Ensemble methods

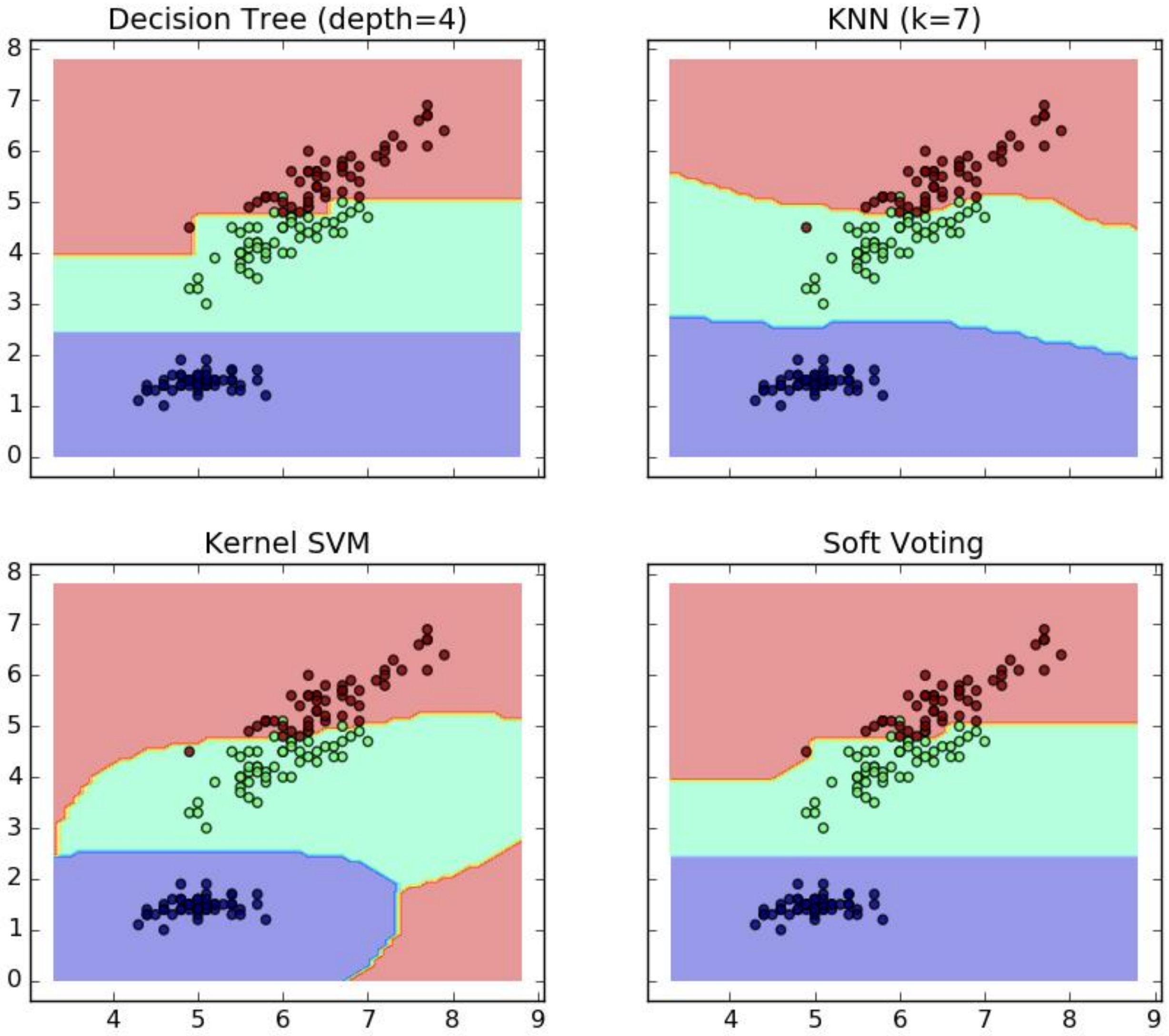
Voting

Hard voting:

Count the number of votes for each class.

Soft voting:

Sum or multiply probabilities.



Random Forests

Train many small trees on random subsamples.

Sampling types:

Pasting – simple sampling with no repeats.

