Bias-variance decomposition in Random Forests

Gilles Louppe @glouppe

Paris ML S02E04, December 8, 2014

Motivation

In supervised learning, combining the predictions of several randomized models often achieves better results than a single non-randomized model.

Why ?

Supervised learning

- The inputs are random variables $X = X_1, \ldots, X_p$;
- The output is a random variable Y .

- Data comes as a finite learning set

$$\mathcal{L} = \{(\mathbf{x}_i, y_i) | i = 0, \ldots, N-1\},$$

where $\mathbf{x}_i \in \mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_p$ and $y_i \in \mathcal{Y}$ are randomly drawn from $P_{X,Y}$.

- The goal is to find a model $\varphi_{\mathcal{L}} : \mathcal{X} \mapsto \mathcal{Y}$ minimizing

$$Err(\varphi_{\mathcal{L}}) = \mathbb{E}_{X,Y}\{L(Y, \varphi_{\mathcal{L}}(X))\}.$$

Performance evaluation

Classification

- Symbolic output (e.g., $\mathcal{Y} = \{\text{yes}, \text{no}\}$)
- Zero-one loss

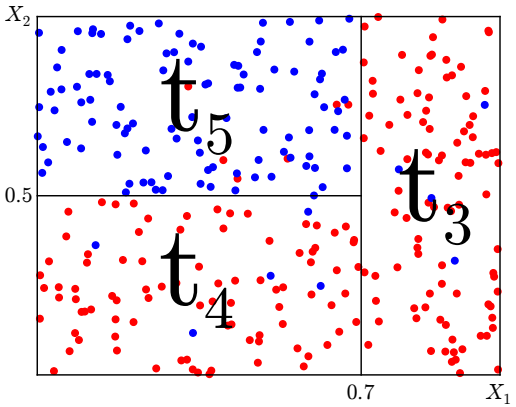
$$L(Y, \varphi_{\mathcal{L}}(X)) = 1(Y \neq \varphi_{\mathcal{L}}(X))$$

