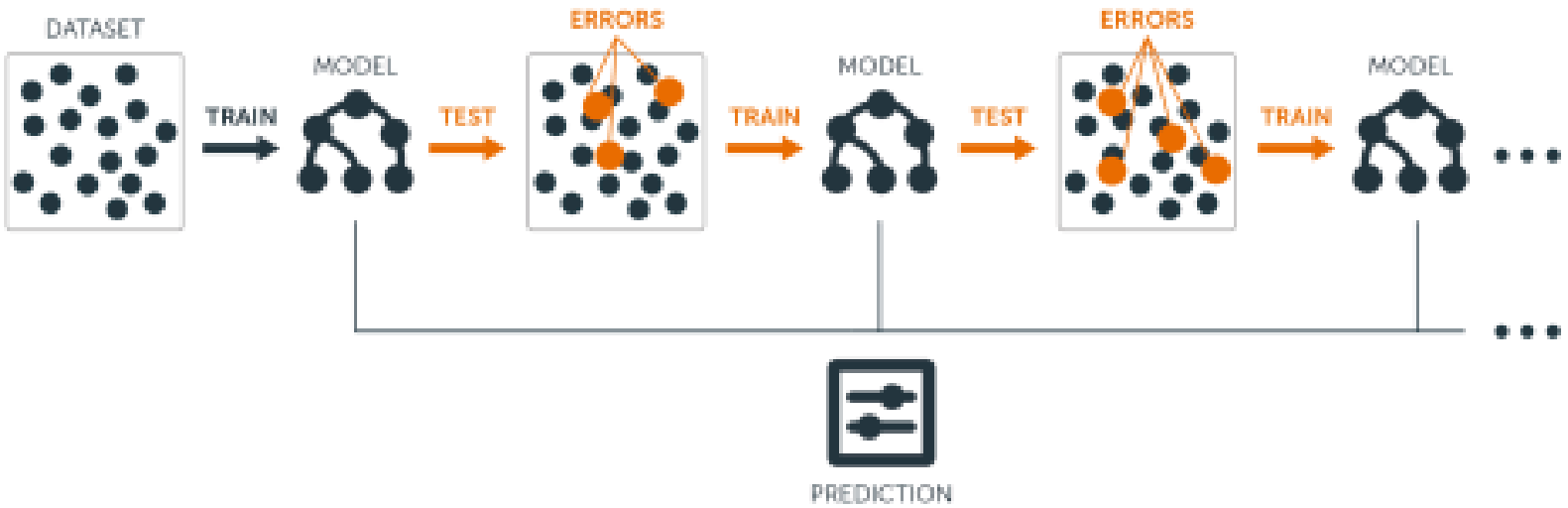
- Often provides predictive accuracy that cannot be beat.
- Lots of flexibility - can optimize on different loss functions and provides several hyperparameter tuning options that make the function fit very flexible.
- No data pre-processing required - often works great with categorical and numerical values as is.
- Handles missing data - imputation not required.

Disadvantages of GBMs

- GBMs will continue improving to minimize all errors. This can overemphasize outliers and cause overfitting. Must use cross-validation to neutralize.
- Computationally expensive - GBMs often require many trees ($>$1000) which can be time and memory exhaustive.
- The high flexibility results in many parameters that interact and influence heavily the behavior of the approach (number of iterations, tree depth, regularization parameters, etc.). This requires a large grid search during tuning.
- Less interpretable although this is easily addressed with various tools (variable importance, partial dependence plots, LIME, etc.).

The idea of GBMs

- Many machine learning models are founded on a single predictive model (i.e. linear regression, penalized models, naive Bayes, svm).
- Other approaches (bagging, random forests) are built on the idea of building an ensemble of models where each individual model predicts the outcome and the ensemble simply averages the predicted values.
- The idea of boosting is to add models to the ensemble sequentially.
- At each particular iteration, a new weak, base-learner model is trained with respect to the error of the whole ensemble learnt so far.