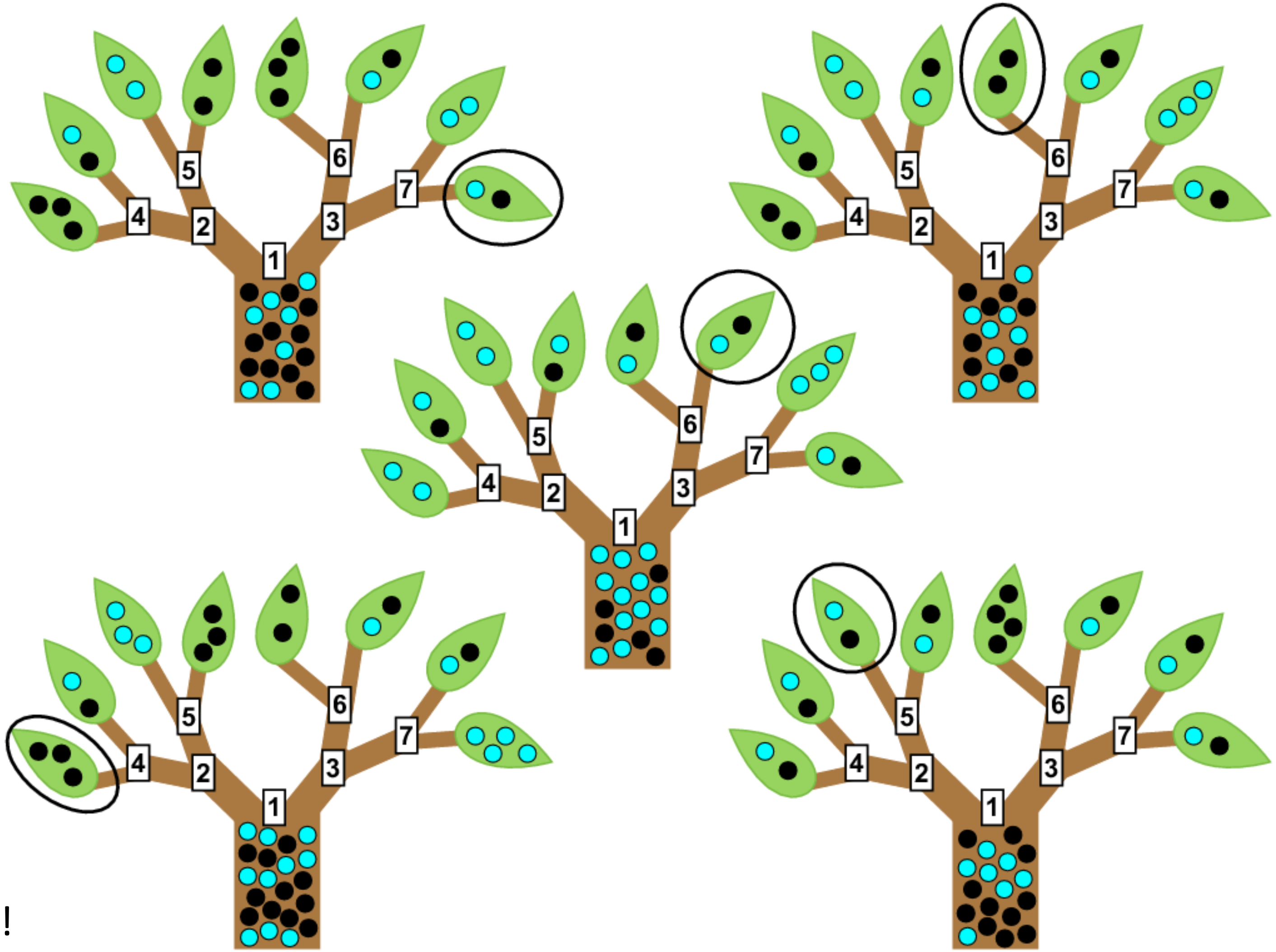
Bagging (bootstrap aggregating) – sample with repeats of the same size as the original dataset.

Random Subspaces – sample of features.

