Combining Models

Jiachun Jin
jinjch@shanghaitech.edu.cn

Outline

- Committees
- Boosting

Committees

- [Definition] A combinations of models. Practically we train M different models, make predictions using the average of the predictions made by each model.

  Predictive model: $y_m(\mathbf{x}), \forall m \in [M]$

  Committee predictive model: $y_{\mathrm{COM}}(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} y_m(\mathbf{x})$

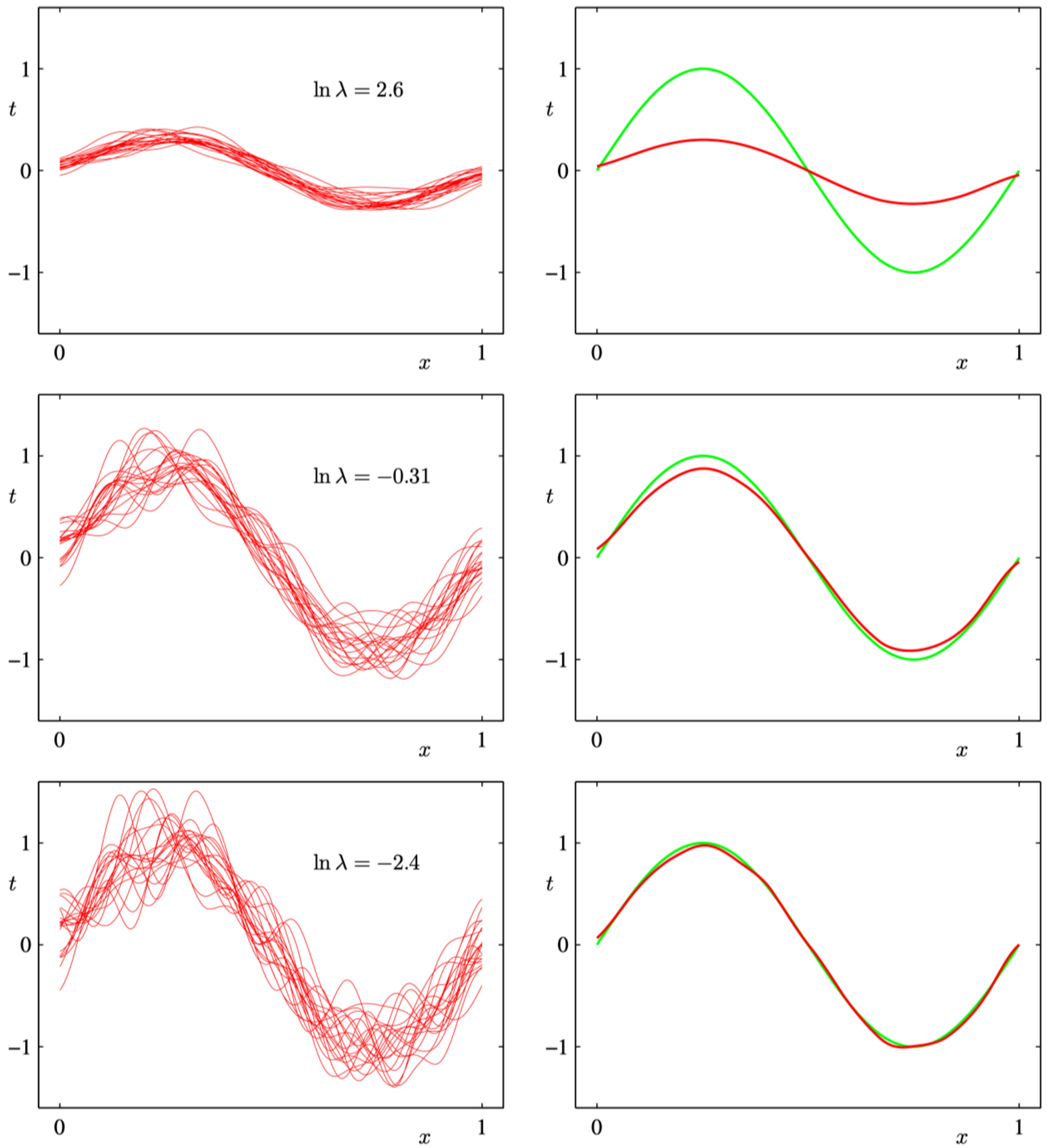
- An example

Committees

- Use polynomial curve to fit sinusoidal function

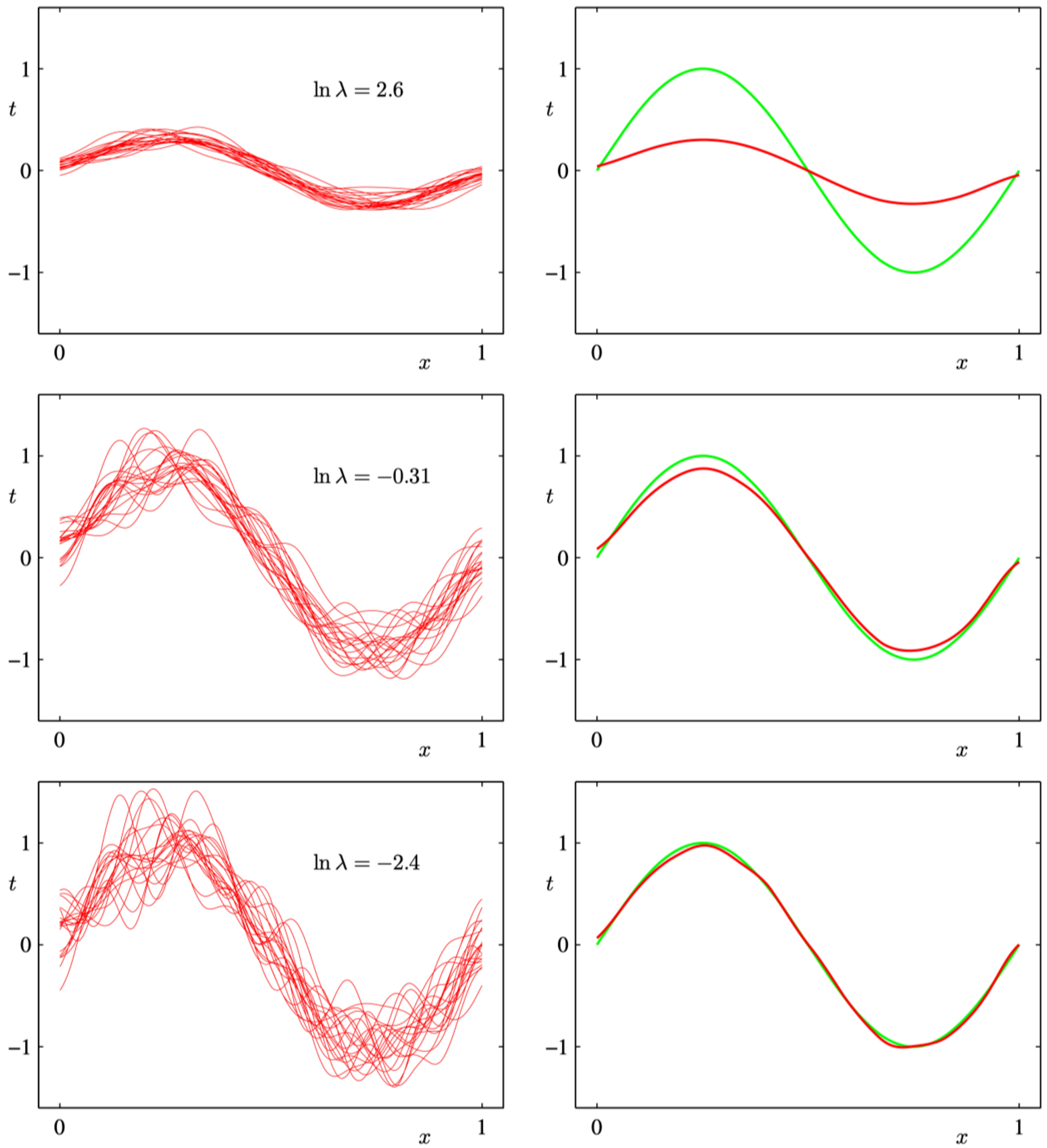
$$\frac{1}{2}\sum_{n=1}^{N}\left\{y_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\left(\mathbf{x}_n\right)\right\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

