Lecture 6: Ensembles

MIPT, 2019

Outline

1. Bias-variance decomposition.
2. Bagging.
3. RSM - Random Subspace Method
4. Random Forest.

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 $a_*(x) = \mathbb{E}[y \mid x] = \int_{\mathbb{Y}} y p(y \mid x) dy$

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

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

Bias-variance decomposition

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

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

Bias-variance decomposition

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

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

Let's return to the risk estimation:

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:

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

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:

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

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

Focus on the last term:

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

Focus on the last term:

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

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

Focus on the last term:

Does not depend on y

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

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

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

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

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

Focus on the last term:

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

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

Focus on the last term:

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

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

Focus on the last term:

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

0

$$\mathbb{E}_x \mathbb{E}_y \Big[\big(y - \mathbb{E}(y \mid x)\big)\big(\mathbb{E}(y \mid x) - a(x)\big) \mid x \Big] =$$
$$= \mathbb{E}_x \Big(\big(\mathbb{E}(y \mid x) - a(x)\big) \mathbb{E}_y \Big[\big(y - \mathbb{E}(y \mid x)\big) \mid x \Big] \Big) =$$
$$= \mathbb{E}_x \Big(\big(\mathbb{E}(y \mid x) - a(x)\big)\big(\mathbb{E}(y \mid x) - \mathbb{E}(y \mid x)\big) \Big) =$$
$$= 0$$

Does not depend on a(x)

So the risk takes form:

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

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

So the optimal regression model with square loss is

$$a_*(x) = \mathbb{E}(y \mid x) = \int_{\mathbb{Y}} y p(y \mid x) dy.$$

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:

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

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

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

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

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

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

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:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

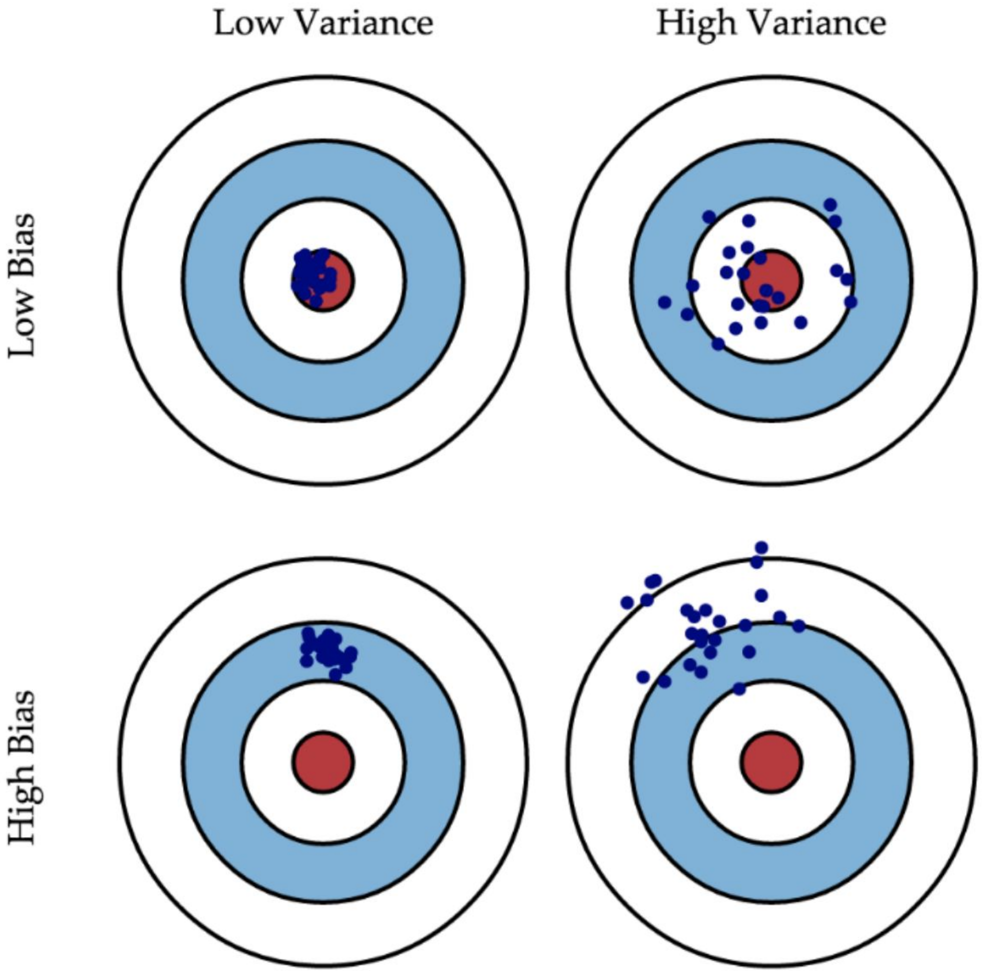
$$=\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x]-\mathbb{E}_X\left[\mu(X)\right]+\mathbb{E}_X\left[\mu(X)\right]-\mu(X)\right)^2\right]\right]=$$