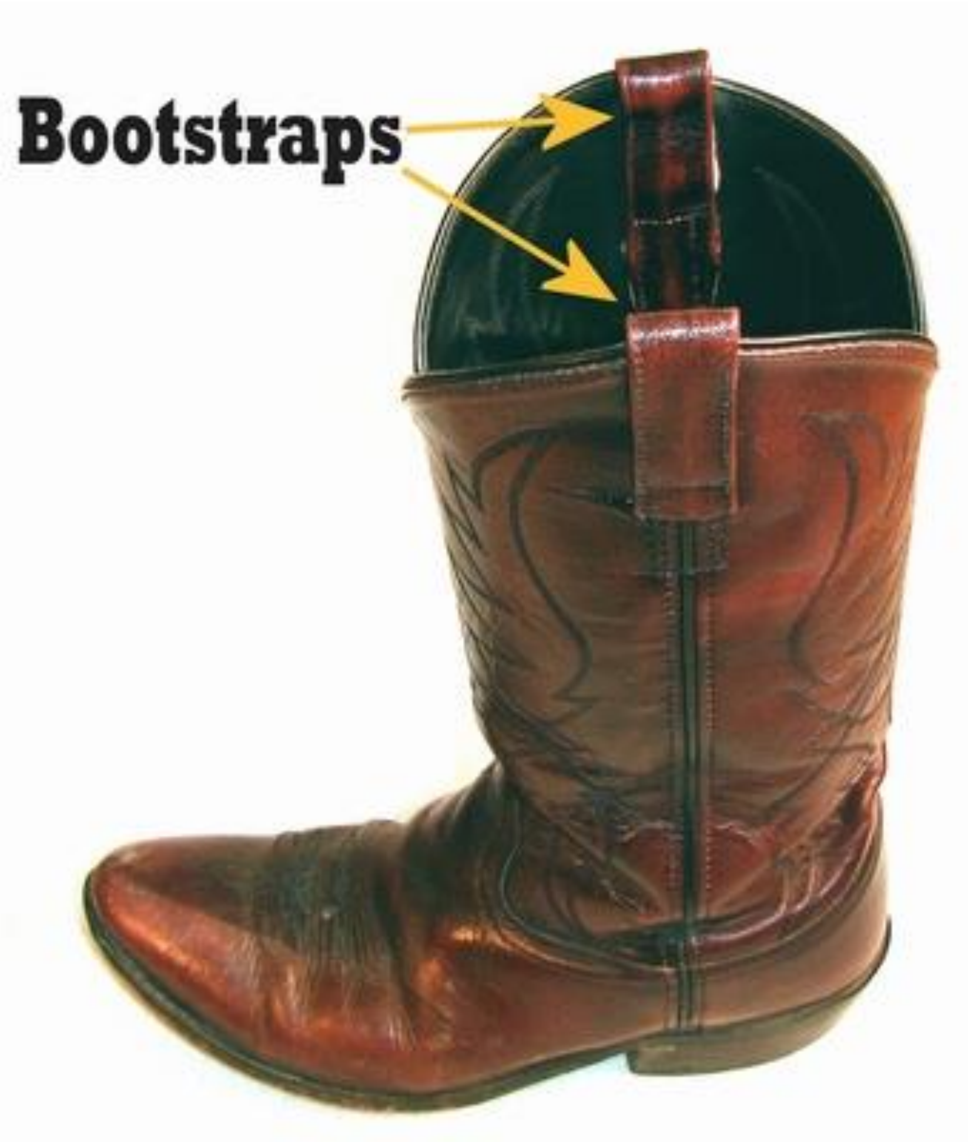
Random Patches – sample both features and examples.

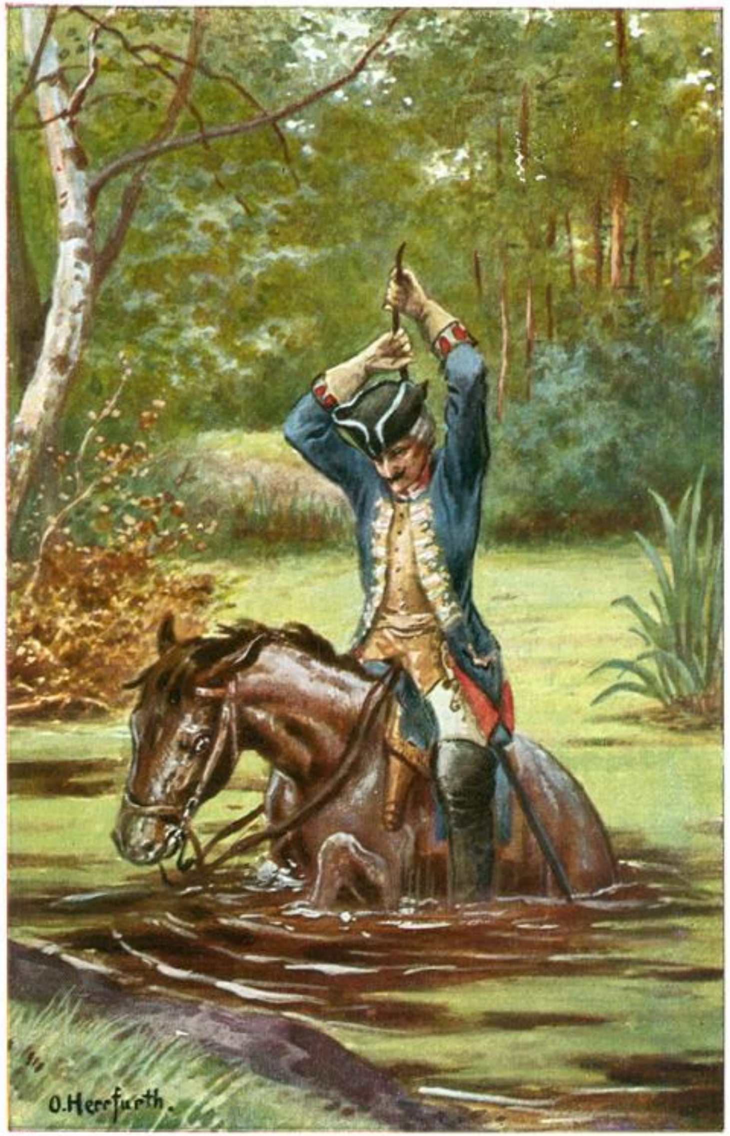
Final result could be determined by either hard or soft voting.