Important concepts

Base-learning models

- Boosting is a framework that iteratively improves any weak learning model.
- Many gradient boosting applications allow you to "plug in" various classes of weak learners at your disposal.
- In practice, boosted algorithms often use decision trees as the base-learner.

Training weak models

- A weak model is one whose error rate is only slightly better than random guessing.
- The idea behind boosting is that each sequential model builds a simple weak model to slightly improve the remaining errors.
- Shallow trees represent a weak learner.
- Trees with only 1-6 splits are used.
- Combining many weak models (versus strong ones) has a few benefits:
- Speed: Constructing weak models is computationally cheap.
- Accuracy improvement: Weak models allow the algorithm to learn slowly; making minor adjustments in new areas where it does not perform well. In general, statistical approaches that learn slowly tend to perform well.
- Avoids overfitting: Due to making only small incremental improvements with each model in the ensemble, this allows us to stop the learning process as soon as overfitting has been detected (typically by using cross-validation).

Sequential training with respect to errors

- Boosted trees are grown sequentially;
- each tree is grown using information from previously grown trees.
- The basic algorithm for boosted regression trees can be generalized to the following where x represents our features and y represents our response:

1.) Fit a decission tree: $F_1(x) = y$

2.) the next decission tree is fixed to the residuals of the previous: $h_1(x) = y - F_1(x)$

3.) Add this new tree to our algorithm: $F_2(x) = F_1(x) + h_1(x)$

4.) The next decission tree is fixed to the residuals of $F_2 : h_2(x) = y - F_2(x)$

5.) Add the new tree to the algorithm: $F_3(x) = F_2(x) + h_1(x)$

Continue this process until some mechanism (i.e. cross validation) tells us to stop.

Basic algorithm for boosted regression trees

The basic algorithm for boosted regression trees can be generalized to the following where the final model is simply a stagewise additive model of b individual regression trees:

$$f(x) = B \sum_{b=1}^{B} f^b(x)$$

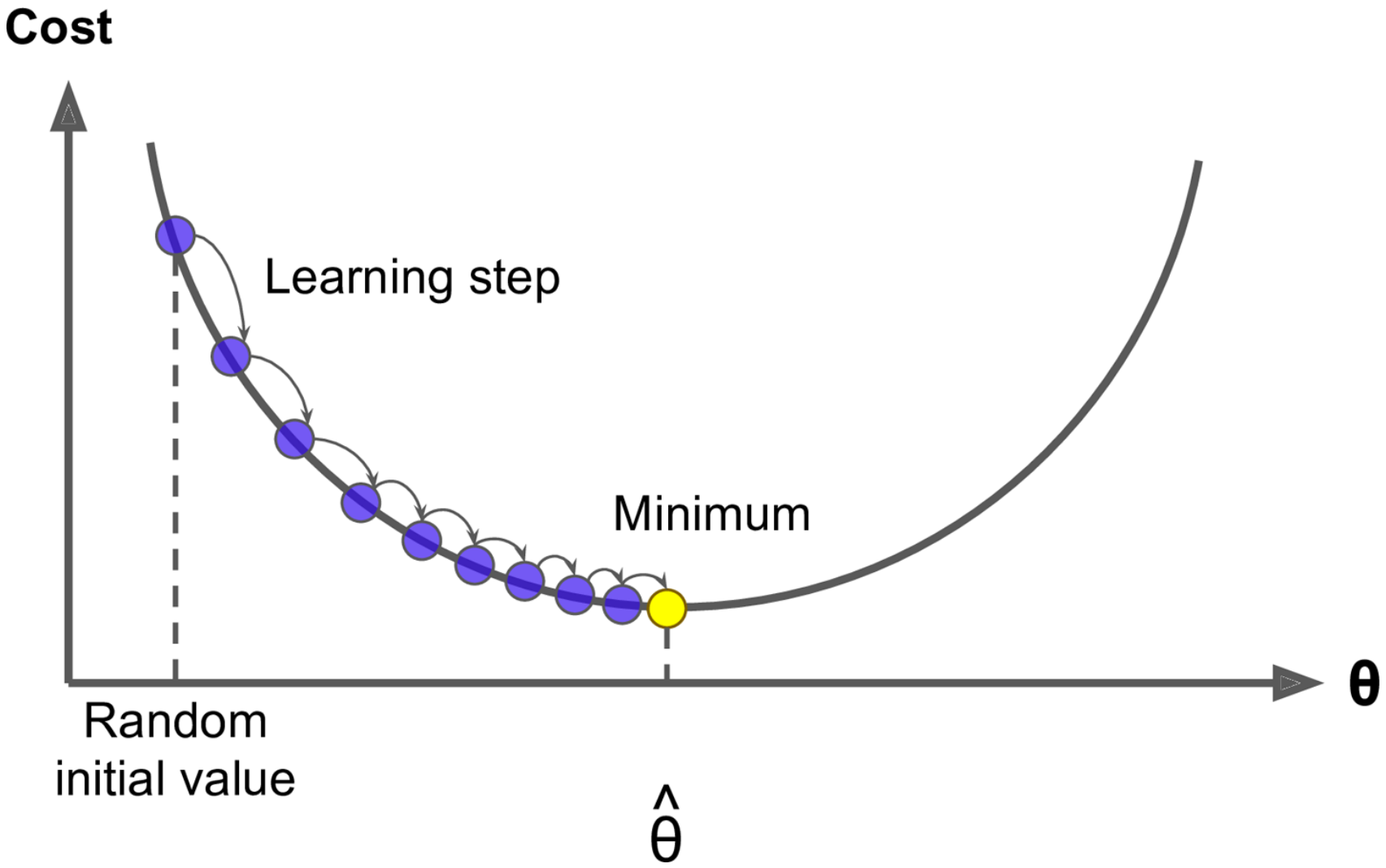
Gradient descent

- Many algorithms, including decision trees, focus on minimizing the residuals and, therefore, emphasize the MSE loss function.
- This algorithm outlines the approach of sequentially fitting regression trees to minimize the errors.
- This specific approach is how gradient boosting minimizes the mean squared error (MSE) loss function.
- 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.

A gradient descent algorithm

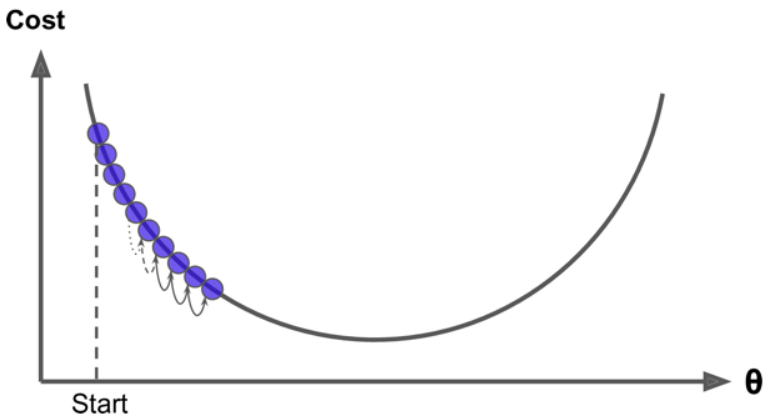
- Gradient boosting is considered a gradient descent algorithm.
- Gradient descent is a very generic optimization algorithm capable of finding optimal solutions to a wide range of problems.
- The general idea of gradient descent is to tweak parameters iteratively in order to minimize a cost function.
- Suppose you are a downhill skier racing your friend.
- A good strategy to beat your friend to the bottom is to take the path with the steepest slope.
- This is exactly what gradient descent does - it measures the local gradient of the loss (cost) function for a given set of parameters (Φ) and takes steps in the direction of the descending gradient.
- Once the gradient is zero, we have reached the minimum.

Gradient descent (Geron, 2017).

