Machine Learning course

Lecture 7: Bias-Variance tradeoff; Stacking, Blending

Radoslav Neychev

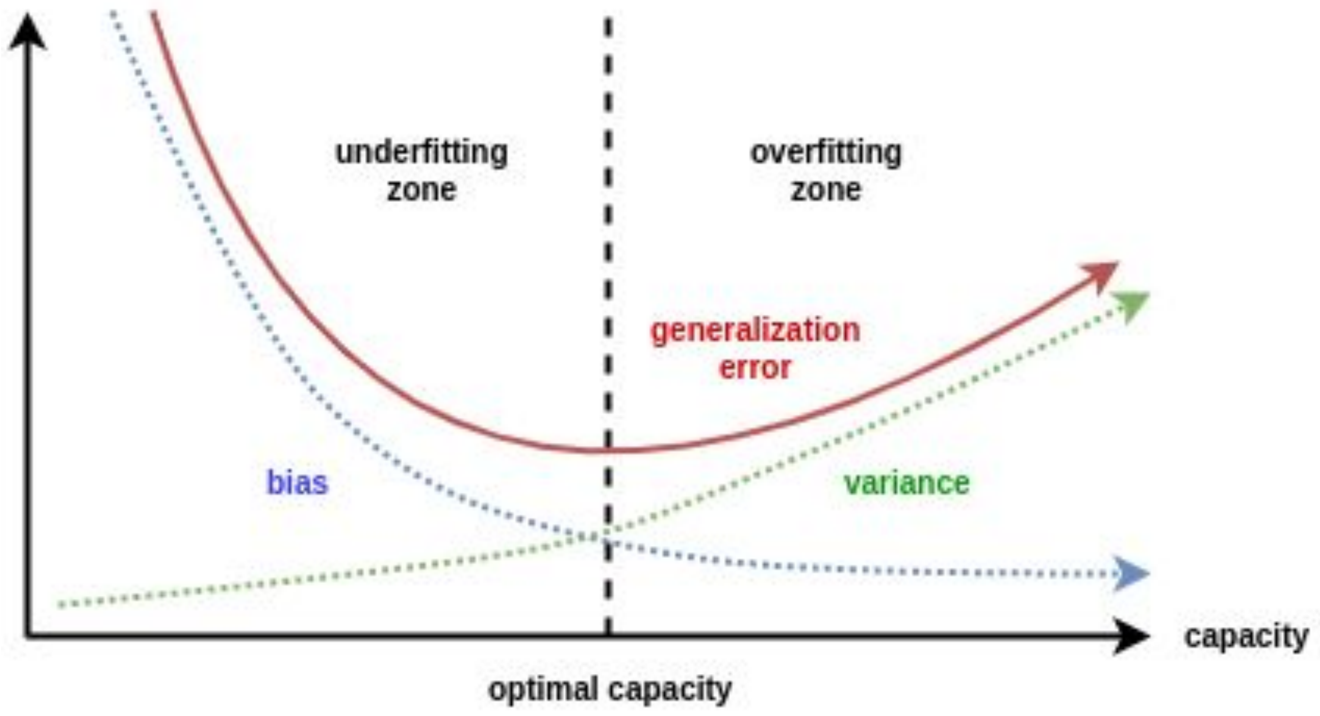
Spring 2021

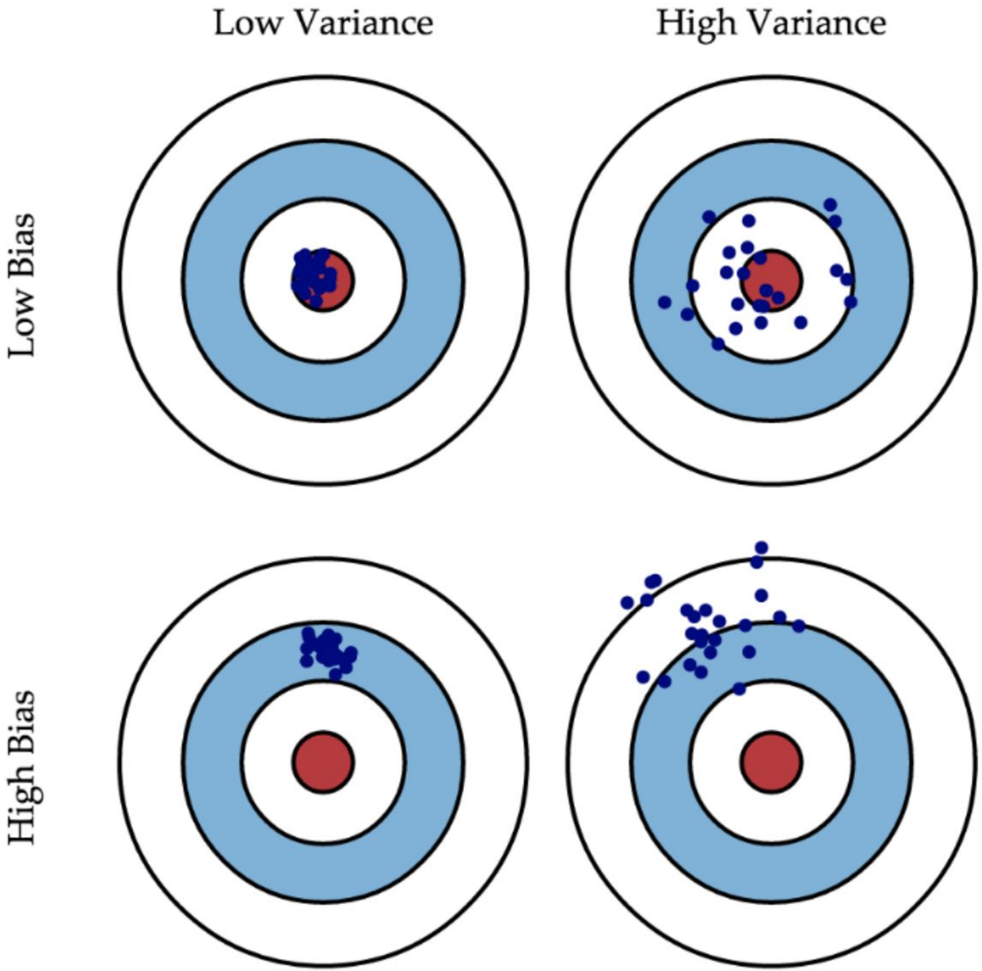
Outline

1. Bias-Variance Tradeoff
2. Blending
3. Stacking

Bias-Variance tradeoff

Bias-variance tradeoff




Low Variance
High Variance
Low Bias
High Bias

Bias-variance decomposition derivation

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 empirical risk takes form:

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

Bias-variance decomposition

Let’s show that

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

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

Returning to the risk estimation:

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

Focus on the last term:

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

Does not depend on y

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

Focus on the last term:

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

0

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

So the risk takes form:

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

Does not depend on a(x)

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

So the optimal regression model with square loss is

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

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.

In further slides (x) is omitted!

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.

In further slides (x) is omitted!

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.

In further slides (x) is omitted!

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.

In further slides (x) is omitted!

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:

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

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

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

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

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

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

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:

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

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

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

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

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

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

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

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

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

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

Just a bit further, we are almost there

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

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

Focus on this term

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

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

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

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

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

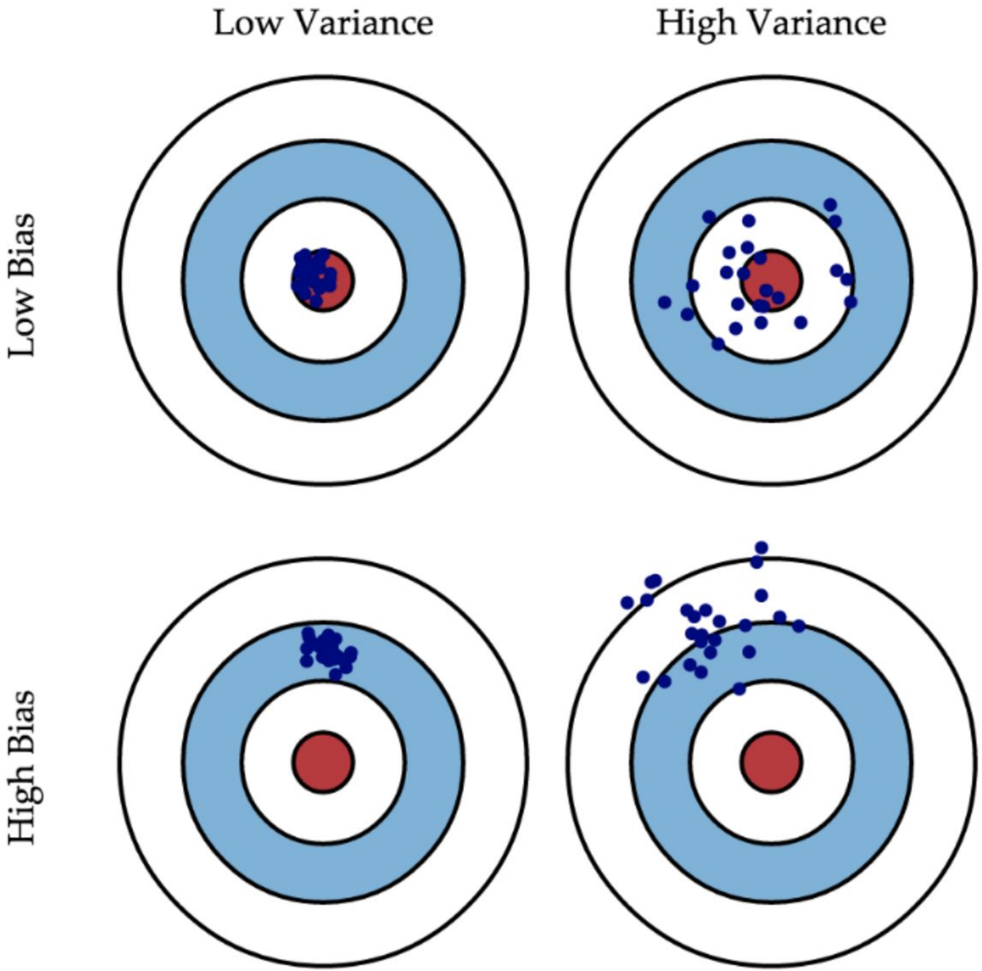
$$= \mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x] - \mathbb{E}_X[\mu(X)] + \mathbb{E}_X[\mu(X)] - \mu(X)\right)^2\right]\right] =$$

