Machine Learning Course
basic track

Lecture 7: Bias-variance decomposition & Feature importances

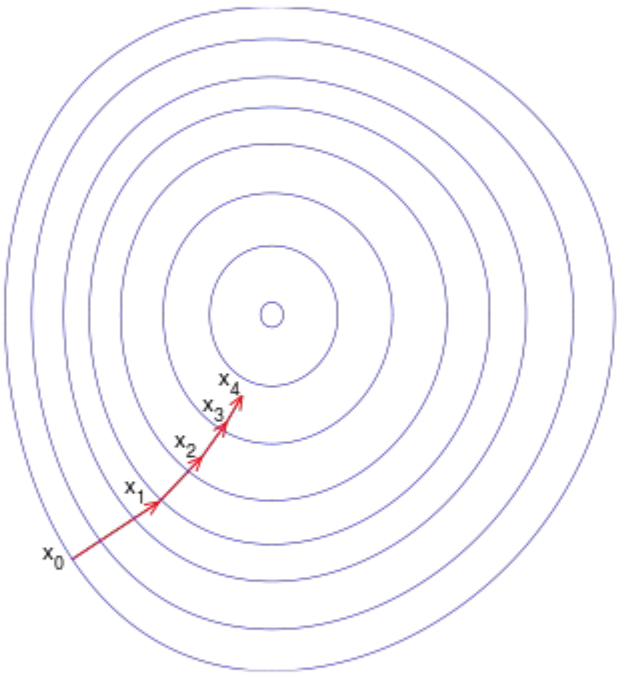
MIPT, 2019

Outline

1. Gradient boosting recap
2. Bias-variance decomposition.
3. Feature importances estimation

Based on ml-mipt materials by Victor Kantor and Evgeny Sokolov HSE 2018 materials

Gradient boosting: recap



We use gradient descent in *space of our models*

Gradient boosting: recap

$$\hat{f}(x) = \sum_{i=0}^{t-1} \hat{f}_i(x),$$

$$r_{it} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}(x)}, \quad \text{for } i = 1, \dots, n,$$

$$\theta_t = \arg\min_{\theta} \sum_{i=1}^{n} (r_{it} - h(x_i, \theta))^2,$$

$$\rho_t = \arg\min_{\rho} \sum_{i=1}^{n} L(y_i, \hat{f}(x_i) + \rho \cdot h(x_i, \theta_t))$$

Gradient boosting: recap

In linear regression case with MSE loss:

$$r_{it} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}(x)} = -2(\hat{y}_i - y_i) \propto \hat{y}_i - y_i$$

Gradient boosting: recap

What we need:

- Data.
- Loss function and its gradient.
- Family of algorithms (with constraints on hyperparameters if necessary).
- Number of iterations M.
- Initial value (GBM by Friedman): constant.

Bias-variance decomposition

The dataset $X = (x_i, y_i)_{i=1}^{\ell}$ with $y_i \in \mathbb{R}$ for regression problem.

Bias-variance decomposition

The dataset $X = (x_i, y_i)_{i=1}^{\ell}$ with $y_i \in \mathbb{R}$ for regression problem.

Denote loss function $L(y, a) = \big(y - a(x)\big)^2$.

Bias-variance decomposition

The dataset $X = (x_i, y_i)_{i=1}^{\ell}$ with $y_i \in \mathbb{R}$ for regression problem.

Denote loss function $L(y, a) = \big(y - a(x)\big)^2$.

The corresponding risk estimation is

$$R(a) = \mathbb{E}_{x,y}\Big[\big(y - a(x)\big)^2\Big] = \int_{\mathbb{X}}\int_{\mathbb{Y}} p(x, y)\big(y - a(x)\big)^2 dx dy.$$

Bias-variance decomposition

Let's show that $a_*(x) = \mathbb{E}[y \mid x] = \int_{\mathbb{Y}} yp(y \mid x)dy$

Bias-variance decomposition

Let's show that $a_*(x) = \mathbb{E}[y \mid x] = \int_{\mathbb{Y}} y p(y \mid x) dy = \arg\min_a R(a).$

Bias-variance decomposition

Let's show that $a_*(x) = \mathbb{E}[y \mid x] = \int_{\mathbb{Y}} yp(y \mid x)dy = \arg\min_a R(a).$

$$L(y, a(x)) = (y - a(x))^2$$

Bias-variance decomposition

Let's show that $$a_*(x) = \mathbb{E}[y \mid x] = \int_{\mathbb{Y}} y p(y \mid x) dy = \arg\min_a R(a).$$

$$L(y, a(x)) = (y - a(x))^2 = (y - \mathbb{E}(y \mid x) + \mathbb{E}(y \mid x) - a(x))^2 =$$

Bias-variance decomposition

Let's show that $a_*(x) = \mathbb{E}[y \,|\, x] = \int_{\mathbb{Y}} yp(y \,|\, x)dy = \arg\min_a R(a).$

$$L(y, a(x)) = (y - a(x))^2 = (y - \mathbb{E}(y \,|\, x) + \mathbb{E}(y \,|\, x) - a(x))^2 =$$
$$= (y - \mathbb{E}(y \,|\, x))^2 + 2\big(y - \mathbb{E}(y \,|\, x)\big)\big(\mathbb{E}(y \,|\, x) - a(x)\big) + (\mathbb{E}(y \,|\, x) - a(x))^2.$$

Let's return to the risk estimation:

Bias-variance decomposition

Let's show that $a_*(x) = \mathbb{E}[y \,|\, x] = \int_{\mathbb{Y}} y p(y \,|\, x) dy = \arg\min_a R(a).$

$$L(y, a(x)) = (y - a(x))^2 = (y - \mathbb{E}(y \,|\, x) + \mathbb{E}(y \,|\, x) - a(x))^2 =$$
$$= (y - \mathbb{E}(y \,|\, x))^2 + 2\big(y - \mathbb{E}(y \,|\, x)\big)\big(\mathbb{E}(y \,|\, x) - a(x)\big) + (\mathbb{E}(y \,|\, x) - a(x))^2.$$

Let's return to the risk estimation:

$$R(a) = \mathbb{E}_{x,y} L(y, a(x))$$

Bias-variance decomposition

Let's show that $a_*(x) = \mathbb{E}[y \mid x] = \int_{\mathbb{Y}} y p(y \mid x) dy = \arg\min_a R(a).$

$$L(y, a(x)) = (y - a(x))^2 = (y - \mathbb{E}(y \mid x) + \mathbb{E}(y \mid x) - a(x))^2 =$$
$$= (y - \mathbb{E}(y \mid x))^2 + 2\big(y - \mathbb{E}(y \mid x)\big)\big(\mathbb{E}(y \mid x) - a(x)\big) + (\mathbb{E}(y \mid x) - a(x))^2.$$