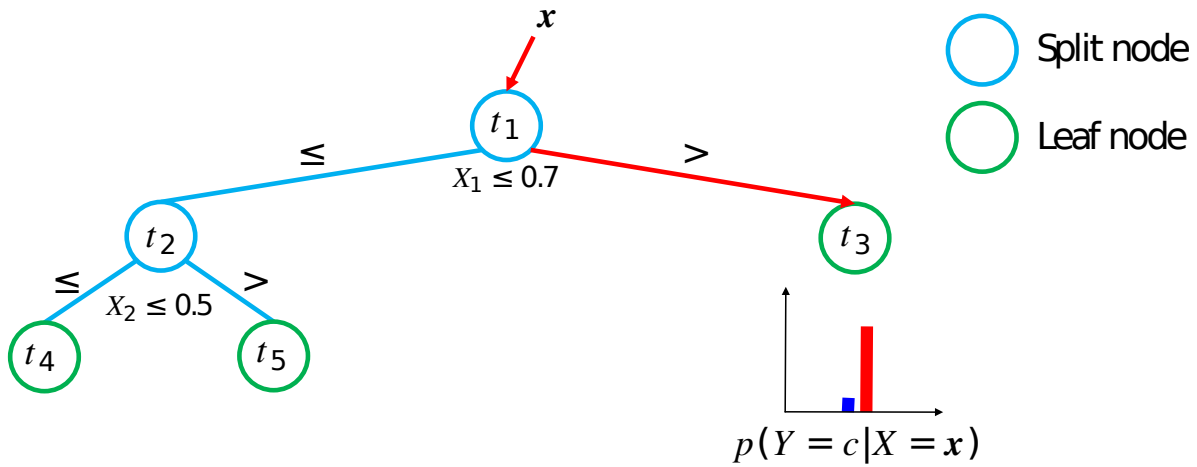
Regression

- Numerical output (e.g., $\mathcal{Y} = \mathbb{R}$)
- Squared error loss

$$L(Y, \varphi_{\mathcal{L}}(X)) = (Y - \varphi_{\mathcal{L}}(X))^2$$

Decision trees





$t \in \varphi$: nodes of the tree φ

X_t : split variable at t

$v_t \in \mathbb{R}$: split threshold at t

$\varphi(\mathbf{x}) = \arg\max_{c \in \mathcal{Y}} p(Y = c|X = \mathbf{x})$

Bias-variance decomposition in regression

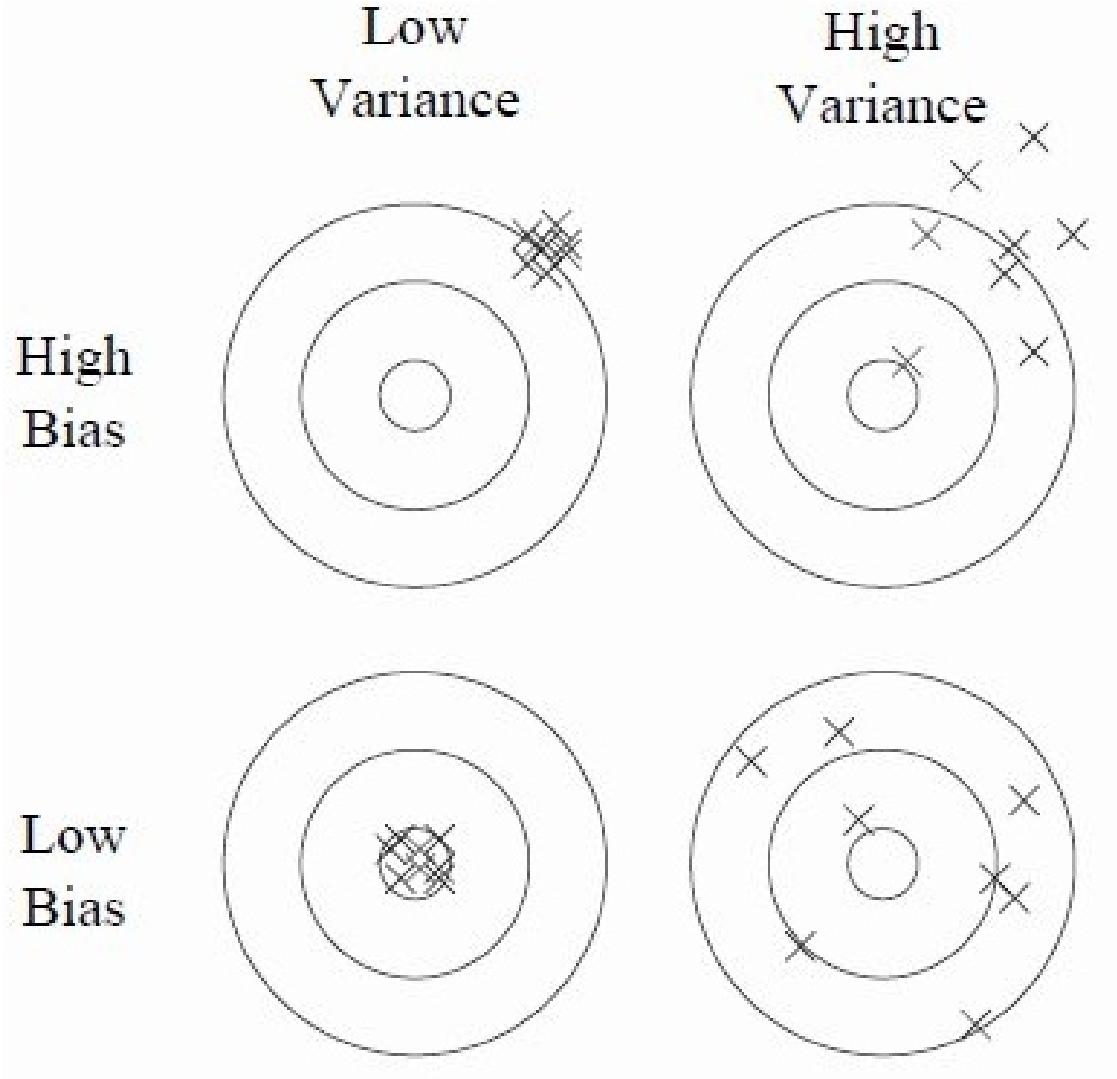
Theorem. For the *squared error loss*, the bias-variance decomposition of the expected generalization error at $X = \mathbf{x}$ is

$$\mathbb{E}_{\mathcal{L}}\{Err(\varphi_{\mathcal{L}}(\mathbf{x}))\} = \text{noise}(\mathbf{x}) + \text{bias}^2(\mathbf{x}) + \text{var}(\mathbf{x})$$

where

$$\begin{aligned}
\text{noise}(\mathbf{x}) &= Err(\varphi_B(\mathbf{x})), \\
\text{bias}^2(\mathbf{x}) &= (\varphi_B(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\})^2, \\
\text{var}(\mathbf{x}) &= \mathbb{E}_{\mathcal{L}}\{(\mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\} - \varphi_{\mathcal{L}}(\mathbf{x}))^2\}.
\end{aligned}$$