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

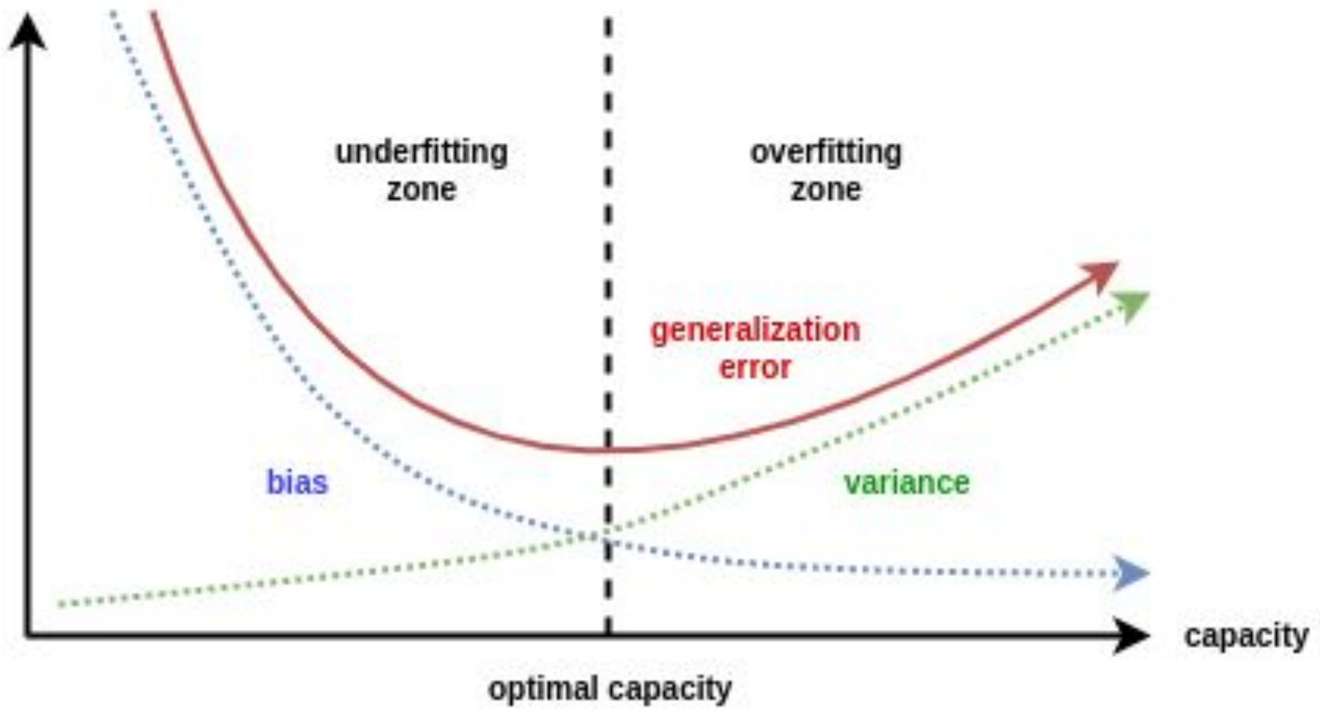
$$\cancel{+ 2\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[\left(\mathbb{E}[y \mid x] - \mathbb{E}_X[\mu(X)]\right)\left(\mathbb{E}_X[\mu(X)] - \mu(X)\right)\right]\right].} \longrightarrow 0$$

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

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


Low Variance
High Variance
Low Bias
High Bias

Bias-variance tradeoff



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

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

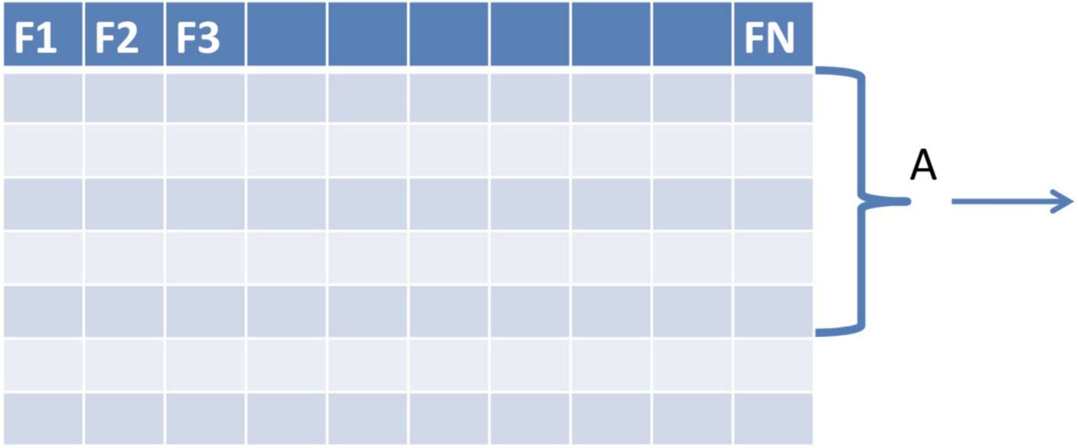
This exact form of bias-variance decomposition is correct for square loss in regression.