Let's return to the risk estimation:

$$\begin{aligned} R(a) &= \mathbb{E}_{x,y} L(y, a(x)) = \\ &= \mathbb{E}_{x,y}(y - \mathbb{E}(y \mid x))^2 + \mathbb{E}_{x,y}(\mathbb{E}(y \mid x) - a(x))^2 + \\ &+ 2\mathbb{E}_{x,y}\big(y - \mathbb{E}(y \mid x)\big)\big(\mathbb{E}(y \mid x) - a(x)\big). \end{aligned}$$

Focus on the last term:

$$
\begin{aligned}
R(a) &= \mathbb{E}_{x,y} L(y, a(x)) = \\
&= \mathbb{E}_{x,y}(y - \mathbb{E}(y \mid x))^2 + \mathbb{E}_{x,y}(\mathbb{E}(y \mid x) - a(x))^2 + \\
&+ 2\mathbb{E}_{x,y}\big(y - \mathbb{E}(y \mid x)\big)\big(\mathbb{E}(y \mid x) - a(x)\big).
\end{aligned}
$$

$$
\begin{aligned}
R(a) &= \mathbb{E}_{x,y} L(y, a(x)) = \\
&= \mathbb{E}_{x,y}(y - \mathbb{E}(y \mid x))^2 + \mathbb{E}_{x,y}(\mathbb{E}(y \mid x) - a(x))^2 + \\
&+ 2\mathbb{E}_{x,y}\big(y - \mathbb{E}(y \mid x)\big)\big(\mathbb{E}(y \mid x) - a(x)\big).
\end{aligned}
$$

Focus on the last term:

$$
\mathbb{E}_x \mathbb{E}_y \Big[\big(y - \mathbb{E}(y \mid x)\big)\big(\mathbb{E}(y \mid x) - a(x)\big) \mid x \Big] =
$$

$$R(a) = \mathbb{E}_{x,y} L(y, a(x)) =$$

$$= \mathbb{E}_{x,y}(y - \mathbb{E}(y \mid x))^2 + \mathbb{E}_{x,y}(\mathbb{E}(y \mid x) - a(x))^2 +$$

Focus on the last term:

$$+ 2\mathbb{E}_{x,y}\big(y - \mathbb{E}(y \mid x)\big)\big(\mathbb{E}(y \mid x) - a(x)\big).$$

Does not depend on y

$$\mathbb{E}_x \mathbb{E}_y \Big[\big(y - \mathbb{E}(y \mid x)\big) \boxed{\big(\mathbb{E}(y \mid x) - a(x)\big)} \mid x \Big] =$$

$$R(a) = \mathbb{E}_{x,y} L(y, a(x)) =$$
$$= \mathbb{E}_{x,y}(y - \mathbb{E}(y \,|\, x))^2 + \mathbb{E}_{x,y}(\mathbb{E}(y \,|\, x) - a(x))^2 +$$

Focus on the last term:

$$+\, 2\mathbb{E}_{x,y}\big(y - \mathbb{E}(y \,|\, x)\big)\big(\mathbb{E}(y \,|\, x) - a(x)\big).$$

Does not depend on y

$$\mathbb{E}_x \mathbb{E}_y \Big[\big(y - \mathbb{E}(y \,|\, x)\big) \boxed{\big(\mathbb{E}(y \,|\, x) - a(x)\big)} \,|\, x \Big] =$$
$$= \mathbb{E}_x \Big(\big(\mathbb{E}(y \,|\, x) - a(x)\big) \mathbb{E}_y \Big[\big(y - \mathbb{E}(y \,|\, x)\big) \,|\, x \Big] \Big) =$$

$$\begin{aligned} R(a) &= \mathbb{E}_{x,y} L(y, a(x)) = \\ &= \mathbb{E}_{x,y}(y - \mathbb{E}(y \,|\, x))^2 + \mathbb{E}_{x,y}(\mathbb{E}(y \,|\, x) - a(x))^2 + \\ &+ 2\mathbb{E}_{x,y}\big(y - \mathbb{E}(y \,|\, x)\big)\big(\mathbb{E}(y \,|\, x) - a(x)\big). \end{aligned}$$

Focus on the last term:

$$\begin{aligned} &\mathbb{E}_x \mathbb{E}_y \Big[\big(y - \mathbb{E}(y \,|\, x)\big)\big(\mathbb{E}(y \,|\, x) - a(x)\big) \,|\, x \Big] = \\ &= \mathbb{E}_x \Big(\big(\mathbb{E}(y \,|\, x) - a(x)\big) \mathbb{E}_y \Big[\big(y - \mathbb{E}(y \,|\, x)\big) \,|\, x \Big] \Big) = \\ &= \mathbb{E}_x \Big(\big(\mathbb{E}(y \,|\, x) - a(x)\big)\big(\mathbb{E}(y \,|\, x) - \mathbb{E}(y \,|\, x)\big) \Big) = \end{aligned}$$

$$
\begin{aligned}
R(a) &= \mathbb{E}_{x,y} L(y, a(x)) = \\
&= \mathbb{E}_{x,y}(y - \mathbb{E}(y \,|\, x))^2 + \mathbb{E}_{x,y}(\mathbb{E}(y \,|\, x) - a(x))^2 + \\
&+ 2\mathbb{E}_{x,y}\big(y - \mathbb{E}(y \,|\, x)\big)\big(\mathbb{E}(y \,|\, x) - a(x)\big).
\end{aligned}
$$

Focus on the last term:

$$
\begin{aligned}
&\mathbb{E}_x \mathbb{E}_y \Big[\big(y - \mathbb{E}(y \,|\, x)\big)\big(\mathbb{E}(y \,|\, x) - a(x)\big) \,|\, x\Big] = \\
&= \mathbb{E}_x \Big(\big(\mathbb{E}(y \,|\, x) - a(x)\big)\mathbb{E}_y \Big[\big(y - \mathbb{E}(y \,|\, x)\big) \,|\, x\Big]\Big) = \\
&= \mathbb{E}_x \Big(\big(\mathbb{E}(y \,|\, x) - a(x)\big)\big(\mathbb{E}(y \,|\, x) - \mathbb{E}(y \,|\, x)\big)\Big) = \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
R(a) = \mathbb{E}_{x,y} L(y, a(x)) = \\
= \mathbb{E}_{x,y}(y - \mathbb{E}(y \,|\, x))^2 + \mathbb{E}_{x,y}(\mathbb{E}(y \,|\, x) - a(x))^2 + \\
+ 2\mathbb{E}_{x,y}\big(y - \mathbb{E}(y \,|\, x)\big)\big(\mathbb{E}(y \,|\, x) - a(x)\big).
\end{aligned}
$$