- $L = 100$ data sets, each having $N = 25$ data points, and $M = 25$ including the bias term
- Green line: the sinusoidal function from which the data sets were generated
- Red line: average of the 100 fits
- Observation: the contribution arising from the variance term tended to cancel, leading to improved predictions
- Idea: averaged a set of low-bias models (corresponding to higher order polynomials)



Committees

- Problem: we have only a single data set
- Solution: use bootstrap datasets
  - Suppose our original data set consists of N data points $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
  - We can create a new data set $\mathbf{X}_B$ by drawing N points at random from $\mathbf{X}$, with replacement, some points in $\mathbf{X}$ may be replicated in $\mathbf{X}_B$
  - repeat M times



Committees

- Predictive model: $y_m(\mathbf{x}), \forall m \in [M]$
- Committee prediction: $y_{\mathrm{COM}}(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} y_m(\mathbf{x})$
- True regression function: $h(\mathbf{x})$
- Output of each model: $y_m(\mathbf{x}) = h(\mathbf{x}) + \epsilon_m(\mathbf{x})$
- Average error made by the models acting individually:

$$E_{\mathrm{AV}} = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{\mathbf{x}}\left[\epsilon_m(\mathbf{x})^2\right]$$

- EPE: trade-off between bias and variance:

$$\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2\right] = \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]}_{\text{variance}}$$

- The expected error from the committee y_{COM}:

$$E_{\mathrm{COM}} = \mathbb{E}_{\mathbf{x}}\left[\left\{\frac{1}{M}\sum_{m=1}^{M} y_m(\mathbf{x}) - h(\mathbf{x})\right\}^2\right] = \mathbb{E}_{\mathbf{x}}\left[\left\{\frac{1}{M}\sum_{m=1}^{M} \epsilon_m(\mathbf{x})\right\}^2\right]$$

- If we assume:

$$\mathbb{E}_{\mathbf{x}}\left[\epsilon_m(\mathbf{x})\right] = 0$$
$$\mathbb{E}_{\mathbf{x}}\left[\epsilon_m(\mathbf{x})\epsilon_l(\mathbf{x})\right] = 0, \qquad m \neq l$$

errors due to the individual models are uncorrelated (not true in practice)

- Then $E_{\mathrm{COM}} = \frac{1}{M} E_{\mathrm{AV}}$

Outline

- Committees
- Boosting

Boosting

AdaBoost

1. Initialize the data weighting coefficients $\{w_n\}$ by setting $w_n^{(1)} = 1/N$ for $n = 1, \ldots, N$.
2. For $m = 1, \ldots, M$:
   (a) Fit a classifier $y_m(\mathbf{x})$ to the training data by minimizing the weighted error function
   $$J_m = \sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n) \tag{14.15}$$
   where $I(y_m(\mathbf{x}_n) \neq t_n)$ is the indicator function and equals 1 when $y_m(\mathbf{x}_n) \neq t_n$ and 0 otherwise.
   (b) Evaluate the quantities
   $$\epsilon_m = \frac{\sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n)}{\sum_{n=1}^{N} w_n^{(m)}} \tag{14.16}$$
   and then use these to evaluate
   $$\alpha_m = \ln\left\{\frac{1-\epsilon_m}{\epsilon_m}\right\}. \tag{14.17}$$
   (c) Update the data weighting coefficients
   $$w_n^{(m+1)} = w_n^{(m)} \exp\{\alpha_m I(y_m(\mathbf{x}_n) \neq t_n)\} \tag{14.18}$$
3. Make predictions using the final model, which is given by
$$Y_M(\mathbf{x}) = \operatorname{sign}\left(\sum_{m=1}^{M} \alpha_m y_m(\mathbf{x})\right). \tag{14.19}$$