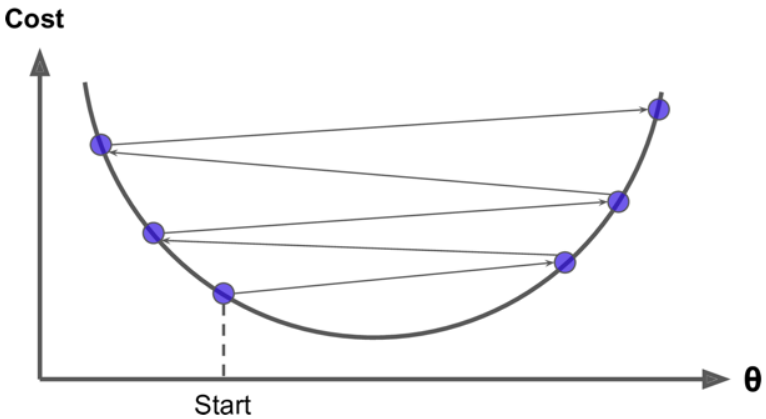


Gradient descent

- Gradient descent can be performed on any loss function that is differentiable.
- 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.
- If the learning rate is too small, then the algorithm will take many iterations to find the minimum.
- But if the learning rate is too high, you might jump cross the minimum and end up further away than when you started.