Focus on the last term:

0

$$
\begin{aligned}
& \mathbb{E}_x \mathbb{E}_y \Big[\big(y - \mathbb{E}(y \,|\, x)\big)\big(\mathbb{E}(y \,|\, x) - a(x)\big) \,|\, x \Big] = \\
& = \mathbb{E}_x \Big(\big(\mathbb{E}(y \,|\, x) - a(x)\big) \mathbb{E}_y \Big[\big(y - \mathbb{E}(y \,|\, x)\big) \,|\, x \Big] \Big) = \\
& = \mathbb{E}_x \Big(\big(\mathbb{E}(y \,|\, x) - a(x)\big)\big(\mathbb{E}(y \,|\, x) - \mathbb{E}(y \,|\, x)\big) \Big) = \\
& = 0
\end{aligned}
$$

Does not depend on a(x)

So the risk takes form:

$$R(a) = \boxed{\mathbb{E}_{x,y}(y - \mathbb{E}(y \,|\, x))^2} + \mathbb{E}_{x,y}(\mathbb{E}(y \,|\, x) - a(x))^2.$$

The minimum is reached when $a(x) = \mathbb{E}(y \,|\, x).$

So the optimal regression model with square loss is

$$a_*(x) = \mathbb{E}(y \,|\, x) = \int_{\mathbb{Y}} yp(y \,|\, x)dy.$$

Denote $\mu : (\mathbb{X} \times \mathbb{Y})^\ell \to \mathcal{A}$, where $\mathcal{A}$ is some family of algorithms.

Denote $\mu : (\mathbb{X} \times \mathbb{Y})^\ell \to \mathcal{A}$, where $\mathcal{A}$ is some family of algorithms.

So $L(\mu) = \mathbb{E}_X \left[\mathbb{E}_{x,y} \left[\left(y - \mu(X)(x) \right)^2 \right] \right]$, where X dataset.

Denote $\mu : (\mathbb{X} \times \mathbb{Y})^\ell \to \mathcal{A}$, where $\mathcal{A}$ is some family of algorithms.

So $L(\mu) = \mathbb{E}_X \left[\mathbb{E}_{x,y} \left[\left(y - \mu(X)(x) \right)^2 \right] \right]$, where X dataset.

If X is fixed, then

$$\mathbb{E}_{x,y} \left[\left(y - \mu(X) \right)^2 \right] = \mathbb{E}_{x,y} \left[\left(y - \mathbb{E}[y \mid x] \right)^2 \right] + \mathbb{E}_{x,y} \left[\left(\mathbb{E}[y \mid x] - \mu(X) \right)^2 \right].$$

Denote $\mu : (\mathbb{X} \times \mathbb{Y})^\ell \to \mathcal{A}$, where $\mathcal{A}$ is some family of algorithms.

So $L(\mu) = \mathbb{E}_X\left[\mathbb{E}_{x,y}\left[\left(y - \mu(X)(x)\right)^2\right]\right]$, where X dataset.

If X is fixed, then

$$\mathbb{E}_{x,y}\left[\left(y - \mu(X)\right)^2\right] = \mathbb{E}_{x,y}\left[\left(y - \mathbb{E}[y \mid x]\right)^2\right] + \mathbb{E}_{x,y}\left[\left(\mathbb{E}[y \mid x] - \mu(X)\right)^2\right].$$

Let’s combine the latter equations:

Denote $\mu : (\mathbb{X} \times \mathbb{Y})^\ell \to \mathcal{A}$, where $\mathcal{A}$ is some family of algorithms.

So $L(\mu) = \mathbb{E}_X \left[\mathbb{E}_{x,y} \left[\left(y - \mu(X)(x) \right)^2 \right] \right]$, where X dataset.

If X is fixed, then

$$\mathbb{E}_{x,y} \left[\left(y - \mu(X) \right)^2 \right] = \mathbb{E}_{x,y} \left[\left(y - \mathbb{E}[y \mid x] \right)^2 \right] + \mathbb{E}_{x,y} \left[\left(\mathbb{E}[y \mid x] - \mu(X) \right)^2 \right].$$

Let's combine the latter equations:

$$L(\mu) = \mathbb{E}_X \left[\underbrace{\mathbb{E}_{x,y} \left[\left(y - \mathbb{E}[y \mid x] \right)^2 \right]}_{\text{Does not depend on X}} + \mathbb{E}_{x,y} \left[\left(\mathbb{E}[y \mid x] - \mu(X) \right)^2 \right] \right]$$

$$L(\mu) = \mathbb{E}_X \Bigg[\underbrace{\mathbb{E}_{x,y}\Big[\big(y - \mathbb{E}[y \mid x]\big)^2\Big]}_{\text{Does not depend on X}} + \mathbb{E}_{x,y}\Big[\big(\mathbb{E}[y \mid x] - \mu(X)\big)^2\Big] \Bigg] =$$

$$L(\mu) = \mathbb{E}_X \Big[\underbrace{\mathbb{E}_{x,y}\Big[\big(y - \mathbb{E}[y \,|\, x]\big)^2 \Big]}_{\text{Does not depend on X}} + \mathbb{E}_{x,y}\Big[\big(\mathbb{E}[y \,|\, x] - \mu(X)\big)^2 \Big] \Big] =$$

$$= \mathbb{E}_{x,y}\Big[\big(y - \mathbb{E}[y \,|\, x]\big)^2 \Big] + \mathbb{E}_{x,y}\Big[\mathbb{E}_X \Big[\big(\mathbb{E}[y \,|\, x] - \mu(X)\big)^2 \Big] \Big].$$

$$L(\mu) = \mathbb{E}_X\left[\mathbb{E}_{x,y}\left[\left(y - \mathbb{E}[y \mid x]\right)^2\right] + \mathbb{E}_{x,y}\left[\left(\mathbb{E}[y \mid x] - \mu(X)\right)^2\right]\right] =$$

$$= \mathbb{E}_{x,y}\left[\left(y - \mathbb{E}[y \mid x]\right)^2\right] + \boxed{\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x] - \mu(X)\right)^2\right]\right]}.$$

