Recap – Random Forest, AdaBoost, GradientBoost

Rina BUOY

AMERICAN UNIVERSITY
OF PHNOM PENH
STUDY LOCALLY. LIVE GLOBALLY.

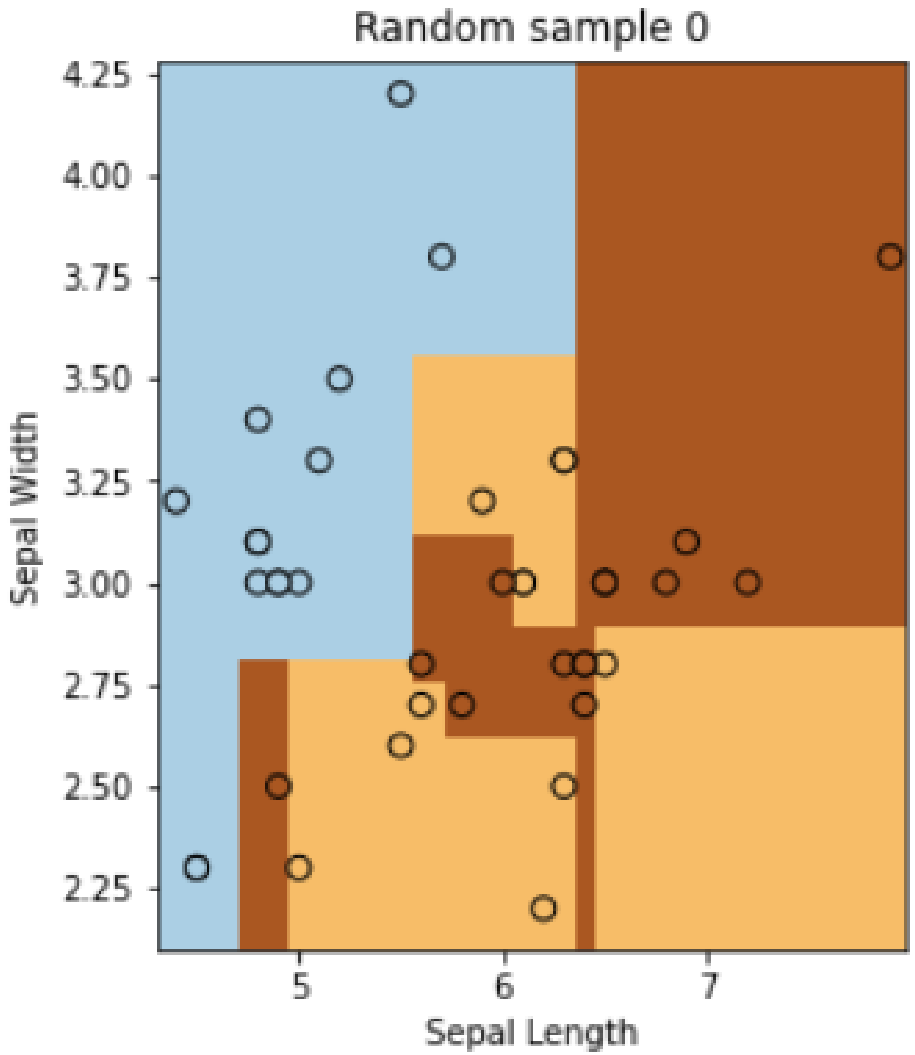
Source: Cornell CS5785

High-Variance Decision Trees

- When the trees have sufficiently high depth, they can quickly overfit the data.

- Tis called the *high variance* problem, because small perturbations of the data lead to large changes in model predictions.