You can even use extremely randomized trees!



Pulling yourself up by your bootstraps





Münchhausen O. Herrfurth pinx

Adaptive Boosting (AdaBoost)

Adaptive Boosting (AdaBoost)

Start with uniform sample weights: $D_1(i) = \frac{1}{N}, for\ (\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_N, y_N)$.

Adaptive Boosting (AdaBoost)

Start with uniform sample weights: $D_1(i) = \frac{1}{N}, for\ (\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_N, y_N)$.

Train a weak hypothesis (weak tree hypothesis are called stumps) h_t by either using a weighted impurity or weighted sampling.

$$E_t = \sum_{i=1}^{N} D_t(i)\ E(h_t(x_i), y_i), \qquad the\ weight\ of\ hypothesis\ h_t\!: \alpha_t = \frac{1}{2}\ln\left(\frac{1 - E_t}{E_t}\right)$$

Adaptive Boosting (AdaBoost)

Start with uniform sample weights: $D_1(i) = \frac{1}{N}, for\ (\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_N, y_N)$.

Train a weak hypothesis (weak tree hypothesis are called stumps) h_t by either using a weighted impurity or weighted sampling.

$$E_t = \sum_{i=1}^{N} D_t(i)\ E(h_t(x_i), y_i), \qquad the\ weight\ of\ hypothesis\ h_t: \alpha_t = \frac{1}{2} \ln\left(\frac{1 - E_t}{E_t}\right)$$

Change the sample weights:

For incorrectly classified points: $D_{t+1}(i) = D_t(i)e^{\alpha_t}$

For correctly classified points: $D_{t+1}(i) = D_t(i)e^{-\alpha_t}$

then normalize

Adaptive Boosting (AdaBoost)

Start with uniform sample weights: $D_1(i) = \frac{1}{N}, for\ (\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_N, y_N)$.

Train a weak hypothesis (weak tree hypothesis are called stumps) h_t by either using a weighted impurity or weighted sampling.

$$E_t = \sum_{i=1}^{N} D_t(i)\ \ E(h_t(x_i), y_i), \qquad the\ weight\ of\ hypothesis\ h_t\colon \alpha_t = \frac{1}{2}\ln\left(\frac{1 - E_t}{E_t}\right)$$

Change the sample weights:

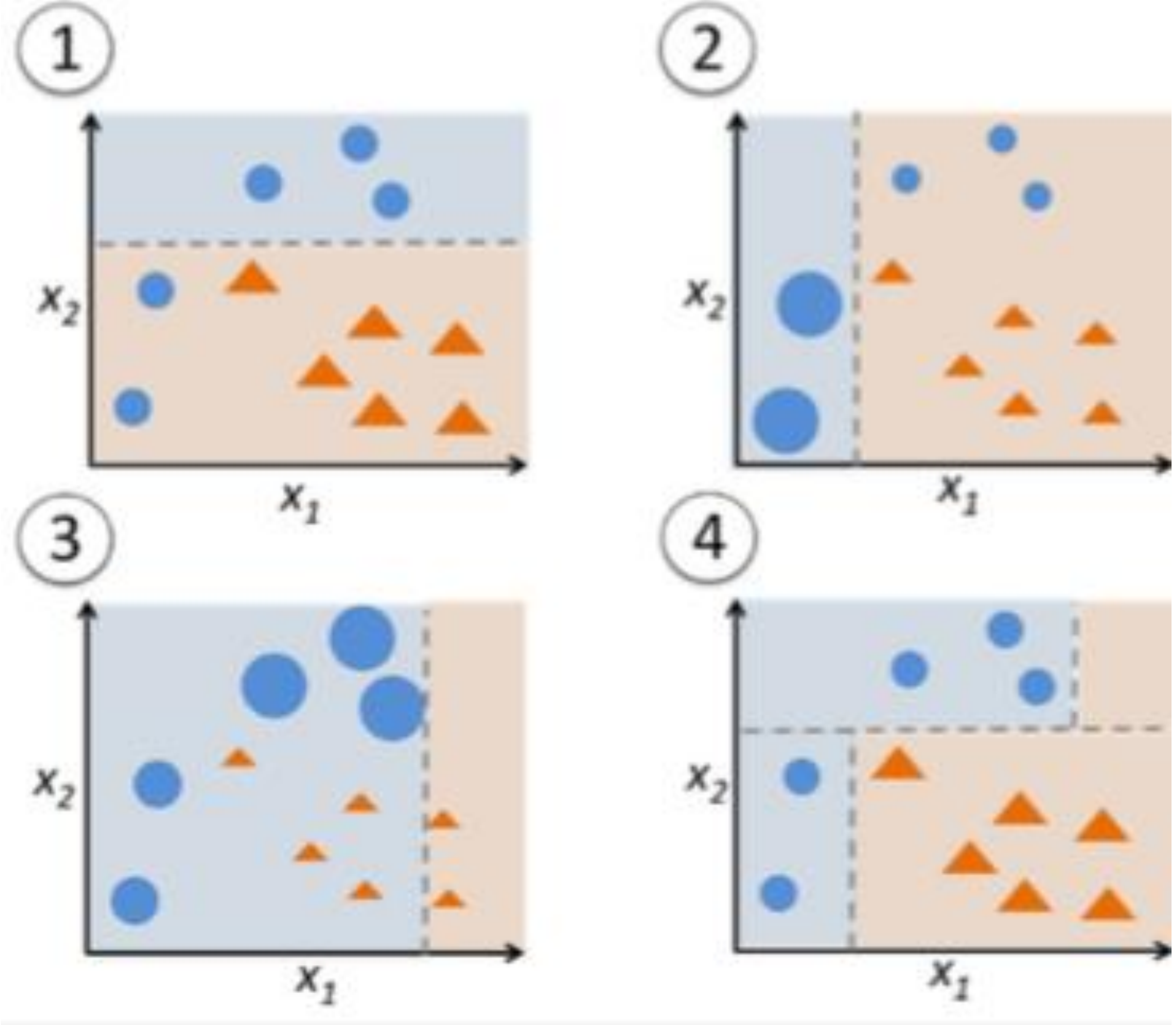
For incorrectly classified points: $D_{t+1}(i) = D_t(i)e^{\alpha_t}$

