Bias-variance decomposition in Random Forests

Gilles Louppe @glouppe

Paris ML S02E04, December 8, 2014
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

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

where $x_i \in \mathcal{X} = X_1 \times \ldots \times X_p$ and $y_i \in \mathcal{Y}$ are randomly drawn from $P_{\mathcal{X}, \mathcal{Y}}$.

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

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

Classification

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

$$L(Y, \varphi_L(X)) = 1(Y \neq \varphi_L(X))$$

Regression

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

$$L(Y, \varphi_L(X)) = (Y - \varphi_L(X))^2$$
Decision trees

\[ t \in \varphi : \text{nodes of the tree } \varphi \]
\[ X_t : \text{split variable at } t \]
\[ \nu_t \in \mathbb{R} : \text{split threshold at } t \]
\[ \varphi(x) = \arg \max_{c \in y} p(Y = c | X = x) \]
Theorem. For the squared error loss, the bias-variance decomposition of the expected generalization error at $X = x$ is

$$\mathbb{E}_L\{Err(\varphi_L(x))\} = \text{noise}(x) + \text{bias}^2(x) + \text{var}(x)$$

where

$$\text{noise}(x) = Err(\varphi_B(x)),$$
$$\text{bias}^2(x) = (\varphi_B(x) - \mathbb{E}_L\{\varphi_L(x)\})^2,$$
$$\text{var}(x) = \mathbb{E}_L\{(\mathbb{E}_L\{\varphi_L(x)\} - \varphi_L(x))^2\}.$$
Bias-variance decomposition
Diagnosing the generalization error of a decision tree

- (Residual error: Lowest achievable error, independent of \( \varphi_L \).)
- 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

\[ x \]

\[ \sum \]

\[ p_{\varphi_1}(Y = c|X = x) \]

\[ p_{\varphi_m}(Y = c|X = x) \]

\[ p_{\psi}(Y = c|X = x) \]

Randomization

- Bootstrap samples
- Random selection of \( K \leq p \) split variables
- Random selection of the threshold

\{ Random Forests \}

\{ Extra-Trees \}
Bias-variance decomposition (cont.)

**Theorem.** For the *squared error loss*, the bias-variance decomposition of the expected generalization error
\[ \mathbb{E}_{\mathcal{L}}\{\text{Err}(\psi_{\mathcal{L},\theta_{1},\ldots,\theta_{M}}(x))\} \]

at \( X = x \) of an ensemble of \( M \) randomized models \( \varphi_{\mathcal{L},\theta_{m}} \) is

\[ \mathbb{E}_{\mathcal{L}}\{\text{Err}(\psi_{\mathcal{L},\theta_{1},\ldots,\theta_{M}}(x))\} = \text{noise}(x) + \text{bias}^{2}(x) + \text{var}(x), \]

where

\[ \text{noise}(x) = \text{Err}(\varphi_{B}(x)), \]
\[ \text{bias}^{2}(x) = (\varphi_{B}(x) - \mathbb{E}_{\mathcal{L},\theta}\{\varphi_{\mathcal{L},\theta}(x)\})^{2}, \]
\[ \text{var}(x) = \rho(x)\sigma_{\mathcal{L},\theta}^{2}(x) + \frac{1 - \rho(x)}{M}\sigma_{\mathcal{L},\theta}^{2}(x). \]

and where \( \rho(x) \) is the Pearson correlation coefficient between the predictions of two randomized trees built on the same learning set.
Interpretation of $\rho(x)$ (Louppe, 2014)

**Theorem.** $\rho(x) = \frac{\nabla_{\mathcal{L}} \{ E_{\theta | \mathcal{L}} \{ \phi_{\mathcal{L}, \theta}(x) \} \}}{\nabla_{\mathcal{L}} \{ E_{\theta | \mathcal{L}} \{ \phi_{\mathcal{L}, \theta}(x) \} \} + E_{\mathcal{L}} \{ \nabla_{\theta | \mathcal{L}} \{ \phi_{\mathcal{L}, \theta}(x) \} \}}$

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 perturbations.

$\rho(x) \rightarrow 1$ when variance is mostly due to the learning set; $\rho(x) \rightarrow 0$ when variance is mostly due to the random perturbations; $\rho(x) \geq 0$. 
Diagnosing the generalization error of random forests

- **Bias**: Identical to the bias of a single randomized tree.
- **Variance**: \( \text{var}(x) = \rho(x) \sigma_{\mathcal{L}, \theta}(x) + \frac{1-\rho(x)}{M} \sigma_{\mathcal{L}, \theta}(x) \)

As \( M \to \infty \), \( \text{var}(x) \to \rho(x) \sigma_{\mathcal{L}, \theta}(x) \)

- The stronger the randomization, \( \rho(x) \to 0 \), \( \text{var}(x) \to 0 \).
- The weaker the randomization, \( \rho(x) \to 1 \), \( \text{var}(x) \to \sigma_{\mathcal{L}, \theta}(x) \).

**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 \( \mathbb{E}_L\{\text{Err}(\varphi_L(x))\} \) at \( X = x \) decomposes as follows:

\[
\mathbb{E}_L\{\text{Err}(\varphi_L(x))\} = P(\varphi_B(x) \neq Y)
+ \Phi\left(\frac{0.5 - \mathbb{E}_L\{\hat{p}_L(Y = \varphi_B(x))\}}{\sqrt{\nabla_L\{\hat{p}_L(Y = \varphi_B(x))\}}}\right)(2P(\varphi_B(x) = Y) - 1)
\]

- For \( \mathbb{E}_L\{\hat{p}_L(Y = \varphi_B(x))\} > 0.5, \nabla_L\{\hat{p}_L(Y = \varphi_B(x))\} \to 0 \) makes \( \Phi \to 0 \) and the expected generalization error tends to the error of the Bayes model.
- Conversely, for \( \mathbb{E}_L\{\hat{p}_L(Y = \varphi_B(x))\} < 0.5, \nabla_L\{\hat{p}_L(Y = \varphi_B(x))\} \to 0 \) makes \( \Phi \to 1 \) and the error is maximal.
Questions ?