Bias-variance decomposition

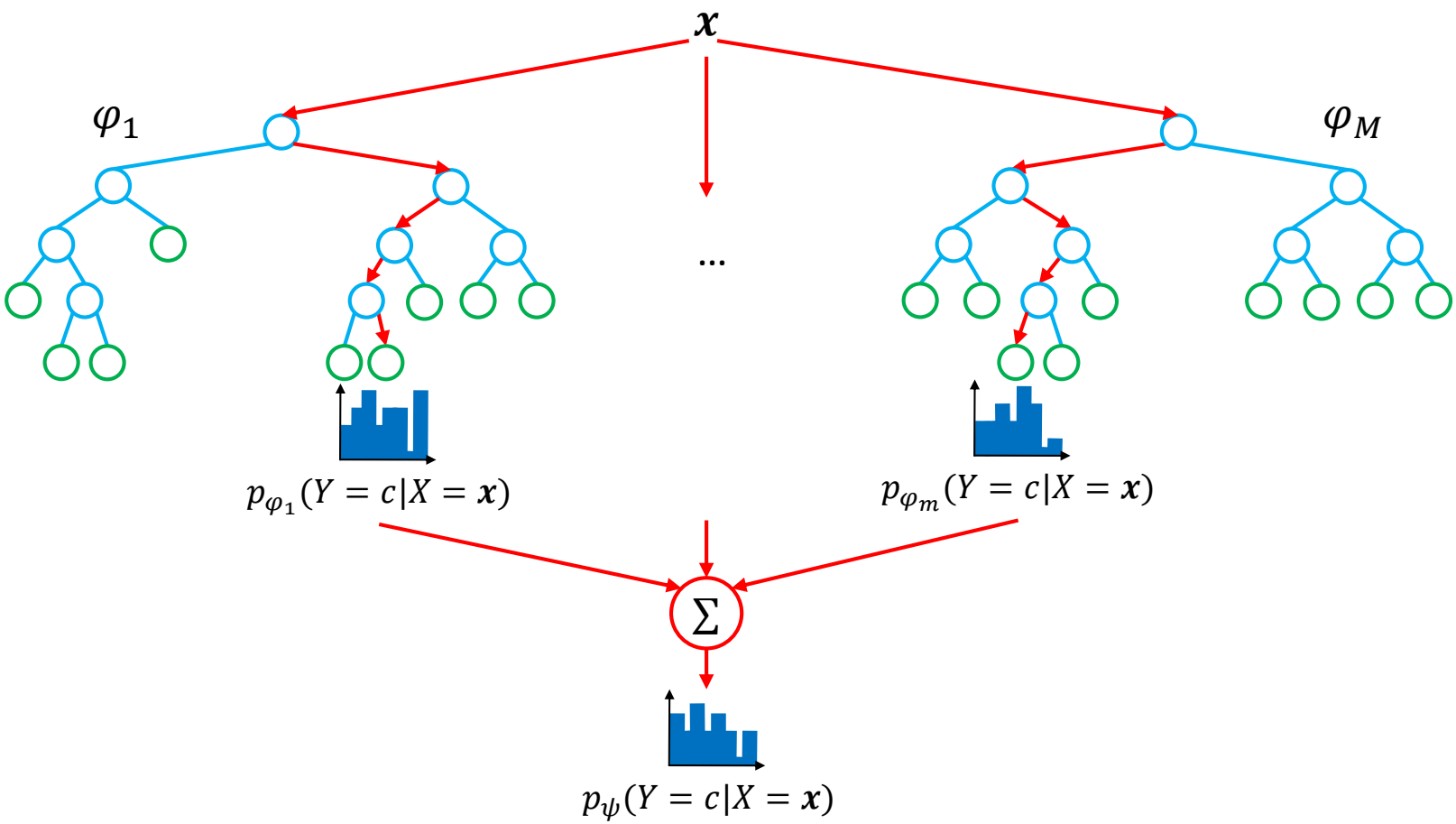


Diagnosing the generalization error of a decision tree

- (Residual error : Lowest achievable error, independent of $\varphi_{\mathcal{L}}$.)
- Bias : Decision trees usually have low bias.
- Variance : They often suffer from high variance.

- Solution : *Combine the predictions of several randomized trees into a single model.*

Random forests



Randomization

- Bootstrap samples
- Random selection of $K \leqslant p$ split variables } Random Forests
- Random selection of the threshold } Extra-Trees

Bias-variance decomposition (cont.)

Theorem. For the *squared error loss*, the bias-variance decomposition of the expected generalization error $\mathbb{E}_{\mathcal{L}}\{Err(\psi_{\mathcal{L},\theta_1,\ldots,\theta_M}(\mathbf{x}))\}$ at $X = \mathbf{x}$ of an ensemble of M randomized models $\varphi_{\mathcal{L},\theta_m}$ is

$$\mathbb{E}_{\mathcal{L}}\{Err(\psi_{\mathcal{L},\theta_1,\ldots,\theta_M}(\mathbf{x}))\} = \text{noise}(\mathbf{x}) + \text{bias}^2(\mathbf{x}) + \text{var}(\mathbf{x}),$$