then normalize

For correctly classified points: $D_{t+1}(i) = D_t(i)e^{-\alpha_t}$

The final hypothesis is summed by hard or soft voting with coefficients α_t .

AdaBoost



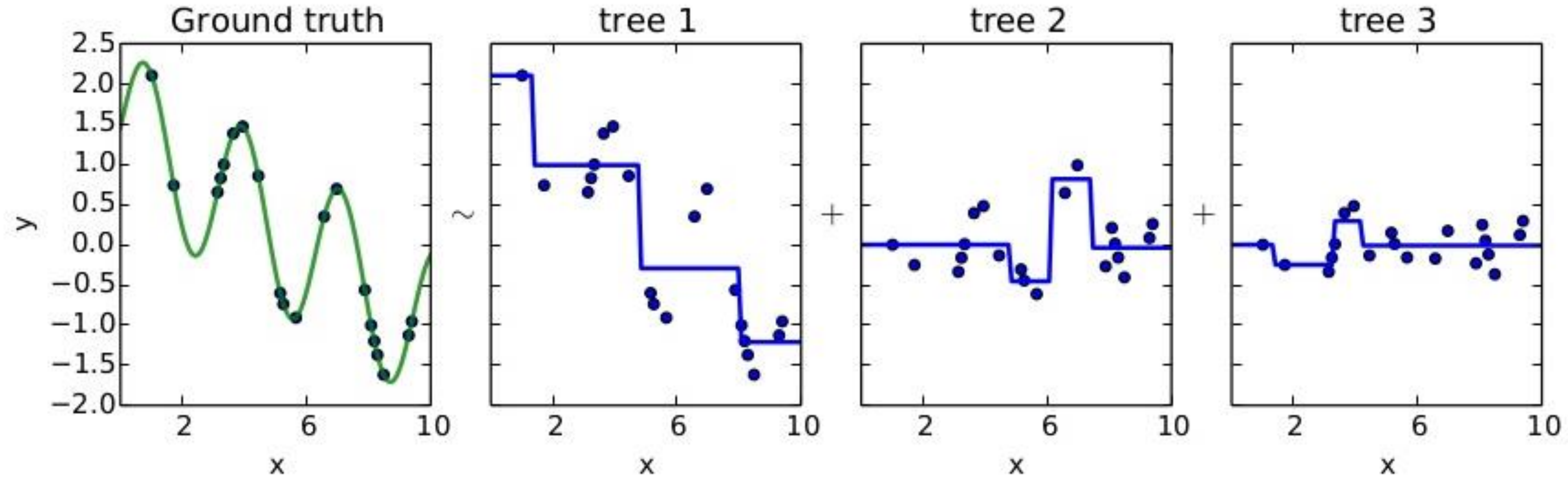
Gradient Boosting

$$H_{t+1}(\mathrm{x}) = H_t(\mathrm{x}) + h_{t+1}(\mathrm{x}) \to y \;\Rightarrow\; h_{t+1}(\mathrm{x}) \to y - H_t(\mathrm{x})$$

Gradient Boosting

$$H_{t+1}(\mathrm{x}) = H_t(\mathrm{x}) + h_{t+1}(\mathrm{x}) \to y \ \Rightarrow \ h_{t+1}(\mathrm{x}) \to y - H_t(\mathrm{x})$$

First – simple!