a) too small

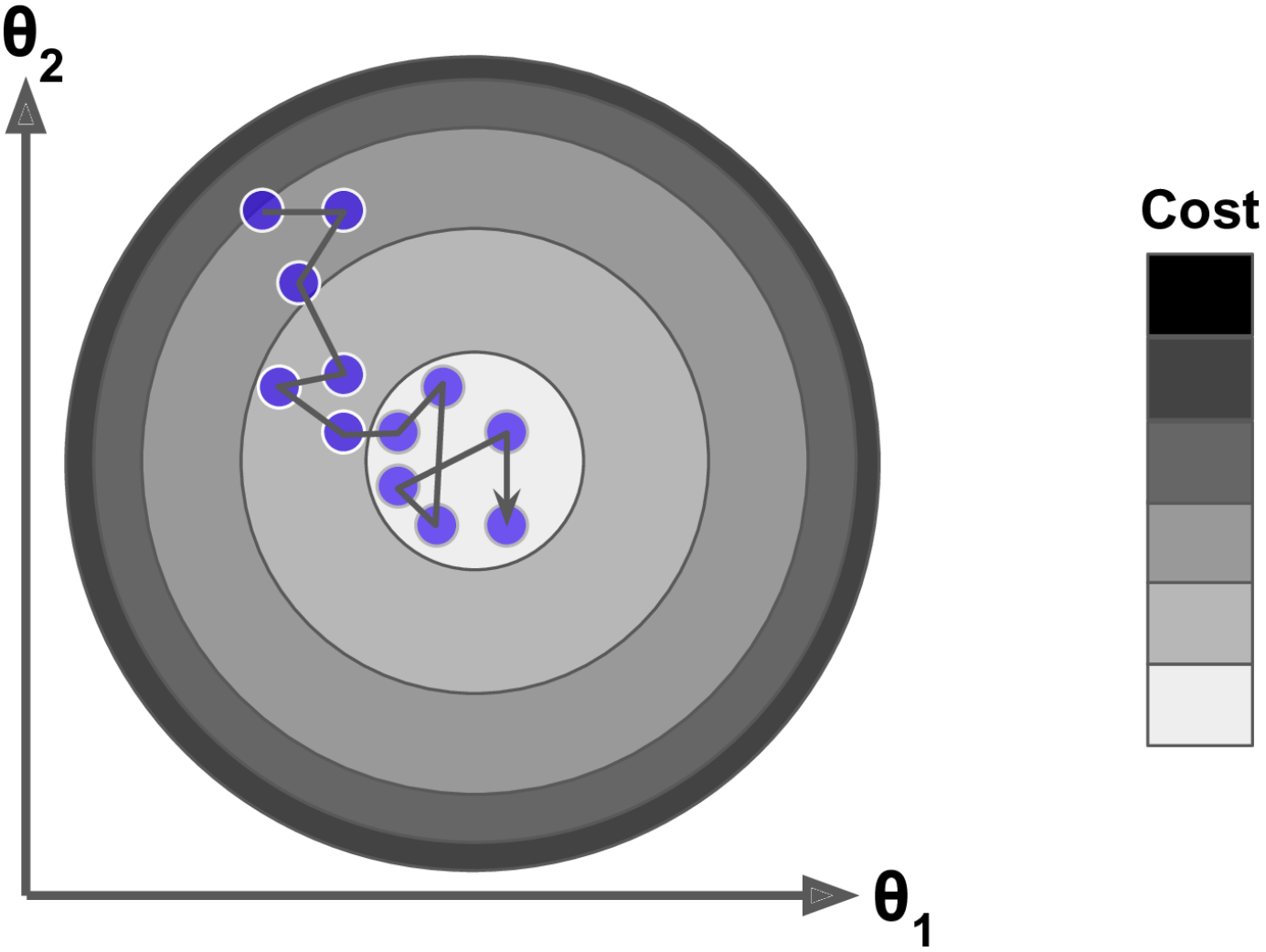


a) too big

Shape of cost functions

- 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.
- This makes the algorithm faster but the stochastic nature of random sampling also adds some random nature in descending the loss function gradient.
- Although this randomness does not allow the algorithm to find the absolute global minimum, it can actually help the algorithm jump out of local minima and off plateaus and get near the global minimum.

Stochastic gradient descent



Tuning GBM

- GBMs are highly flexible - many tuning parameters
- It is time consuming to find the optimal combination of hyperparameters

Number of trees

- GBMs often require many trees; however, unlike random forests GBMs can overfit so the goal is to find the optimal number of trees that minimize the loss function of interest with cross validation.

Tuning parameters

Depth of trees

- The number d of splits in each tree, which controls the complexity of the boosted ensemble.
- Often $d = 1$ works well, in which case each tree is a stump consisting of a single split. More commonly, d is greater than 1 but it is unlikely $d > 10$ will be required.

Learning rate

- The number d of splits in each tree, which controls the complexity of the boosted ensemble.
- Often $d = 1$ works well, in which case each tree is a stump consisting of a single split. -
- Normally, d is greater than 1 but it is unlikely $d > 10$ will be required.

Tuning parameters (II)

Subsampling

- Controls if a fraction of the available training observations is used.
- Using less than 100% of the training observations means you are implementing **stochastic gradient descent**.
- This can help to minimize overfitting and keep from getting stuck in a local minimum or plateau of the loss function gradient.

Package implementation

The most popular implementations of GBM in R:

GBM
The original R implementation of GBMs

XGBOOST
A fast and efficient gradient boosting framework (C++ backend).

H2O
A powerful java-based interface that provides parallel distributed algorithms and efficient productionalization.

The R-package gbm

The gbm R package is an implementation of extensions to Freund and Schapire's AdaBoost algorithm and Friedman's gradient boosting machine. This is the original R implementation of GBM.

Basic implementation

- two primary training functions: `gbm::gbm` and `gbm::gbm.fit`.
- `gbm::gbm` uses the formula interface to specify your model
- `gbm::gbm.fit` requires the separated `x` and `y` matrices (more efficient with many variables).

- The default settings in `gbm` includes a learning rate (shrinkage) of 0.001.
- This is a very small learning rate and typically requires a large number of trees to find the minimum MSE.
- `gbm` uses a default number of trees of 100, which is rarely sufficient. - - The default depth of each tree (interaction.depth) is 1, which means we are ensembling a bunch of stumps. Lastly, I also include cv.folds to perform a 5 fold cross validation. The model took about 90 seconds to run and the results show that our MSE loss function is minimized with 10,000 trees.