where

$$\begin{aligned}
\text{noise}(\mathbf{x}) &= Err(\varphi_B(\mathbf{x})), \\
\text{bias}^2(\mathbf{x}) &= (\varphi_B(\mathbf{x}) - \mathbb{E}_{\mathcal{L},\theta}\{\varphi_{\mathcal{L},\theta}(\mathbf{x})\})^2, \\
\text{var}(\mathbf{x}) &= \rho(\mathbf{x})\sigma^2_{\mathcal{L},\theta}(\mathbf{x}) + \frac{1-\rho(\mathbf{x})}{M}\sigma^2_{\mathcal{L},\theta}(\mathbf{x}).
\end{aligned}$$

and where $\rho(\mathbf{x})$ is the Pearson correlation coefficient between the predictions of two randomized trees built on the same learning set.

Interpretation of $\rho(\mathbf{x})$ (Louppe, 2014)

Theorem. $\rho(\mathbf{x}) = \frac{\mathbb{V}_{\mathcal{L}}\{\mathbb{E}_{\theta|\mathcal{L}}\{\varphi_{\mathcal{L},\theta}(\mathbf{x})\}\}}{\mathbb{V}_{\mathcal{L}}\{\mathbb{E}_{\theta|\mathcal{L}}\{\varphi_{\mathcal{L},\theta}(\mathbf{x})\}\}+\mathbb{E}_{\mathcal{L}}\{\mathbb{V}_{\theta|\mathcal{L}}\{\varphi_{\mathcal{L},\theta}(\mathbf{x})\}\}}$

In other words, it is the ratio between

- the variance due to the learning set and
- the total variance, accounting for random effects due to both the learning set and the random perburbations.

$\rho(\mathbf{x}) \to 1$ when variance is mostly due to the learning set ;
$\rho(\mathbf{x}) \to 0$ when variance is mostly due to the random perturbations ;
$\rho(\mathbf{x}) \geqslant 0$.

Diagnosing the generalization error of random forests

- Bias : Identical to the bias of a single randomized tree.
- Variance : $\text{var}(\mathbf{x}) = \rho(\mathbf{x})\sigma^2_{\mathcal{L},\theta}(\mathbf{x}) + \frac{1-\rho(\mathbf{x})}{M}\sigma^2_{\mathcal{L},\theta}(\mathbf{x})$
  As $M \to \infty$, $\text{var}(\mathbf{x}) \to \rho(\mathbf{x})\sigma^2_{\mathcal{L},\theta}(\mathbf{x})$
  - The stronger the randomization, $\rho(\mathbf{x}) \to 0$, $\text{var}(\mathbf{x}) \to 0$.
  - The weaker the randomization, $\rho(\mathbf{x}) \to 1$, $\text{var}(\mathbf{x}) \to \sigma^2_{\mathcal{L},\theta}(\mathbf{x})$

Bias-variance trade-off. Randomization increases bias but makes it possible to reduce the variance of the corresponding ensemble model through averaging.

The crux of the problem is to find the right trade-off.

Tips : tune `max_features` in Random Forests.

Bias-variance decomposition in classification

Theorem. For the *zero-one loss* and binary classification, the expected generalization error $\mathbb{E}_{\mathcal{L}}\{Err(\varphi_{\mathcal{L}}(\mathbf{x}))\}$ at $X = \mathbf{x}$ decomposes as follows :

$$\mathbb{E}_{\mathcal{L}}\{Err(\varphi_{\mathcal{L}}(\mathbf{x}))\} = P(\varphi_B(\mathbf{x}) \neq Y) + \Phi\left(\frac{0.5 - \mathbb{E}_{\mathcal{L}}\{\widehat{p}_{\mathcal{L}}(Y = \varphi_B(\mathbf{x}))\}}{\sqrt{\mathbb{V}_{\mathcal{L}}\{\widehat{p}_{\mathcal{L}}(Y = \varphi_B(\mathbf{x}))\}}}\right)(2P(\varphi_B(\mathbf{x}) = Y) - 1)$$

- For $\mathbb{E}_{\mathcal{L}}\{\widehat{p}_{\mathcal{L}}(Y = \varphi_B(\mathbf{x})\} > 0.5$, $\mathbb{V}_{\mathcal{L}}\{\widehat{p}_{\mathcal{L}}(Y = \varphi_B(\mathbf{x}))\} \to 0$ makes $\Phi \to 0$ and the expected generalization error tends to the error of the Bayes model.
- Conversely, for $\mathbb{E}_{\mathcal{L}}\{\widehat{p}_{\mathcal{L}}(Y = \varphi_B(\mathbf{x})\} < 0.5$, $\mathbb{V}_{\mathcal{L}}\{\widehat{p}_{\mathcal{L}}(Y = \varphi_B(\mathbf{x}))\} \to 0$ makes $\Phi \to 1$ and the error is maximal.

Questions ?