$t_n \in \{-1,1\}$ $\quad$ $y(\mathbf{x}) \in \{-1,1\}$

the exponential error function： $E = \sum_{n=1}^{N} \exp\left\{-t_n f_m\left(\mathbf{x}_n\right)\right\}$

linear combination of base classifiers： $f_m(\mathbf{x}) = \frac{1}{2}\sum_{l=1}^{m} \alpha_l y_l(\mathbf{x})$

Q: How do they come up with this idea?

A: Motivated by finding a local optimal of E

Boosting

(a) Fit a classifier $y_m(\mathbf{x})$ to the training data by minimizing the weighted error function

$$J_m = \sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n) \qquad (14.15)$$

where $I(y_m(\mathbf{x}_n) \neq t_n)$ is the indicator function and equals 1 when $y_m(\mathbf{x}_n) \neq t_n$ and 0 otherwise.

the exponential error function: $E = \sum_{n=1}^{N} \exp\left\{-t_n f_m\left(\mathbf{x}_n\right)\right\}$

linear combination of base classifiers: $f_m(\mathbf{x}) = \frac{1}{2}\sum_{l=1}^{m} \alpha_l y_l(\mathbf{x})$

$t_n \in \{-1,1\} \qquad y(\mathbf{x}) \in \{-1,1\}$

Goal: minimize E with respect to both the weighting coefficients α_l and the parameters of the base classifiers $y_l(\mathbf{x})$

Suppose: $y_1(\mathbf{x})$, …, $y_{m-1}(\mathbf{x})$, α_1, …, α_{m-1} are fixed, minimize only with respect to $y_m(\mathbf{x})$ and α_m

$$\begin{aligned}
E &= \sum_{n=1}^{N} \exp\left\{-t_n f_{m-1}\left(\mathbf{x}_n\right) - \frac{1}{2} t_n \alpha_m y_m\left(\mathbf{x}_n\right)\right\} \\
&= \sum_{n=1}^{N} w_n^{(m)} \exp\left\{-\frac{1}{2} t_n \alpha_m y_m\left(\mathbf{x}_n\right)\right\} \\
&= e^{-\alpha_m/2} \sum_{n \in \mathcal{T}_m} w_n^{(m)} + e^{\alpha_m/2} \sum_{n \in \mathcal{M}_m} w_n^{(m)} \\
&= \left(e^{\alpha_m/2} - e^{-\alpha_m/2}\right) \sum_{n=1}^{N} w_n^{(m)} I\left(y_m\left(\mathbf{x}_n\right) \neq t_n\right) + e^{-\alpha_m/2} \sum_{n=1}^{N} w_n^{(m)}
\end{aligned}$$

minimize E w.r.t $y_m(\mathbf{x})$:

$$\hat{y}_m = \arg\min_{y_m} \sum_{n=1}^{N} w_n^{(m)} I\left(y_m\left(\mathbf{x}_n\right) \neq t_n\right)$$

Boosting

AdaBoost

1. Initialize the data weighting coefficients $\{w_n\}$ by setting $w_n^{(1)} = 1/N$ for $n = 1, \ldots, N$.
2. For $m = 1, \ldots, M$:
   (a) Fit a classifier $y_m(\mathbf{x})$ to the training data by minimizing the weighted error function

$$J_m = \sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n) \tag{14.15}$$

   where $I(y_m(\mathbf{x}_n) \neq t_n)$ is the indicator function and equals 1 when $y_m(\mathbf{x}_n) \neq t_n$ and 0 otherwise.