Train GBM model

```
set.seed(123)
gbm.fit <- gbm(formula = Sale_Price ~ .,distribution = "gaussian
  data = ames_train,n.trees = 10000,interaction.depth = 1,
  shrinkage = 0.001,cv.folds = 5,
  n.cores = NULL, # will use all cores by default
  verbose = FALSE)
# print results
print(gbm.fit)
```

```
## gbm(formula = Sale_Price ~ ., distribution = "gaussian", data
##     n.trees = 10000, interaction.depth = 1, shrinkage = 0.001
##     cv.folds = 5, verbose = FALSE, n.cores = NULL)
## A gradient boosted model with gaussian loss function.
## 10000 iterations were performed.
## The best cross-validation iteration was 10000.
## There were 80 predictors of which 45 had non-zero influence.
```

The output object...

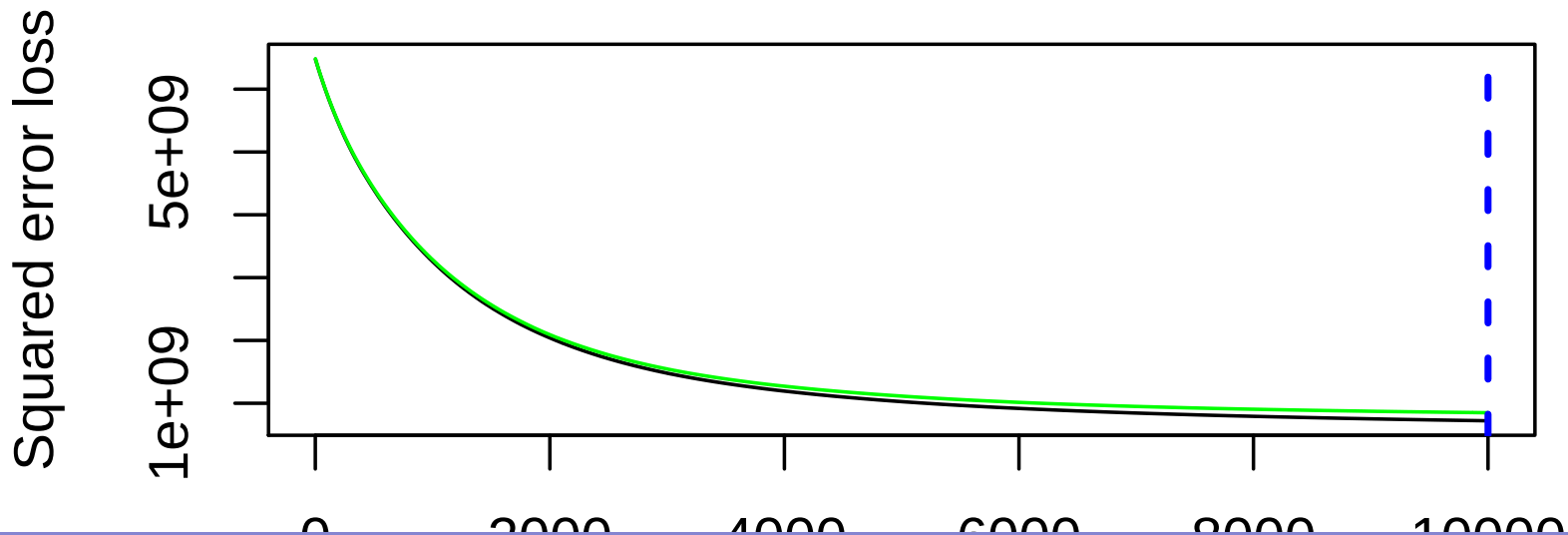
... is a list containing several modelling and results information. We can access this information with regular indexing; I recommend you take some time to dig around in the object to get comfortable with its components. Here, we see that the minimum CV RMSE is 29133 (this means on average our model is about $29,133 off from the actual sales price) but the plot also illustrates that the CV error is still decreasing at 10,000 trees.