Gradient Boosting Learning Rate

$$h_{t+1}(\mathrm{x}) \rightarrow \overbrace{y - H_t(\mathrm{x})}^{\text{Pseudo Residual}}$$

$$H_{t+1}(\mathrm{x}) = H_t(\mathrm{x}) + \underset{\substack{\uparrow \\ \text{Learning Rate}}}{\alpha} h_{t+1}(\mathrm{x})$$

Now – complicated!

$$H_t(\mathrm{x}) = H_{t-1}(\mathrm{x}) + h_t(\mathrm{x}) = \sum_{j=1}^{t} h_j(\mathrm{x})$$

eXtreme Gradient Boosting (XGBoost)

$$H_t(\mathrm{x}) = H_{t-1}(\mathrm{x}) + h_t(\mathrm{x}) = \sum_{j=1}^{t} h_j(\mathrm{x})$$

We want to minimize: $E_t = \sum_{i=1}^{N} L(H_t(\mathrm{x}_i), y_i) + \sum_{j=1}^{t} \Omega(h_j)$

Regularization

eXtreme Gradient Boosting (XGBoost)

$$H_t(\mathrm{x}) = H_{t-1}(\mathrm{x}) + h_t(\mathrm{x}) = \sum_{j=1}^{t} h_j(\mathrm{x})$$

We want to minimize: $E_t = \sum_{i=1}^{N} L(H_t(\mathrm{x}_i), y_i) + \sum_{j=1}^{t} \Omega(h_j) = \sum_{i=1}^{N} L\left((H_{t-1}(\mathrm{x}_i) + h_t(\mathrm{x}_i)), y_i\right) + \sum_{j=1}^{t-1} \Omega(h_j) + \Omega(h_t)$

Regularization

XGBoost

$$E_t = \sum_{i=1}^{N} L\left(\left(H_{t-1}(\mathrm{x}_i) + \ h_t(\mathrm{x}_i)\right), y_i\right) + \sum_{j=1}^{t-1} \Omega(h_j) + \Omega(h_t)$$

XGBoost

$$E_t = \sum_{i=1}^{N} L\left(\left(H_{t-1}(\mathrm{x}_i) + \; h_t(\mathrm{x}_i)\right), y_i\right) + \sum_{j=1}^{t-1} \Omega(h_j) + \Omega(h_t)$$

In the general case: $E_t = \sum_{i=1}^{N} \left(L(H_{t-1}(\mathrm{x}_i), y_i) + u_i h_t(\mathrm{x}_i) + \frac{1}{2} v_i \left(h_t(\mathrm{x}_i)\right)^2 \right) + \Omega(h_t) + const,$

XGBoost

$$E_t = \sum_{i=1}^{N} L\left(\left(H_{t-1}(\mathrm{x}_i) + \ h_t(\mathrm{x}_i)\right), y_i\right) + \sum_{j=1}^{t-1} \Omega(h_j) + \Omega(h_t)$$

In the general case: $E_t = \sum_{i=1}^{N} \left(L(H_{t-1}(\mathrm{x}_i), y_i) + u_i h_t(\mathrm{x}_i) + \frac{1}{2} v_i \left(h_t(\mathrm{x}_i)\right)^2 \right) + \Omega(h_t) + const,$

$$u_i = \partial_{H_{t-1}(\mathrm{x}_i)}\left(L(H_{t-1}(\mathrm{x}_i), y_i)\right)$$

$$v_i = \partial^2_{H_{t-1}(\mathrm{x}_i)}\left(L(H_{t-1}(\mathrm{x}_i), y_i)\right)$$

XGBoost

$$E_t = \sum_{i=1}^{N} L\left(\left(H_{t-1}(\mathrm{x}_i) + \ h_t(\mathrm{x}_i)\right), y_i\right) + \sum_{j=1}^{t-1} \Omega(h_j) + \Omega(h_t)$$

In the general case: $E_t = \sum_{i=1}^{N} \left(L(H_{t-1}(\mathrm{x}_i), y_i) + u_i h_t(\mathrm{x}_i) + \frac{1}{2} v_i \left(h_t(\mathrm{x}_i)\right)^2 \right) + \Omega(h_t) + const,$

$$u_i = \partial_{H_{t-1}(\mathrm{x}_i)} \left(L(H_{t-1}(\mathrm{x}_i), y_i) \right)$$

$$v_i = \partial^2_{H_{t-1}(\mathrm{x}_i)} \left(L(H_{t-1}(\mathrm{x}_i), y_i) \right)$$

We want to minimize: $E_t = \sum_{i=1}^{N} \left(u_i h_t(\mathrm{x}_i) + \frac{1}{2} v_i \left(h_t(\mathrm{x}_i)\right)^2 \right) + \Omega(h_t)$

XGBoost

$$E_t = \sum_{i=1}^{N} L\left(\left(H_{t-1}(\mathrm{x}_i) + \ h_t(\mathrm{x}_i)\right), y_i\right) + \sum_{j=1}^{t-1} \Omega(h_j) + \Omega(h_t)$$