High-Variance Decision Trees

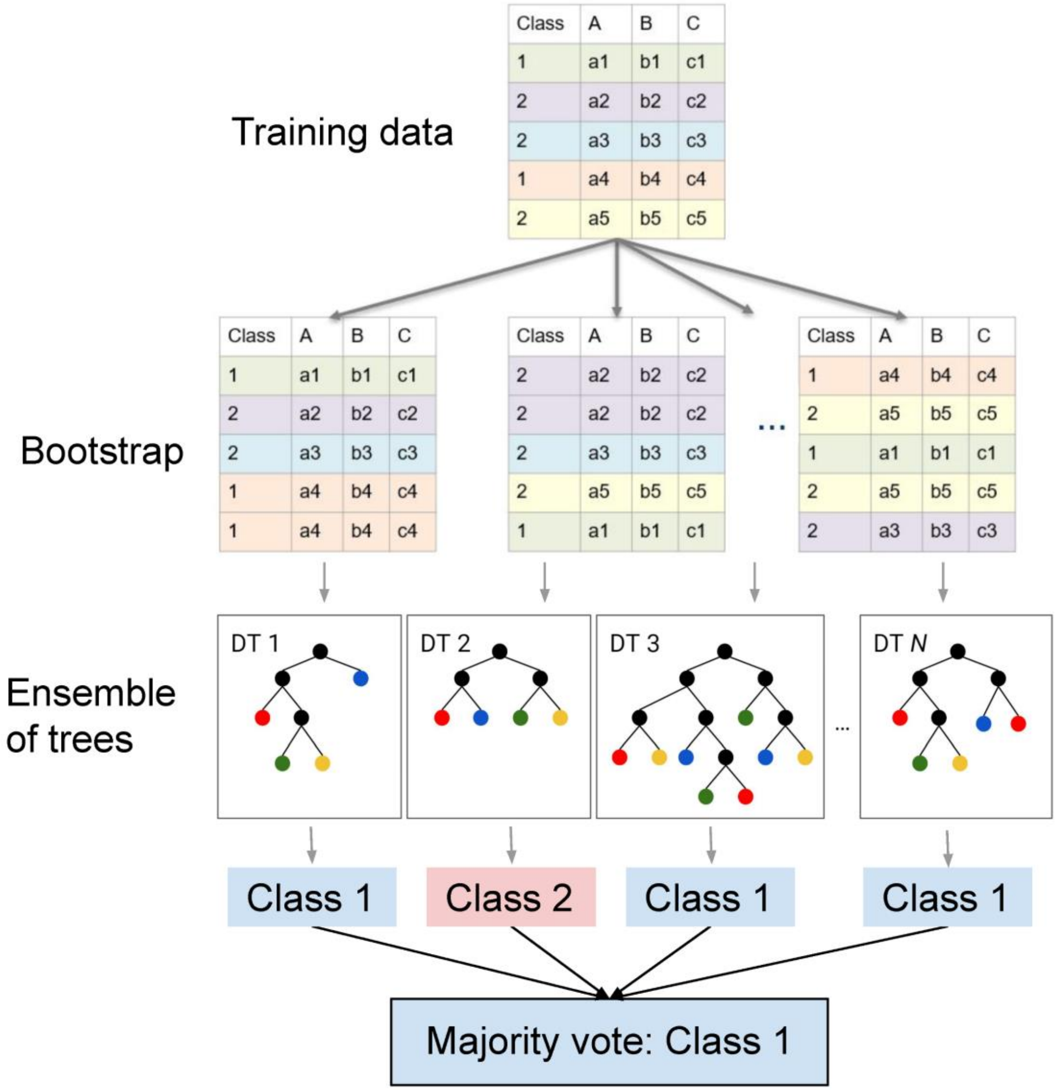


Random Forests

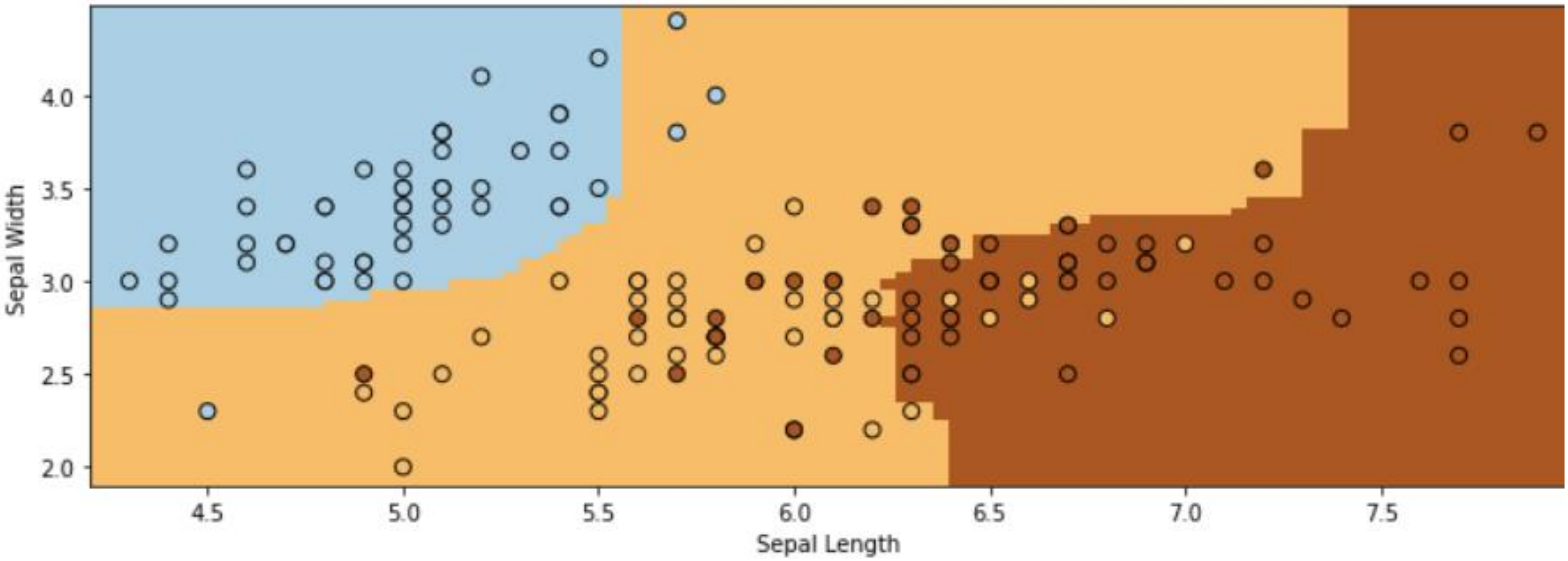
- In order to reduce the variance of the basic decision tree, we apply bagging -- the variance reduction technique that we have seen earlier.

- We refer to bagged decision trees as **Random Forests**.

Random Forests



Random Forests



Pros and Cons of Random Forests

- Random forests remain a popular machine learning algorithm:

  * They require little data preparation (no rescaling, handle continuous and discrete features, work well for classification and regression).

  * They are often quite accurate.

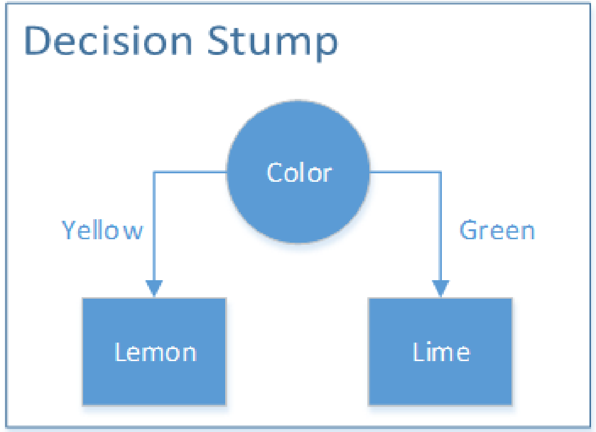
- Their main disadvantages are that:

  * They are not interpretable.

  * They do not work with unstructured data (images, audio).

Underfitting

- Underfitting is another common problem in machine learning.
- The model is too simple to fit the data well.
- Training performance is low, hence test performance is low.
- Intuitively, a weak model is a model that is slightly better than random.
- Examples of weak learners include: small linear models, small decision trees.



Boosting

- The idea of **boosting** is to reduce *underfitting* by combining models that correct each others' errors.

- As in bagging, we combine many weak models into one strong model.

- Each weak model fits the points where the previous models made errors.

Adaboost

One of the first practical boosting algorithms was *Adaboost*.

We start with uniform $w^{(i)} = 1/n$ and $f = 0$. Then for $t = 1, 2, ..., T$:

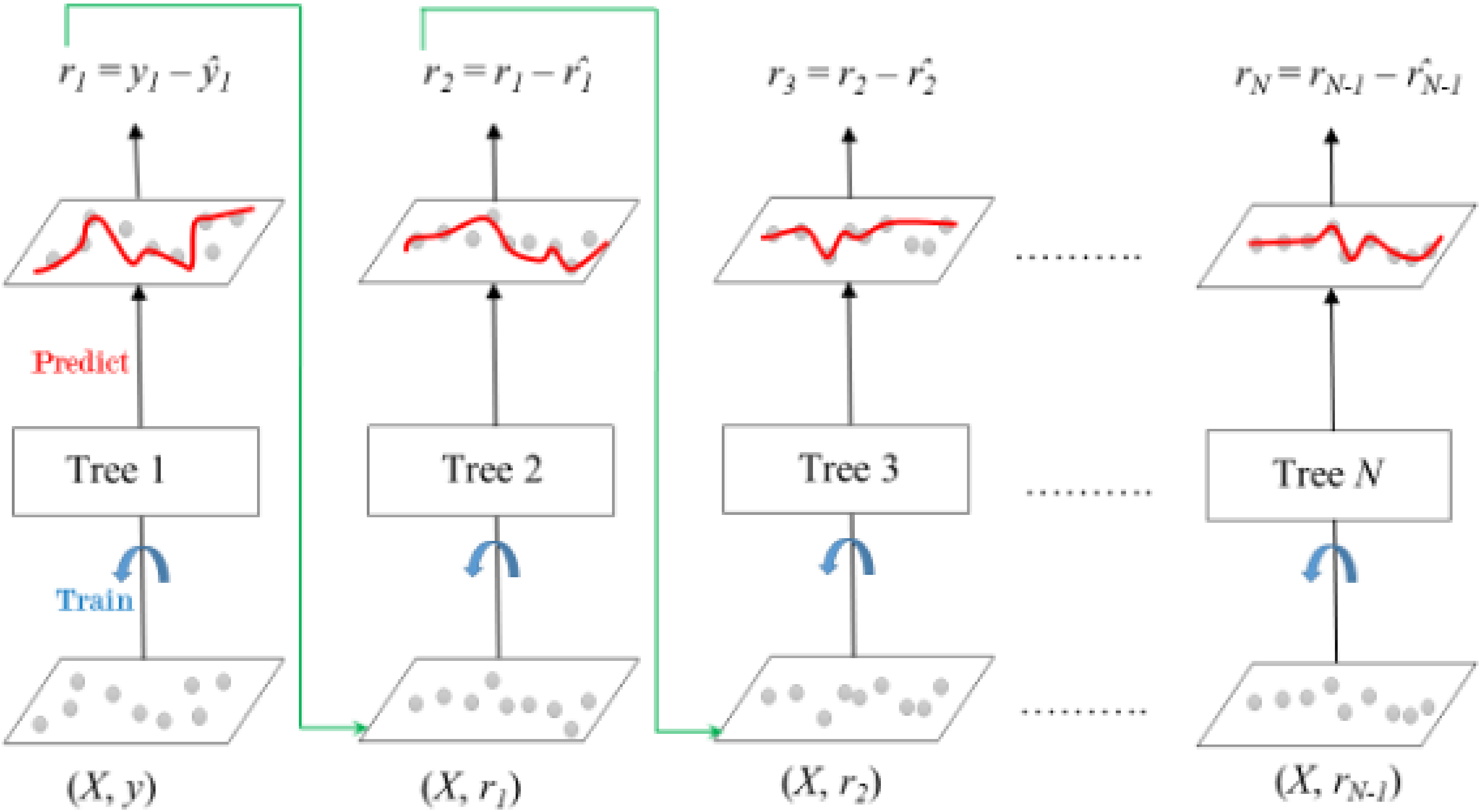
1. Fit weak learner g_t on $\mathcal{D}$ with weights $w^{(i)}$.
2. Compute misclassification error $e_t = \frac{\sum_{i=1}^{n} w^{(i)} \mathbb{I}\{y^{(i)} \neq f(x^{(i)})\}}{\sum_{i=1}^{n} w^{(i)}}$
3. Compute model weight $\alpha_t = \log[(1 - e_t)/e_t]$. Set $f \leftarrow f + \alpha_t g_t$.
4. Compute new data weights $w^{(i)} \leftarrow w^{(i)} \exp[\alpha_t \mathbb{I}\{y^{(i)} \neq f(x^{(i)})\}]$.

Adaboost

- Boosting algorithms generalize Adaboost and offer many advantages:
  - High accuracy via a highly expressive non-linear model family.
  - Low pre-processing requirements if trees are used as weak learners.

- Disadvantages include:
  - Large ensembles can be expensive to train.
  - The interpretability of the weak learners is lost.

Gradient boosting



https://www.geeksforgeeks.org/ml-gradient-boosting/

Gradient boosting

- Gradient boosted decision trees (GBTs) are one of the best off-the-shelf ML algorithms that exist, often on par with deep learning.
  - Attain state-of-the-art performance.
  - GBTs rule on Kaggle.
  - Require little data pre-processing and tuning.
- Their main limitations are:
  - GBTs don't work with unstructured data like images, audio.
  - Implementations not as flexible as modern neural net libraries.