Focus on the second term:

$$L(\mu) = \mathbb{E}_X\left[\mathbb{E}_{x,y}\left[\left(y - \mathbb{E}[y \mid x]\right)^2\right] + \mathbb{E}_{x,y}\left[\left(\mathbb{E}[y \mid x] - \mu(X)\right)^2\right]\right] =$$

$$= \mathbb{E}_{x,y}\left[\left(y - \mathbb{E}[y \mid x]\right)^2\right] + \boxed{\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x] - \mu(X)\right)^2\right]\right]}.$$

Focus on the second term:

$$\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x] - \mu(X)\right)^2\right]\right] =$$

$$L(\mu) = \mathbb{E}_X\left[\mathbb{E}_{x,y}\left[\left(y - \mathbb{E}[y \mid x]\right)^2\right] + \mathbb{E}_{x,y}\left[\left(\mathbb{E}[y \mid x] - \mu(X)\right)^2\right]\right] =$$

$$= \mathbb{E}_{x,y}\left[\left(y - \mathbb{E}[y \mid x]\right)^2\right] + \mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x] - \mu(X)\right)^2\right]\right].$$

Focus on the second term:

$$\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x] - \mu(X)\right)^2\right]\right] =$$
$$= \mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x] - \mathbb{E}_X\left[\mu(X)\right] + \mathbb{E}_X\left[\mu(X)\right] - \mu(X)\right)^2\right]\right]$$

$$\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y\,|\,x]-\mu(X)\right)^2\right]\right]=$$

$$=\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y\,|\,x]-\mathbb{E}_X\left[\mu(X)\right]+\mathbb{E}_X\left[\mu(X)\right]-\mu(X)\right)^2\right]\right]=$$

$$
\begin{aligned}
&\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x] - \mu(X)\right)^2\right]\right] = \\
&= \mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x] - \mathbb{E}_X\left[\mu(X)\right] + \mathbb{E}_X\left[\mu(X)\right] - \mu(X)\right)^2\right]\right] = \\
&= \mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\underbrace{\left(\mathbb{E}[y \mid x] - \mathbb{E}_X\left[\mu(X)\right]\right)^2}\right]\right] + \mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}_X\left[\mu(X)\right] - \mu(X)\right)^2\right]\right] + \\
&\qquad + 2\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x] - \mathbb{E}_X\left[\mu(X)\right]\right)\left(\mathbb{E}_X\left[\mu(X)\right] - \mu(X)\right)\right]\right].
\end{aligned}
$$

$$\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x] - \mu(X)\right)^2\right]\right] =$$

$$= \mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x] - \mathbb{E}_X\left[\mu(X)\right] + \mathbb{E}_X\left[\mu(X)\right] - \mu(X)\right)^2\right]\right] =$$

$$= \mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\underbrace{\left(\mathbb{E}[y \mid x] - \mathbb{E}_X\left[\mu(X)\right]\right)^2}_{\text{Does not depend on X}}\right]\right] + \mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}_X\left[\mu(X)\right] - \mu(X)\right)^2\right]\right] +$$

$$+ 2\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x] - \mathbb{E}_X\left[\mu(X)\right]\right)\left(\mathbb{E}_X\left[\mu(X)\right] - \mu(X)\right)\right]\right].$$

$$
\begin{aligned}
&\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x]-\mu(X)\right)^2\right]\right]= \\
&=\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x]-\mathbb{E}_X\left[\mu(X)\right]+\mathbb{E}_X\left[\mu(X)\right]-\mu(X)\right)^2\right]\right]= \\
&=\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\underbrace{\left(\mathbb{E}[y \mid x]-\mathbb{E}_X\left[\mu(X)\right]\right)^2}_{\text{Does not depend on X}}\right]\right]+\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}_X\left[\mu(X)\right]-\mu(X)\right)^2\right]\right]+ \\
&\qquad +2\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x]-\mathbb{E}_X\left[\mu(X)\right]\right)\left(\mathbb{E}_X\left[\mu(X)\right]-\mu(X)\right)\right]\right].
\end{aligned}
$$

Just a bit further, we are almost there

$$\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y\,|\,x] - \mu(X)\right)^2\right]\right] =$$

$$= \mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y\,|\,x] - \mathbb{E}_X\left[\mu(X)\right] + \mathbb{E}_X\left[\mu(X)\right] - \mu(X)\right)^2\right]\right] =$$

$$= \mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y\,|\,x] - \mathbb{E}_X\left[\mu(X)\right]\right)^2\right]\right] + \mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}_X\left[\mu(X)\right] - \mu(X)\right)^2\right]\right] +$$

$$\boxed{+\, 2\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y\,|\,x] - \mathbb{E}_X\left[\mu(X)\right]\right)\left(\mathbb{E}_X\left[\mu(X)\right] - \mu(X)\right)\right]\right]}.$$

Focus on this term

$$\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x] - \mathbb{E}_X\left[\mu(X)\right]\right)\left(\mathbb{E}_X\left[\mu(X)\right] - \mu(X)\right)\right] =$$

$$
\begin{aligned}
&\mathbb{E}_X\Big[\big(\mathbb{E}[y\,|\,x]-\mathbb{E}_X\left[\mu(X)\right]\big)\big(\mathbb{E}_X\left[\mu(X)\right]-\mu(X)\big)\Big] = \\
&\quad = \big(\mathbb{E}[y\,|\,x]-\mathbb{E}_X\left[\mu(X)\right]\big)\mathbb{E}_X\Big[\mathbb{E}_X\left[\mu(X)\right]-\mu(X)\Big] =
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_X\Big[\big(\mathbb{E}[y \mid x] - \mathbb{E}_X[\mu(X)]\big)\big(\mathbb{E}_X[\mu(X)] - \mu(X)\big)\Big] &= \\
&= \big(\mathbb{E}[y \mid x] - \mathbb{E}_X[\mu(X)]\big)\mathbb{E}_X\Big[\mathbb{E}_X[\mu(X)] - \mu(X)\Big] = \\
&= \big(\mathbb{E}[y \mid x] - \mathbb{E}_X[\mu(X)]\big)\Big[\mathbb{E}_X[\mu(X)] - \mathbb{E}_X[\mu(X)]\Big] =
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_X\Big[\big(\mathbb{E}[y \mid x] - \mathbb{E}_X[\mu(X)]\big)\big(\mathbb{E}_X[\mu(X)] - \mu(X)\big)\Big] &= \\
&= \big(\mathbb{E}[y \mid x] - \mathbb{E}_X[\mu(X)]\big)\mathbb{E}_X\Big[\mathbb{E}_X[\mu(X)] - \mu(X)\Big] = \\
&= \big(\mathbb{E}[y \mid x] - \mathbb{E}_X[\mu(X)]\big)\Big[\mathbb{E}_X[\mu(X)] - \mathbb{E}_X[\mu(X)]\Big] = \\
&= 0.
\end{aligned}
$$

$$
\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x] - \mu(X)\right)^2\right]\right] =
$$