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

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

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

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


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:

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

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

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

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

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

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

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

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

The variance:

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

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

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

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

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

One algorithm variance * 1/N

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

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

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

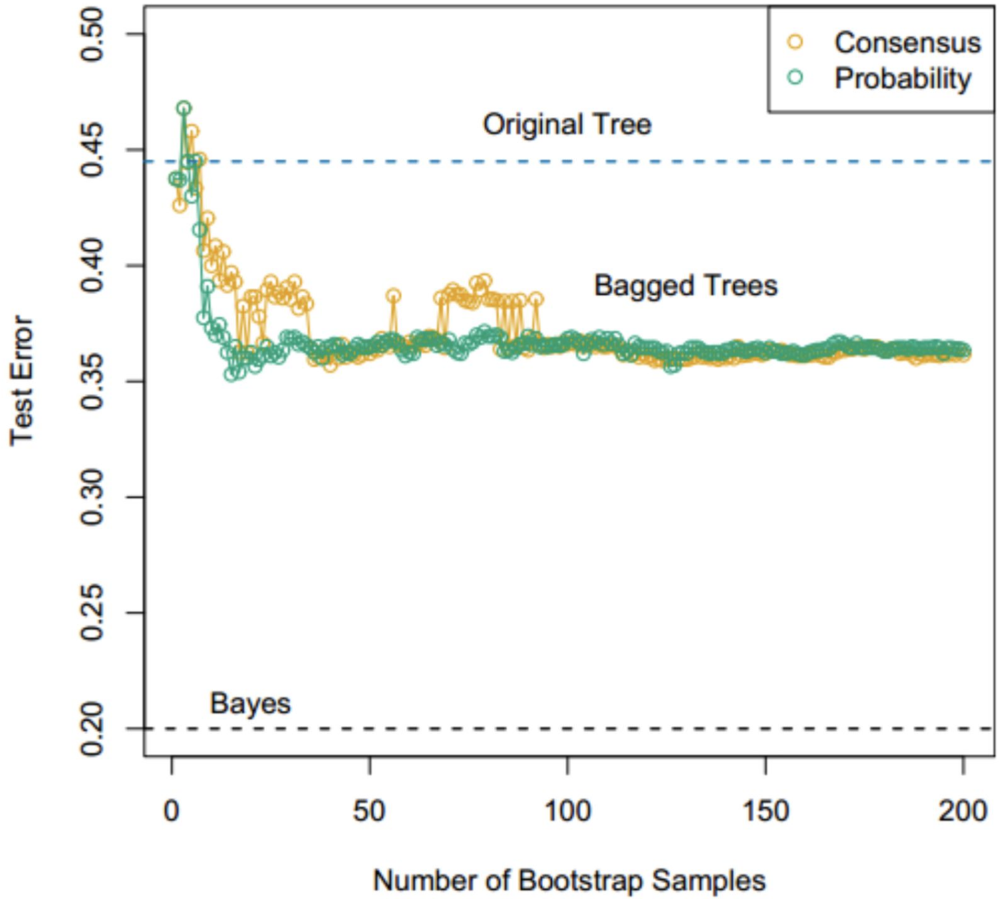
The variance:

$$
\begin{aligned}
&\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\frac{1}{N^2}\sum_{n=1}^{N}\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)^2+\right.\right.\\
&\quad\left.\left.+\frac{1}{N^2}\sum_{n_1\neq n_2}\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)\right]\right]=\\
&=\frac{1}{N^2}\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\sum_{n=1}^{N}\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)^2\right]\right]+\\
&\quad+\frac{1}{N^2}\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\sum_{n_1\neq n_2}\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)\times\right.\right.\\
&\qquad\qquad\left.\left.\times\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)\right]\right]=\\
&=\boxed{\frac{1}{N}\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)^2\right]\right]}+\\
&\quad\boxed{\begin{aligned}&+\frac{N(N-1)}{N^2}\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)\times\right.\right.\\
&\qquad\left.\left.\times\Big(\tilde{\mu}(X)(x)-\mathbb{E}_X\left[\tilde{\mu}(X)(x)\right]\Big)\right]\right]\end{aligned}}
\end{aligned}
$$

One algorithm variance * 1/N

Basic algorithms covariance

Bagging = Bootstrap aggregating



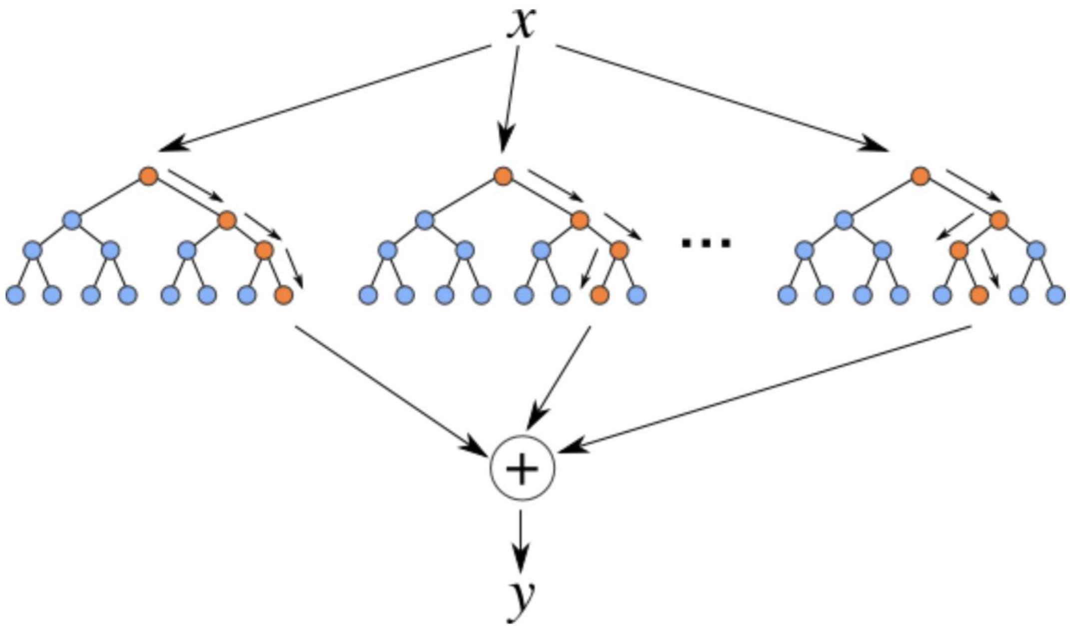
Bagging = Bootstrap aggregating

Decreases the variance if the basic algorithms are not correlated.

RSM - Random Subspace Method

Same approach, but with features.

Bagging + RSM = Random Forest



Random Forest

- One of the greatest “universal” models.

Random Forest

- One of the greatest “universal” models.
- There are some modifications: Extremely Randomized Trees, Isolation Forest, etc.
-