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

Stacking and blending

Blending

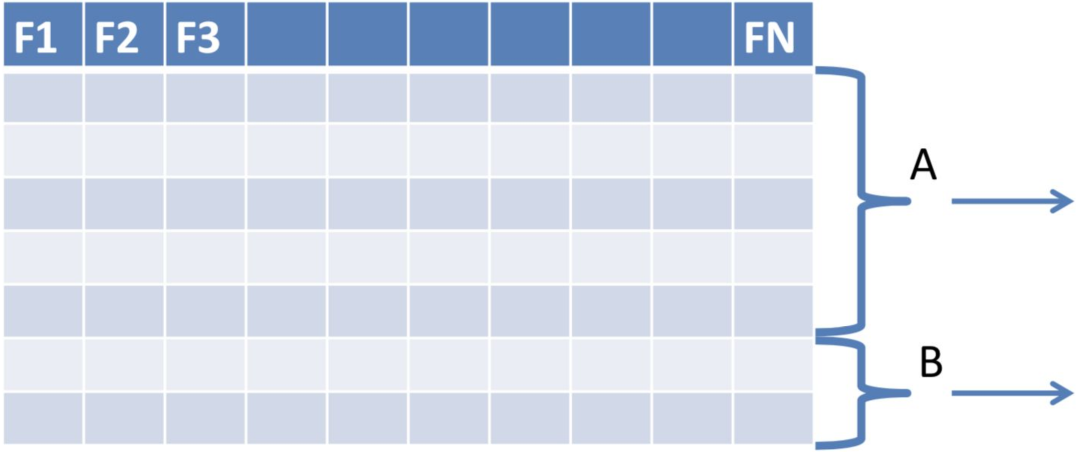
How to build an ensemble from *different* models?



Tune M basic algorithms

Blending

How to build an ensemble from *different* models?

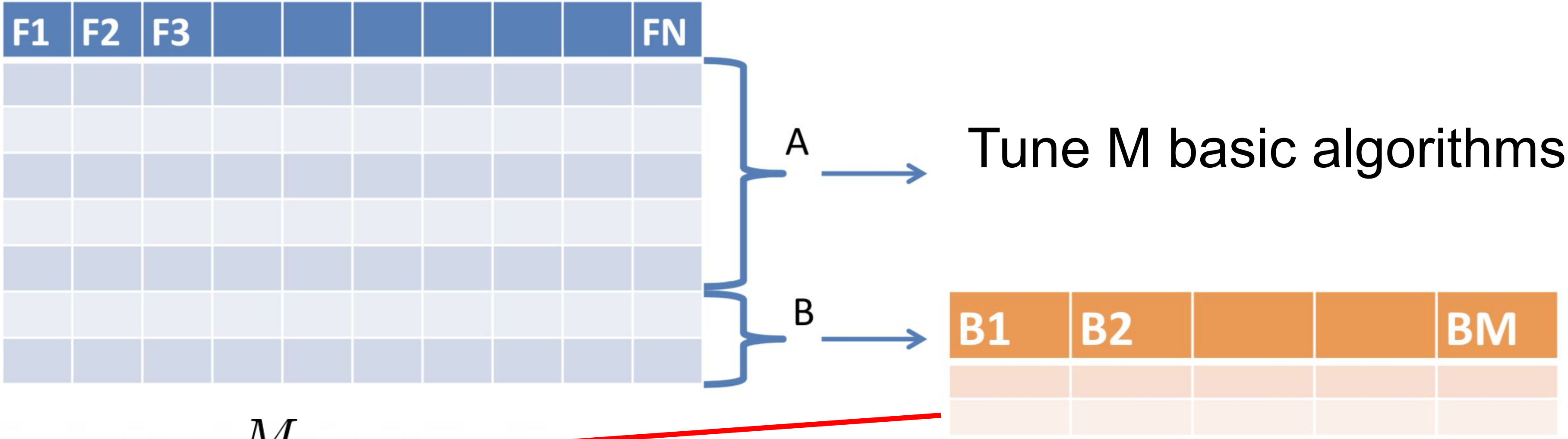


Tune M basic algorithms

Tune new model on the outputs of the basic algorithms

Blending

How to build an ensemble from *different* models?