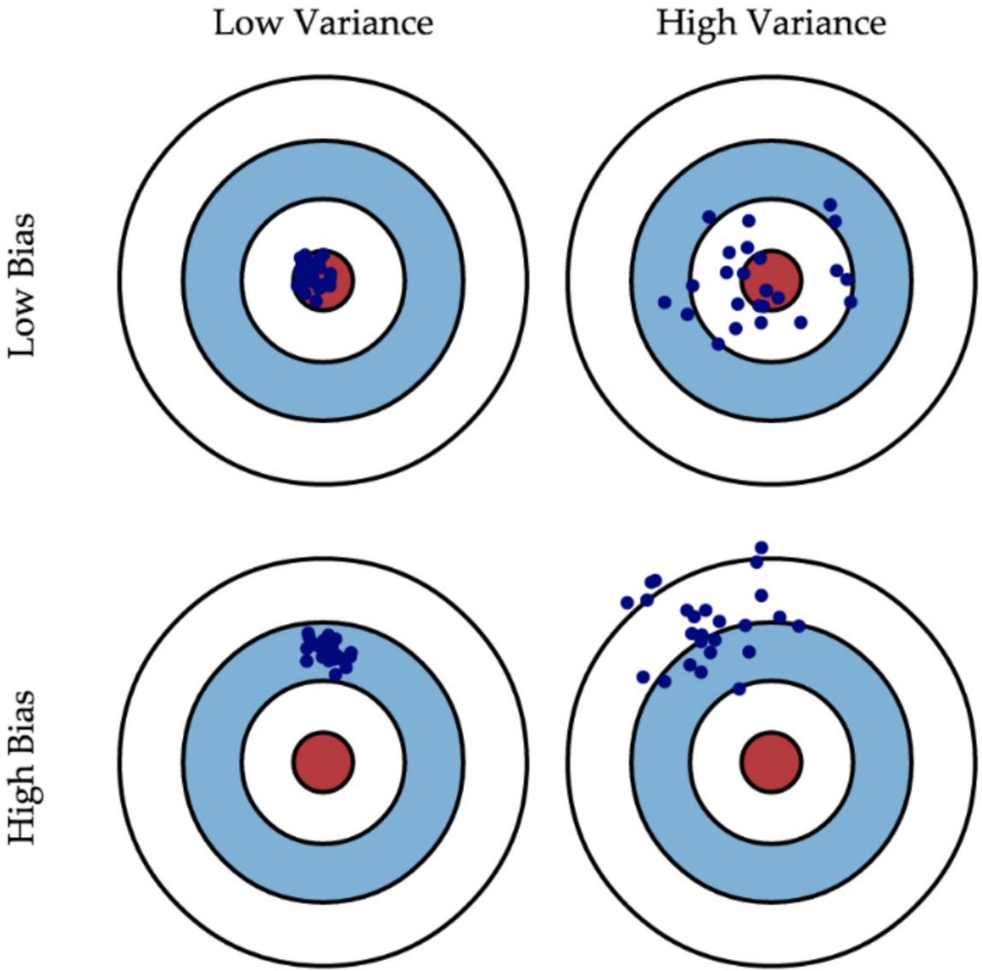
$$
= \mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x] - \mathbb{E}_X\left[\mu(X)\right] + \mathbb{E}_X\left[\mu(X)\right] - \mu(X)\right)^2\right]\right] =
$$

$$
= \mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x] - \mathbb{E}_X\left[\mu(X)\right]\right)^2\right]\right] + \mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}_X\left[\mu(X)\right] - \mu(X)\right)^2\right]\right] +
$$

$$
\cancel{+ 2\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x] - \mathbb{E}_X\left[\mu(X)\right]\right)\left(\mathbb{E}_X\left[\mu(X)\right] - \mu(X)\right)\right]\right].} \rightarrow 0
$$

$$L(\mu) = \underbrace{\mathbb{E}_{x,y}\left[\left(y - \mathbb{E}[y\,|\,x]\right)^2\right]}_{\text{noise}} +$$

$$+ \underbrace{\mathbb{E}_x\left[\left(\mathbb{E}_X\left[\mu(X)\right] - \mathbb{E}[y\,|\,x]\right)^2\right]}_{\text{bias}} + \underbrace{\mathbb{E}_x\left[\mathbb{E}_X\left[\left(\mu(X) - \mathbb{E}_X\left[\mu(X)\right]\right)^2\right]\right]}_{\text{variance}}.$$


Low Variance
High Variance
Low Bias
High Bias

$$L(\mu) = \underbrace{\mathbb{E}_{x,y}\left[\left(y - \mathbb{E}[y \mid x]\right)^2\right]}_{\text{noise}} +$$

$$+ \underbrace{\mathbb{E}_x\left[\left(\mathbb{E}_X\left[\mu(X)\right] - \mathbb{E}[y \mid x]\right)^2\right]}_{\text{bias}} + \underbrace{\mathbb{E}_x\left[\mathbb{E}_X\left[\left(\mu(X) - \mathbb{E}_X\left[\mu(X)\right]\right)^2\right]\right]}_{\text{variance}}.$$

This exact form of bias-variance decomposition is correct for square loss in regression.

However, it is much more general. See extra materials for more exotic cases.

Bagging = Bootstrap aggregating

Denote dataset $\tilde{X}$ bootstrapped from X.

Denote μ: $\tilde{\mu}(X) = \mu(\tilde{X})$. Let $b_n(x)$ be basic algorithm.