Random Forest

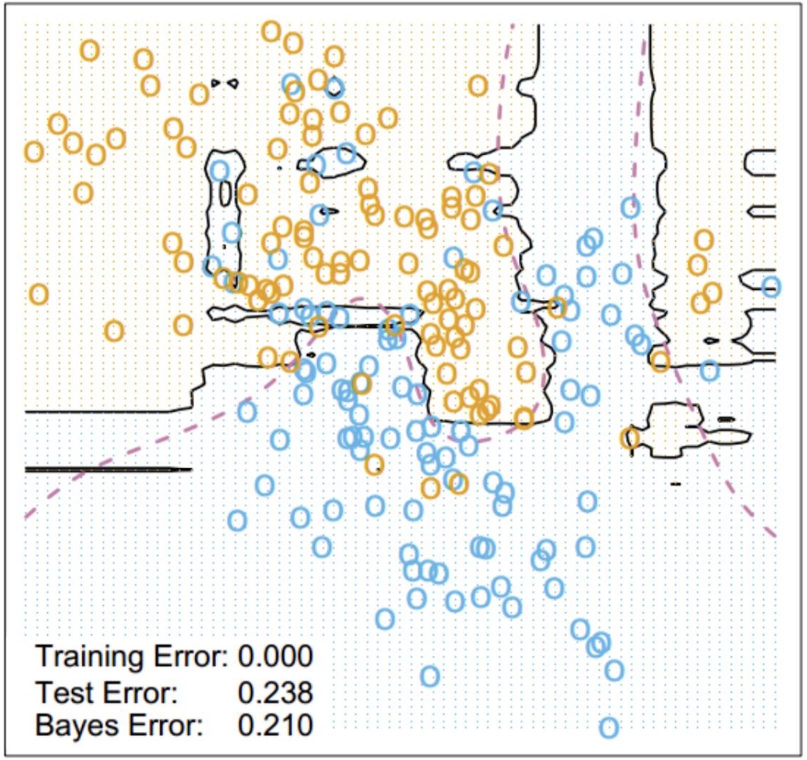
- One of the greatest “universal” models.
- There are some modifications: Extremely Randomized Trees, Isolation Forest, etc.
- Allows to use train data for validation: OOB

Random Forest

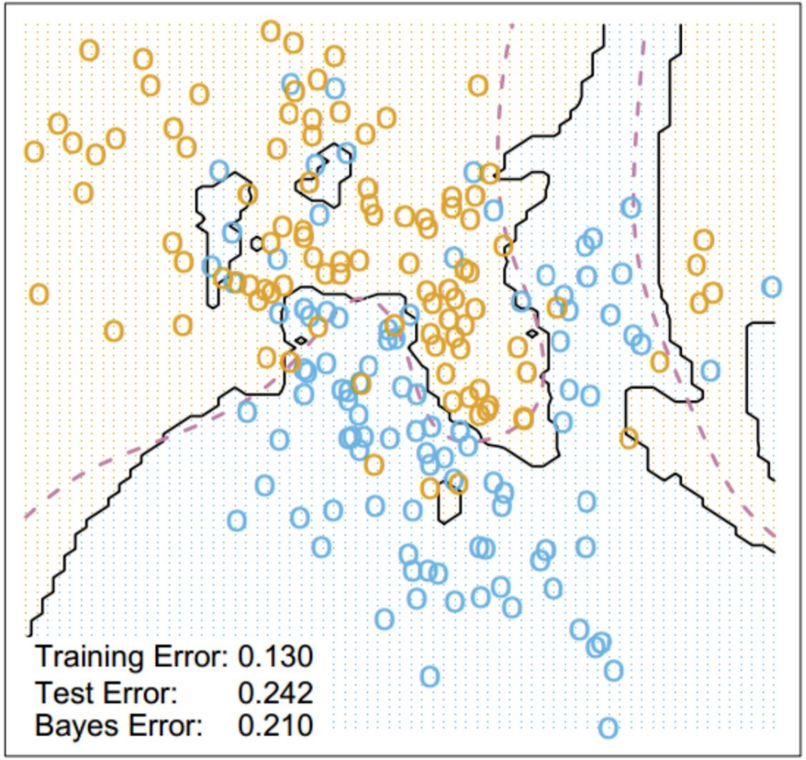
- One of the greatest “universal” models.
- There are some modifications: Extremely Randomized Trees, Isolation Forest, etc.
- Allows to use train data for validation: OOB

$$\mathrm{OOB} = \sum_{i=1}^{\ell} L\left(y_i, \frac{1}{\sum_{n=1}^{N}[x_i \notin X_n]} \sum_{n=1}^{N}[x_i \notin X_n] b_n(x_i)\right)$$

Random Forest Classifier



3−Nearest Neighbors



Boosting is coming next time.

Stay tuned.