   (b) Evaluate the quantities

$$\epsilon_m = \frac{\sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n)}{\sum_{n=1}^{N} w_n^{(m)}} \tag{14.16}$$

   and then use these to evaluate

$$\alpha_m = \ln\left\{\frac{1-\epsilon_m}{\epsilon_m}\right\}. \tag{14.17}$$

   (c) Update the data weighting coefficients

$$w_n^{(m+1)} = w_n^{(m)} \exp\{\alpha_m I(y_m(\mathbf{x}_n) \neq t_n)\} \tag{14.18}$$

3. Make predictions using the final model, which is given by

$$Y_M(\mathbf{x}) = \operatorname{sign}\left(\sum_{m=1}^{M} \alpha_m y_m(\mathbf{x})\right). \tag{14.19}$$

$t_n \in \{-1,1\}$ $\quad$ $y(\mathbf{x}) \in \{-1,1\}$

the exponential error function： $E = \sum_{n=1}^{N} \exp\{-t_n f_m(\mathbf{x}_n)\}$

linear combination of base classifiers： $f_m(\mathbf{x}) = \frac{1}{2}\sum_{l=1}^{m} \alpha_l y_l(\mathbf{x})$

Q: How do they come up with this idea?

A: Motivated by finding a local optimal of E

Boosting

(c) Update the data weighting coefficients

$$w_n^{(m+1)} = w_n^{(m)} \exp\{\alpha_m I(y_m(\mathbf{x}_n) \neq t_n)\} \qquad (14.18)$$

the exponential error function： $E = \sum_{n=1}^{N} \exp\{-t_n f_m(\mathbf{x}_n)\}$

linear combination of base classifiers： $f_m(\mathbf{x}) = \frac{1}{2}\sum_{l=1}^{m} \alpha_l y_l(\mathbf{x})$

$t_n \in \{-1,1\} \qquad y(\mathbf{x}) \in \{-1,1\}$

Goal: minimize E with respect to both the weighting coefficients α_l and the parameters of the base classifiers $y_l(\mathbf{x})$

Suppose: $y_1(\mathbf{x}), \ldots, y_{m-1}(\mathbf{x}), \alpha_1, \ldots, \alpha_{m-1}$ are fixed, minimize only with respect to $y_m(\mathbf{x})$ and α_m

$$E = \sum_{n=1}^{N} \exp\left\{-t_n f_{m-1}(\mathbf{x}_n) - \frac{1}{2} t_n \alpha_m y_m(\mathbf{x}_n)\right\}$$
$$= \sum_{n=1}^{N} w_n^{(m)} \exp\left\{-\frac{1}{2} t_n \alpha_m y_m(\mathbf{x}_n)\right\}$$

$$W_n^{(m+1)} = \exp\{-t_n f_m(x_n)\}$$
$$= \exp\{-t_n(f_{m-1}(x_n) + \tfrac{1}{2}\alpha_m y_m(x_n))\}$$
$$= \exp\{-t_n f_{m-1}(x_n) - \tfrac{1}{2} t_n \alpha_m y_m(x_n)\}$$
$$= W_n^{(m)} \exp\{-\tfrac{1}{2} t_n \alpha_m y_m(x_n)\}$$

$t_n y_m(x_n) = 1 - 2\mathbb{I}(y_m(x_n) \neq t_n)$

$\Rightarrow$

$$W_n^{(m+1)} = W_n^{(m)} \exp\{-\tfrac{1}{2}\alpha_m(1 - 2\mathbb{I}(y_m(x_n) \neq t_n))\}$$
$$= W_n^{(m)} \underbrace{\exp\{-\tfrac{1}{2}\alpha_m\}}_{\text{same for all } n \in [N]} \exp\{\alpha_m \mathbb{I}(y_m(x_n) \neq t_n\}$$