Denote the ensemble:

$$a_N(x) = \frac{1}{N}\sum_{n=1}^{N} b_n(x) = \frac{1}{N}\sum_{n=1}^{N} \tilde{\mu}(X)(x).$$

$$L(\mu) = \underbrace{\mathbb{E}_{x,y}\left[\left(y - \mathbb{E}[y \mid x]\right)^2\right]}_{\text{noise}} +$$

$$+ \underbrace{\mathbb{E}_x\left[\left(\mathbb{E}_X\left[\mu(X)\right] - \mathbb{E}[y \mid x]\right)^2\right]}_{\text{bias}} + \underbrace{\mathbb{E}_x\left[\mathbb{E}_X\left[\left(\mu(X) - \mathbb{E}_X\left[\mu(X)\right]\right)^2\right]\right]}_{\text{variance}}.$$

The bias term takes the following form:

$$\mathbb{E}_{x,y}\left[\left(\mathbb{E}_X\left[\frac{1}{N}\sum_{n=1}^{N}\tilde{\mu}(X)(x)\right] - \mathbb{E}[y \mid x]\right)^2\right] =$$

The bias term takes the following form:

$$\mathbb{E}_{x,y}\left[\left(\mathbb{E}_X\left[\frac{1}{N}\sum_{n=1}^{N}\tilde{\mu}(X)(x)\right] - \mathbb{E}[y \mid x]\right)^2\right] =$$

$$= \mathbb{E}_{x,y}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_X[\tilde{\mu}(X)(x)] - \mathbb{E}[y \mid x]\right)^2\right] =$$

The bias term takes the following form:

$$\mathbb{E}_{x,y}\left[\left(\mathbb{E}_X\left[\frac{1}{N}\sum_{n=1}^{N}\tilde{\mu}(X)(x)\right] - \mathbb{E}[y \mid x]\right)^2\right] =$$

$$= \mathbb{E}_{x,y}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_X[\tilde{\mu}(X)(x)] - \mathbb{E}[y \mid x]\right)^2\right] =$$

$$= \mathbb{E}_{x,y}\left[\left(\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right] - \mathbb{E}[y \mid x]\right)^2\right].$$

The bias term takes the following form:

$$\mathbb{E}_{x,y}\left[\left(\mathbb{E}_X\left[\frac{1}{N}\sum_{n=1}^{N}\tilde{\mu}(X)(x)\right]-\mathbb{E}[y\,|\,x]\right)^2\right]=$$

$$=\mathbb{E}_{x,y}\left[\left(\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_X[\tilde{\mu}(X)(x)]-\mathbb{E}[y\,|\,x]\right)^2\right]=$$

$$=\boxed{\mathbb{E}_{x,y}\left[\left(\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]-\mathbb{E}[y\,|\,x]\right)^2\right]}.$$

One algorithm bias

The variance: $$\mathbb{E}_{x,y}\Big[\mathbb{E}_X\Big[\Big(\frac{1}{N}\sum_{n=1}^{N}\tilde{\mu}(X)(x)-\mathbb{E}_X\Big[\frac{1}{N}\sum_{n=1}^{N}\tilde{\mu}(X)(x)\Big]\Big)^2\Big]\Big].$$

$$\Big(\frac{1}{N}\sum_{n=1}^{N}\tilde{\mu}(X)(x)-\mathbb{E}_X\Big[\frac{1}{N}\sum_{n=1}^{N}\tilde{\mu}(X)(x)\Big]\Big)^2=$$

$$=\frac{1}{N^2}\Big(\sum_{n=1}^{N}\Big[\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big]\Big)^2=$$

$$=\frac{1}{N^2}\sum_{n=1}^{N}\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)^2+$$

$$+\frac{1}{N^2}\sum_{n_1\neq n_2}\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)$$

The variance:

$$\mathbb{E}_{x,y}\Big[\mathbb{E}_X\Big[\frac{1}{N^2}\sum_{n=1}^{N}\Big(\tilde{\mu}(X)(x) - \mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)^2 +$$

$$+\frac{1}{N^2}\sum_{n_1\neq n_2}\Big(\tilde{\mu}(X)(x) - \mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)\Big(\tilde{\mu}(X)(x) - \mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)\Big]\Big] =$$

$$= \frac{1}{N^2}\mathbb{E}_{x,y}\Big[\mathbb{E}_X\Big[\sum_{n=1}^{N}\Big(\tilde{\mu}(X)(x) - \mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)^2\Big]\Big]+$$

$$+\frac{1}{N^2}\mathbb{E}_{x,y}\Big[\mathbb{E}_X\Big[\sum_{n_1\neq n_2}\Big(\tilde{\mu}(X)(x) - \mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)\times$$

$$\times\Big(\tilde{\mu}(X)(x) - \mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)\Big]\Big] =$$

One algorithm variance * 1/N

$$= \boxed{\frac{1}{N}\mathbb{E}_{x,y}\Big[\mathbb{E}_X\Big[\Big(\tilde{\mu}(X)(x) - \mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)^2\Big]\Big]}+$$

$$+\frac{N(N-1)}{N^2}\mathbb{E}_{x,y}\Big[\mathbb{E}_X\Big[\Big(\tilde{\mu}(X)(x) - \mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)\times$$

$$\times\Big(\tilde{\mu}(X)(x) - \mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)\Big]\Big]$$

The variance:

$$\mathbb{E}_{x,y}\Bigg[\mathbb{E}_X\Bigg[\frac{1}{N^2}\sum_{n=1}^{N}\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)^2+$$

$$+\frac{1}{N^2}\sum_{n_1\neq n_2}\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)\Bigg]\Bigg]=$$

$$=\frac{1}{N^2}\mathbb{E}_{x,y}\Bigg[\mathbb{E}_X\Bigg[\sum_{n=1}^{N}\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)^2\Bigg]\Bigg]+$$

$$+\frac{1}{N^2}\mathbb{E}_{x,y}\Bigg[\mathbb{E}_X\Bigg[\sum_{n_1\neq n_2}\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)\times$$

$$\times\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)\Bigg]\Bigg]=$$

One algorithm variance * 1/N

$$=\boxed{\frac{1}{N}\mathbb{E}_{x,y}\Bigg[\mathbb{E}_X\Bigg[\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)^2\Bigg]\Bigg]}+$$

$$\boxed{+\frac{N(N-1)}{N^2}\mathbb{E}_{x,y}\Bigg[\mathbb{E}_X\Bigg[\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)\times\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)\Bigg]\Bigg]}$$