In the general case: $E_t = \sum_{i=1}^{N} \left(L(H_{t-1}(\mathrm{x}_i), y_i) + u_i h_t(\mathrm{x}_i) + \frac{1}{2} v_i \left(h_t(\mathrm{x}_i)\right)^2 \right) + \Omega(h_t) + const,$

$$u_i = \partial_{H_{t-1}(\mathrm{x}_i)}\left(L(H_{t-1}(\mathrm{x}_i), y_i)\right)$$

$$v_i = \partial^2_{H_{t-1}(\mathrm{x}_i)}\left(L(H_{t-1}(\mathrm{x}_i), y_i)\right)$$

We want to minimize: $E_t = \sum_{i=1}^{N} \left(u_i h_t(\mathrm{x}_i) + \frac{1}{2} v_i \left(h_t(\mathrm{x}_i)\right)^2 \right) + \Omega(h_t)$

For MSE: $E_t = \sum_{i=1}^{N} \left(\left(H_{t-1}(\mathrm{x}_i) + \ h_t(\mathrm{x}_i)\right) - y_i\right)^2 + \sum_{j=1}^{t} \Omega(h_j) = \sum_{i=1}^{N} \left(2(H_{t-1}(\mathrm{x}_i) - y_i) h_t(\mathrm{x}_i) + \left(h_t(\mathrm{x}_i)\right)^2 \right) + \Omega(h_t) + const$

XGBoost

$$\Omega(f) = \gamma M + \frac{1}{2}\lambda \sum_{j=1}^{M} w_j^2 \,, M \ - number\ of\ leaves, w_j \ - otput\ number\ in\ the\ leaf\ j$$

XGBoost

$$\Omega(f) = \gamma M + \frac{1}{2}\lambda \sum_{j=1}^{M} w_j^2\ , M\ - number\ of\ leaves, w_j\ - otput\ number\ in\ the\ leaf\ j$$

XGBoost

$$\Omega(f) = \gamma M + \frac{1}{2}\lambda\sum_{j=1}^{M} w_j^2\,, M \ - number\ of\ leaves, w_j \ - otput\ number\ in\ the\ leaf\ j$$

$$E_t = \sum_{i=1}^{N}\left(u_i h_t(\mathrm{x}_i) + \frac{1}{2}v_i\big(h_t(\mathrm{x}_i)\big)^2\right) + \Omega(h_t) = \sum_{i=1}^{N}\left(u_i w_{q(\mathrm{x}_i)} + \frac{1}{2}v_i\big(w_{q(\mathrm{x}_i)}\big)^2\right) + \gamma M + \frac{1}{2}\lambda\sum_{j=1}^{M} w_j^2$$

$q(\mathbf{x}_i)$ is the leaf of $\mathbf{x}_i$.

XGBoost

$$\Omega(f) = \gamma M + \frac{1}{2}\lambda \sum_{j=1}^{M} w_j^2\,, M\; - number\ of\ leaves, w_j\; - otput\ number\ in\ the\ leaf\ j$$

$$E_t = \sum_{i=1}^{N}\left(u_i h_t(\mathrm{x}_i) + \frac{1}{2} v_i \big(h_t(\mathrm{x}_i)\big)^2\right) + \Omega(h_t) = \sum_{i=1}^{N}\left(u_i w_{q(\mathrm{x}_i)} + \frac{1}{2} v_i \big(w_{q(\mathrm{x}_i)}\big)^2\right) + \gamma M + \frac{1}{2}\lambda \sum_{j=1}^{M} w_j^2$$

$q(\mathbf{x}_i)$ is the leaf of $\mathbf{x}_i$.

Group by leaves:

$$E_t = \sum_{j=1}^{M}\left(\sum_{q(\mathrm{x}_i)=j} u_i\, w_j + \frac{1}{2}\left(\sum_{q(\mathrm{x}_i)=j} v_i + \lambda\right) w_j^2\right) + \gamma M$$

XGBoost

$$\Omega(f) = \gamma M + \frac{1}{2}\lambda\sum_{j=1}^{M} w_j^2\ , M\ - number\ of\ leaves, w_j\ - otput\ number\ in\ the\ leaf\ j$$

$$E_t = \sum_{i=1}^{N}\left(u_i h_t(\mathrm{x}_i) + \frac{1}{2}v_i\big(h_t(\mathrm{x}_i)\big)^2\right) + \Omega(h_t) = \sum_{i=1}^{N}\left(u_i w_{q(\mathrm{x}_i)} + \frac{1}{2}v_i\big(w_{q(\mathrm{x}_i)}\big)^2\right) + \gamma M + \frac{1}{2}\lambda\sum_{j=1}^{M} w_j^2$$

$q(\mathbf{x}_i)$ is the leaf of $\mathbf{x}_i$.