Get MSE and compute RMSE

```
sqrt(min(gbm.fit$cv.error))
```

```
## [1] 29133.33
```

```
# plot loss function as a result of n trees added to the ensembl
gbm.perf(gbm.fit, method = "cv")
```



Tuning GBMs

- The learning rate is increased to take larger steps down the gradient descent,
- The number of trees is reduced (since we are reducing the learning rate), and increase the depth of each tree from using a single split to 3 splits.

```
set.seed(123)
gbm.fit2 <- gbm(formula = Sale_Price ~ .,
  distribution = "gaussian",data = ames_train,
  n.trees = 5000,interaction.depth = 3,shrinkage = 0.1,
  cv.folds = 5,n.cores = NULL,verbose = FALSE)
# find index for n trees with minimum CV error
min_MSE <- which.min(gbm.fit2$cv.error)

# get MSE and compute RMSE
sqrt(gbm.fit2$cv.error[min_MSE])

## [1] 23112.1
```

Grid search

- the minimum number of observations allowed in the trees terminal nodes (`n.minobsinnode`) is varied

```
hyper_grid <- expand.grid(
  shrinkage = c(.01, .1, .3),
  interaction.depth = c(1, 3, 5),
  n.minobsinnode = c(5, 10, 15),
  bag.fraction = c(.65, .8, 1),
  optimal_trees = 0,# a place to dump results
  min_RMSE = 0
)

# total number of combinations
nrow(hyper_grid)
```

```
## [1] 81
```

- We loop through each hyperparameter combination (5,000 trees).
- To speed up the tuning process, instead of performing 5-fold CV I train on 75% of the training observations and evaluate performance on the remaining 25%.
- When using `train.fraction` it will take the first XX% of the data so its important to randomize your rows in case their is any logic behind the ordering of the data (i.e. ordered by neighborhood).

- First, our top model has better performance than our previously fitted model above, with the RMSE nearly $3,000 lower. Second, looking at the top 10 models we see that:
- none of the top models used a learning rate of 0.3; small incremental steps down the gradient descent appears to work best,
- none of the top models used stumps (`interaction.depth = 1`); there are likely stome important interactions that the deeper trees are able to capture,
- adding a stochastic component with bag.fraction < 1 seems to help; there may be some local minimas in our loss function gradient,

Randomize data and loop over hyperparameter grid

```
# randomize data
random_index <- sample(1:nrow(ames_train), nrow(ames_train))
random_ames_train <- ames_train[random_index, ]
# grid search
for(i in 1:nrow(hyper_grid)) {
  set.seed(123)
  # train model
  gbm.tune <- gbm(
    formula = Sale_Price ~ .,
    distribution = "gaussian",
    data = random_ames_train,
    n.trees = 5000,
    interaction.depth = hyper_grid$interaction.depth[i],
    shrinkage = hyper_grid$shrinkage[i],
    n.minobsinnode = hyper_grid$n.minobsinnode[i],
    bag.fraction = hyper_grid$bag.fraction[i],
```

Refine the search - adjust the grid

```
# modify hyperparameter grid
hyper_grid <- expand.grid(
  shrinkage = c(.01, .05, .1),
  interaction.depth = c(3, 5, 7),
  n.minobsinnode = c(5, 7, 10),
  bag.fraction = c(.65, .8, 1),
  optimal_trees = 0,                    # a place to dump results
  min_RMSE = 0                          # a place to dump results
)

# total number of combinations
nrow(hyper_grid)


## [1] 81
```

The final model

```
set.seed(123)

# train GBM model
gbm.fit.final <- gbm(
  formula = Sale_Price ~ .,
  distribution = "gaussian",
  data = ames_train,
  n.trees = 483,
  interaction.depth = 5,
  shrinkage = 0.1,
  n.minobsinnode = 5,
  bag.fraction = .65,
  train.fraction = 1,
  n.cores = NULL, # will use all cores by default
  verbose = FALSE
  )
```

Visualizing - Variable importance

- cBars allows you to adjust the number of variables to show

```
par(mar = c(5, 8, 1, 1))
summary(gbm.fit.final,cBars = 10,
  # also can use permutation.test.gbm
  method = relative.influence,las = 2)
```