Basic algorithms covariance

Feature importance estimation

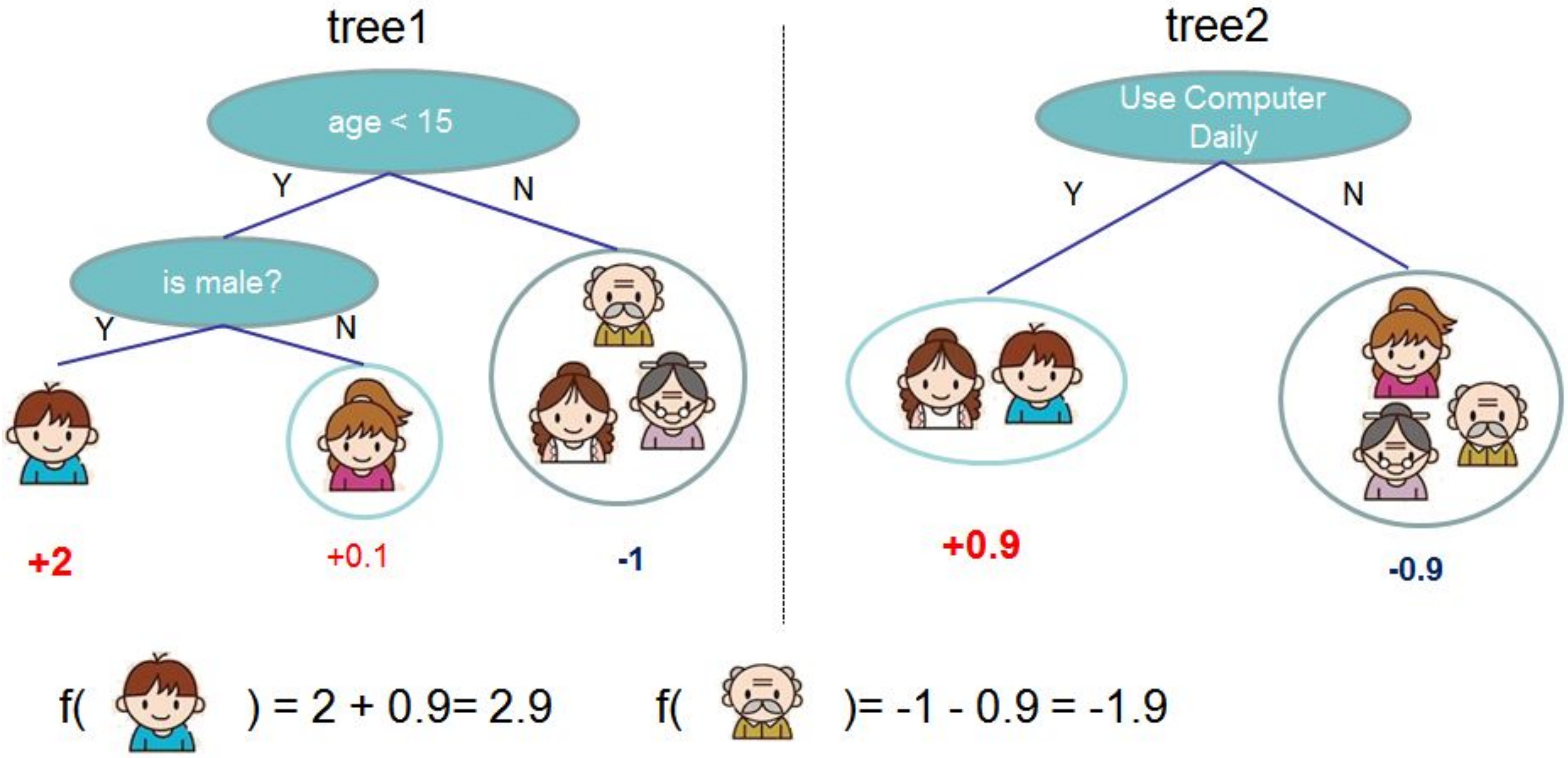


Image source: https://habr.com/ru/post/428213/

Feature importance estimation

1. Permutation importance
2. Partial Dependence Plots (PDP)
3. Tree specific:
    a. Gain
    b. Frequency (Split Count)
    c. Cover (weighted Split Count)
4. Shap

Permutation importance

Height at age 20 (cm)	Height at age 10 (cm)	...	Socks owned at age 10
182	155	...	20
175	147	...	10
...	...	...	...
156	142	...	8
153	130	...	24

Image source: https://www.kaggle.com/dansbecker/permutation-importance

Permutation importance

Height at age 20 (cm)	Height at age 10 (cm)	...	Socks owned at age 10
182	155	...	20
175	147	...	10
...	...	...	...
156	142	...	8
153	130	...	24