Group by leaves:

$$E_t = \sum_{j=1}^{M}\left(\sum_{q(\mathrm{x}_i)=j} u_i\, w_j + \frac{1}{2}\left(\sum_{q(\mathrm{x}_i)=j} v_i + \lambda\right) w_j^2\right) + \gamma M$$

$$U_j = \sum_{q(\mathrm{x}_i)=j} u_i \qquad V_j = \sum_{q(\mathrm{x}_i)=j} v_i$$

XGBoost

$$\Omega(f) = \gamma M + \frac{1}{2}\lambda \sum_{j=1}^{M} w_j^2\,, M \; - number\ of\ leaves, w_j \; - otput\ number\ in\ the\ leaf\ j$$

$$E_t = \sum_{i=1}^{N}\left(u_i h_t(\mathrm{x}_i) + \frac{1}{2} v_i \big(h_t(\mathrm{x}_i)\big)^2\right) + \Omega(h_t) = \sum_{i=1}^{N}\left(u_i w_{q(\mathrm{x}_i)} + \frac{1}{2} v_i \big(w_{q(\mathrm{x}_i)}\big)^2\right) + \gamma M + \frac{1}{2}\lambda \sum_{j=1}^{M} w_j^2$$

$q(\mathbf{x}_i)$ is the leaf of $\mathbf{x}_i$.

Group by leaves:

$$E_t = \sum_{j=1}^{M}\left(\sum_{q(\mathrm{x}_i)=j} u_i\, w_j + \frac{1}{2}\left(\sum_{q(\mathrm{x}_i)=j} v_i + \lambda\right) w_j^2\right) + \gamma M = \sum_{j=1}^{M}\left(U_j w_j + \frac{1}{2}(V_j + \lambda) w_j^2\right) + \gamma M$$

$$U_j = \sum_{q(\mathrm{x}_i)=j} u_i \qquad V_j = \sum_{q(\mathrm{x}_i)=j} v_i$$

XGBoost

$$E_t = \sum_{j=1}^{M} \left(U_j w_j + \frac{1}{2} \left(V_j + \lambda \right) w_j^2 \right) + \gamma M$$

XGBoost

$$E_t = \sum_{j=1}^{M} \left(U_j w_j + \frac{1}{2} (V_j + \lambda) w_j^2 \right) + \gamma M$$

$$w_j^{opt} = -\frac{U_j}{V_j + \lambda} \qquad E_t^{opt} = -\frac{1}{2} \sum_{j=1}^{M} \frac{U_j^2}{V_j + \lambda} + \gamma M$$

XGBoost

$$E_t = \sum_{j=1}^{M} \left(U_j w_j + \frac{1}{2} (V_j + \lambda) w_j^2 \right) + \gamma M$$

$$w_j^{opt} = -\frac{U_j}{V_j + \lambda}$$

$$E_t^{opt} = -\frac{1}{2} \sum_{j=1}^{M} \frac{U_j^2}{V_j + \lambda} + \gamma M$$

For MSE, the average of residuals!

MSE!

$$Gain = \frac{1}{2} \left[\frac{U_L^2}{V_L + \lambda} + \frac{U_R^2}{V_R + \lambda} - \frac{(U_L + U_R)^2}{V_L + V_R + \lambda} \right] - \gamma$$

Gradient Boosting Learning Rate

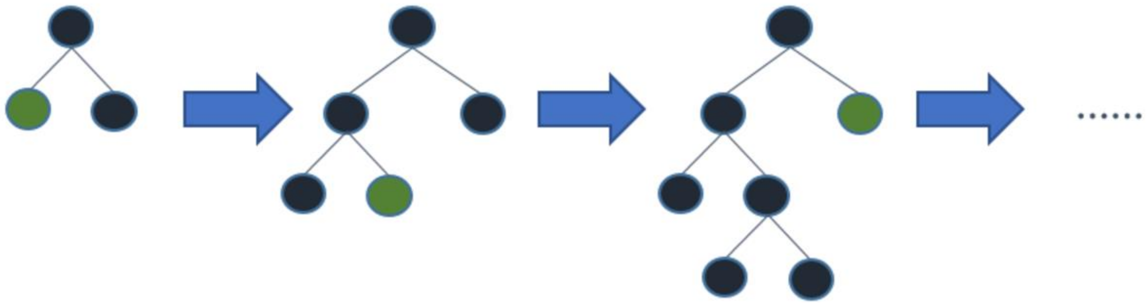
$$h_{t+1}(\mathrm{x}) \rightarrow \overbrace{y - H_t(\mathrm{x})}^{\text{Pseudo Residual}}$$

$$H_{t+1}(\mathrm{x}) = H_t(\mathrm{x}) + \alpha h_{t+1}(\mathrm{x})$$

GB libraries

dmlc XGBoost

LightGBM



CatBoost