Boosting

AdaBoost

1. Initialize the data weighting coefficients $\{w_n\}$ by setting $w_n^{(1)} = 1/N$ for $n = 1, \dots, N$.
2. For $m = 1, \dots, M$:
   (a) Fit a classifier $y_m(\mathbf{x})$ to the training data by minimizing the weighted error function

$$J_m = \sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n) \tag{14.15}$$

   where $I(y_m(\mathbf{x}_n) \neq t_n)$ is the indicator function and equals 1 when $y_m(\mathbf{x}_n) \neq t_n$ and 0 otherwise.

   (b) Evaluate the quantities

$$\epsilon_m = \frac{\sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n)}{\sum_{n=1}^{N} w_n^{(m)}} \tag{14.16}$$

   and then use these to evaluate

$$\alpha_m = \ln \left\{ \frac{1 - \epsilon_m}{\epsilon_m} \right\}. \tag{14.17}$$

   (c) Update the data weighting coefficients

$$w_n^{(m+1)} = w_n^{(m)} \exp \{ \alpha_m I(y_m(\mathbf{x}_n) \neq t_n) \} \tag{14.18}$$

3. Make predictions using the final model, which is given by

$$Y_M(\mathbf{x}) = \operatorname{sign}\left(\sum_{m=1}^{M} \alpha_m y_m(\mathbf{x}) \right). \tag{14.19}$$

$t_n \in \{-1,1\}$ $\quad$ $y(\mathbf{x}) \in \{-1,1\}$

the exponential error function： $E = \sum_{n=1}^{N} \exp\{-t_n f_m(\mathbf{x}_n)\}$

linear combination of base classifiers： $f_m(\mathbf{x}) = \frac{1}{2} \sum_{l=1}^{m} \alpha_l y_l(\mathbf{x})$

Q: How do they come up with this idea?

A: Motivated by finding a local optimal of E

Boosting

(b) Evaluate the quantities

$$\epsilon_m = \frac{\sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n)}{\sum_{n=1}^{N} w_n^{(m)}} \tag{14.16}$$

and then use these to evaluate

$$\alpha_m = \ln\left\{\frac{1-\epsilon_m}{\epsilon_m}\right\}. \tag{14.17}$$

the exponential error function： $E = \sum_{n=1}^{N} \exp\left\{-t_n f_m\left(\mathbf{x}_n\right)\right\}$

linear combination of base classifiers： $f_m(\mathbf{x}) = \frac{1}{2}\sum_{l=1}^{m} \alpha_l y_l(\mathbf{x})$

$t_n \in \{-1,1\} \qquad y(\mathbf{x}) \in \{-1,1\}$

Goal: minimize E with respect to both the weighting coefficients α_l and the parameters of the base classifiers $y_l(\mathbf{x})$

Suppose: $y_1(\mathbf{x}), \ldots, y_{m-1}(\mathbf{x}), \alpha_1, \ldots, \alpha_{m-1}$ are fixed, minimize only with respect to $y_m(\mathbf{x})$ and α_m

$$\begin{aligned} E &= \sum_{n=1}^{N} \exp\left\{-t_n f_{m-1}\left(\mathbf{x}_n\right) - \frac{1}{2} t_n \alpha_m y_m\left(\mathbf{x}_n\right)\right\} \\ &= \sum_{n=1}^{N} w_n^{(m)} \exp\left\{-\frac{1}{2} t_n \alpha_m y_m\left(\mathbf{x}_n\right)\right\} \\ &= e^{-\alpha_m/2} \sum_{n \in \mathcal{T}_m} w_n^{(m)} + e^{\alpha_m/2} \sum_{n \in \mathcal{M}_m} w_n^{(m)} \\ &= \left(e^{\alpha_m/2} - e^{-\alpha_m/2}\right) \sum_{n=1}^{N} w_n^{(m)} I\left(y_m\left(\mathbf{x}_n\right) \neq t_n\right) + e^{-\alpha_m/2} \sum_{n=1}^{N} w_n^{(m)} \end{aligned}$$

minimize E w.r.t α_m:

Boosting

(b) Evaluate the quantities

$$\epsilon_m = \frac{\sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n)}{\sum_{n=1}^{N} w_n^{(m)}} \quad (14.16)$$

and then use these to evaluate

$$\alpha_m = \ln\left\{\frac{1-\epsilon_m}{\epsilon_m}\right\}. \quad (14.17)$$

We want to minimize E w.r.t α_m, where

$$E = \left(e^{\alpha_m/2} - e^{-\alpha_m/2}\right) \sum_{n=1}^{N} w_n^{(m)} I\left(y_m\left(\mathbf{x}_n\right) \neq t_n\right) + e^{-\alpha_m/2} \sum_{n=1}^{N} w_n^{(m)}$$

$$\frac{\partial E}{\partial \alpha_m} = \frac{1}{2}\left(\left(e^{\alpha_m/2} + e^{-\alpha_m/2}\right) \sum_{n=1}^{N} w_n^{(m)} I\left(y_m\left(\mathbf{x}_n\right) \neq t_n\right) - e^{-\alpha_m/2} \sum_{n=1}^{N} w_n^{(m)}\right)$$

set it to zero and rearrange:

$$\frac{\sum_n w_n^{(m)} I\left(y_m\left(\mathbf{x}_n\right) \neq t_n\right)}{\sum_n w_n^{(m)}} = \frac{e^{-\alpha_m/2}}{e^{\alpha_m/2} + e^{-\alpha_m/2}} = \frac{1}{e^{\alpha_m} + 1} = \epsilon_m$$

$$\Rightarrow e^{\alpha_m} = \frac{1-\epsilon_m}{\epsilon_m}$$

$$\Rightarrow \alpha_m = \log \frac{1-\epsilon_m}{\epsilon_m}$$

Boosting

AdaBoost

1. Initialize the data weighting coefficients $\{w_n\}$ by setting $w_n^{(1)} = 1/N$ for $n = 1, \ldots, N$.
2. For $m = 1, \ldots, M$:
   (a) Fit a classifier $y_m(\mathbf{x})$ to the training data by minimizing the weighted error function

$$J_m = \sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n) \tag{14.15}$$

   where $I(y_m(\mathbf{x}_n) \neq t_n)$ is the indicator function and equals 1 when $y_m(\mathbf{x}_n) \neq t_n$ and 0 otherwise.

   (b) Evaluate the quantities

$$\epsilon_m = \frac{\sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n)}{\sum_{n=1}^{N} w_n^{(m)}} \tag{14.16}$$

   and then use these to evaluate

$$\alpha_m = \ln\left\{\frac{1 - \epsilon_m}{\epsilon_m}\right\}. \tag{14.17}$$

   (c) Update the data weighting coefficients

$$w_n^{(m+1)} = w_n^{(m)} \exp\{\alpha_m I(y_m(\mathbf{x}_n) \neq t_n)\} \tag{14.18}$$

3. Make predictions using the final model, which is given by

$$Y_M(\mathbf{x}) = \operatorname{sign}\left(\sum_{m=1}^{M} \alpha_m y_m(\mathbf{x})\right). \tag{14.19}$$

$t_n \in \{-1,1\}$ $\quad y(\mathbf{x}) \in \{-1,1\}$

This is trivial

the exponential error function: $E = \sum_{n=1}^{N} \exp\left\{-t_n f_m\left(\mathbf{x}_n\right)\right\}$

linear combination of base classifiers: $f_m(\mathbf{x}) = \frac{1}{2}\sum_{l=1}^{m} \alpha_l y_l(\mathbf{x})$

Q: How do they come up with this idea?

A: Motivated by finding a local optimal of E

Reference

- PRML chapter 14