Train model

Observe changes caused by feature random permutations

Image source: https://www.kaggle.com/dansbecker/permutation-importance

Partial Dependence Plots

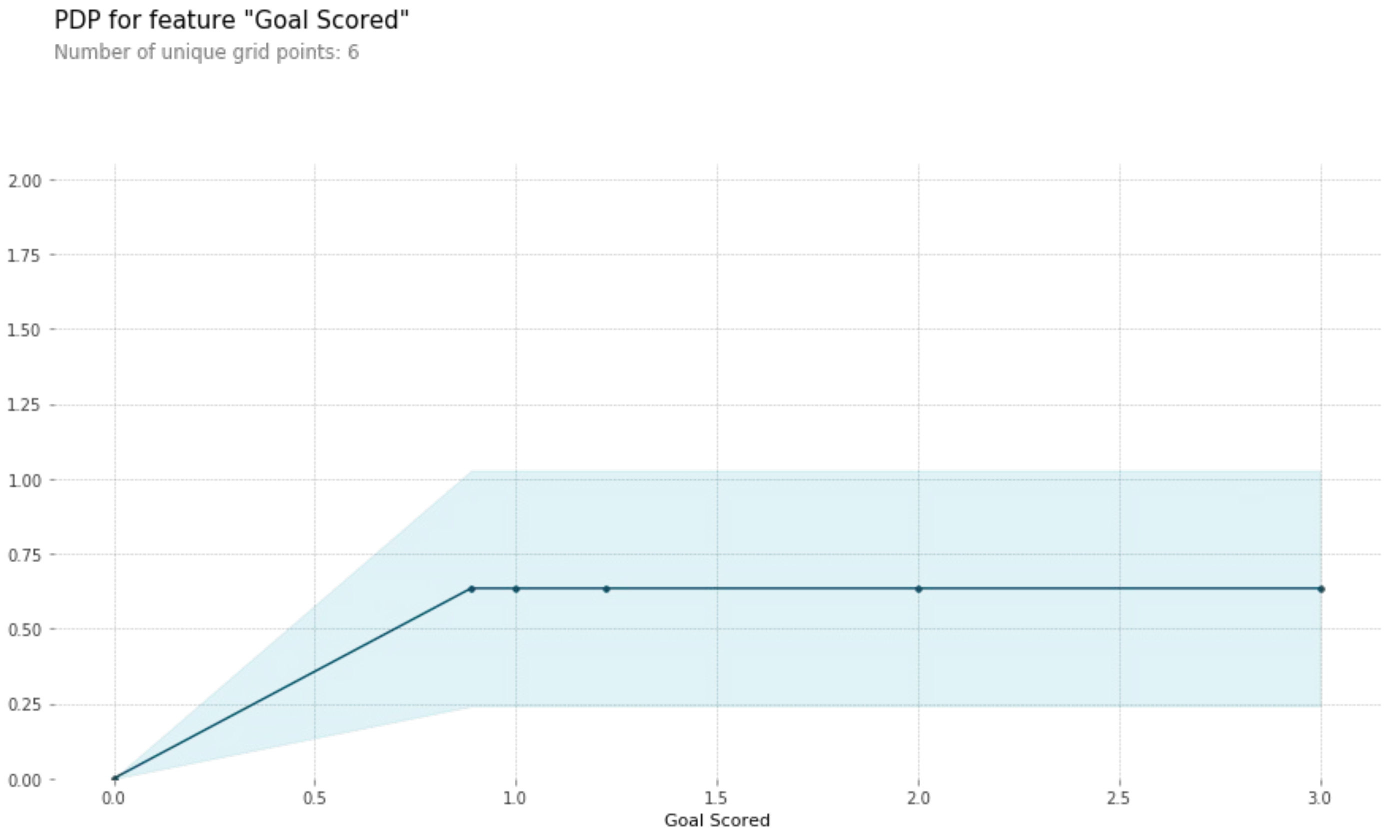


Image source: https://www.kaggle.com/dansbecker/partial-plots

Importance estimation problems

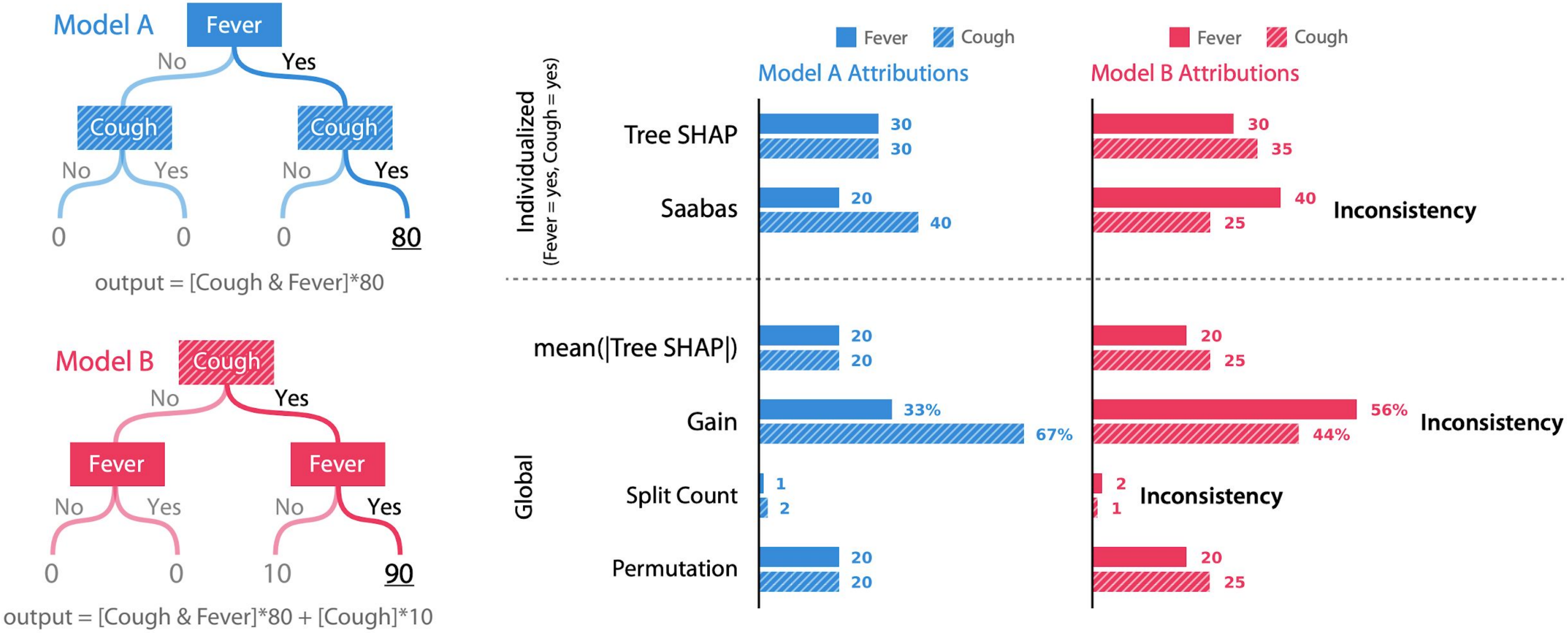


Image source: [“Consistent Individualized Feature Attribution for Tree Ensembles”] paper

Shap values

Consider *i-th* feature. Shap value will be

$$\phi_i(p) = \sum_{S \subseteq N/\{i\}} \frac{|S|!(n-|S|-1)!}{n!}(p(S \cup \{i\}) - p(S))$$

where $p(S \cup \{i\})$ is model prediction on feature subset S with *i-th* feature added.

Shap values

Consider *i-th* feature. Shap value will be

$$\phi_i(p) = \sum_{S \subseteq N/\{i\}} \frac{|S|!(n - |S| - 1)!}{n!}(p(S \cup \{i\}) - p(S))$$

where $p(S \cup \{i\})$ is model prediction on feature subset S with *i-th* feature added.

SHAP values are the only consistent and locally accurate individualized feature attributions

See Consistent Individualized Feature Attribution for Tree Ensembles paper for more info

Outro

1. Remember the bias-varience decomposition
2. Consider using SHAP values to estimate feature importances.

Great demo: http://arogozhnikov.github.io/2016/06/24/gradient_boosting_explained.html