$$\hat{f}(x) = \sum_{i=1}^{M} \rho_i f_i(x)$$

$$\sum_{i=1}^{M} \rho_i = 1, \quad \rho_i \in [0;1] \;\; \forall i$$

Blending

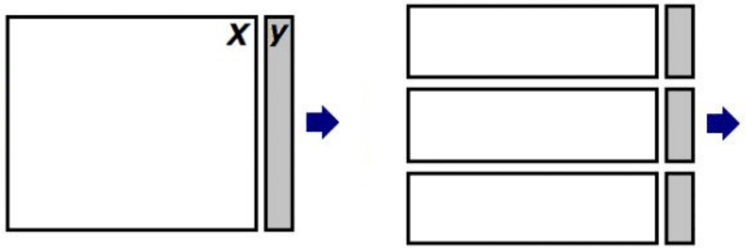
Just combine several *strong/complex* models.

$$\hat{f}(x) = \sum_{i=1}^{M} \rho_i f_i(x), \qquad \sum_{i=1}^{M} \rho_i = 1, \quad \rho_i \in [0;1] \ \ \forall i$$

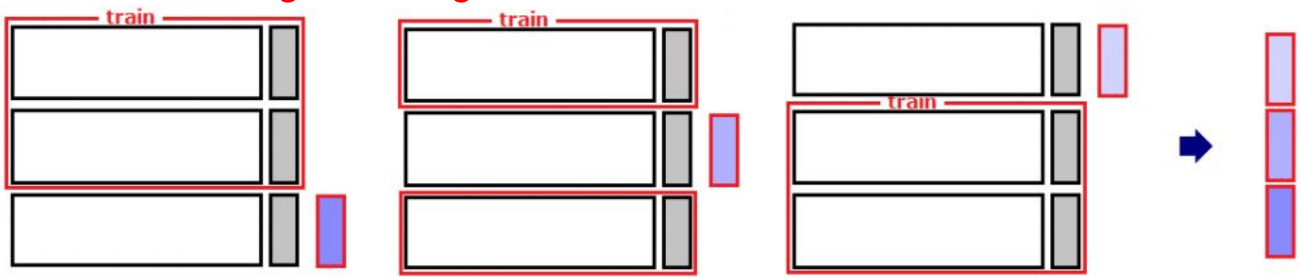
- Pros:
  - Simple and intuitive ensembling method.
  - Average several blendings to achieve better results.
- Cons:
  - Linear composition is not always enough.
  - Need to split the data. How to fix it?

Stacking

1. Split data into folds



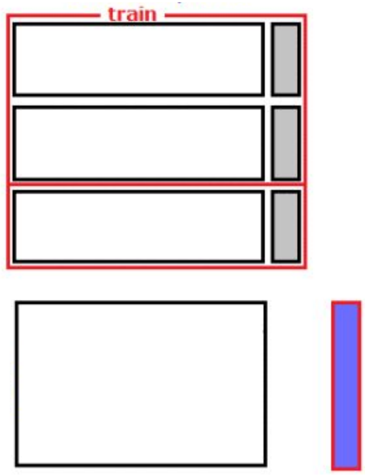
Fit using folds to get meta-features on train



2. Tune models on different groups of folds, predict on left out

3. Tune the new model on the “meta”-features

Fit using all data meta-features on test



Picture source: https://dyakonov.org/2017/03/10/стекинг-stacking-и-блендинг-blending/

Stacking

- Train base algorithm(s) on different groups of folds leaving one fold out.
- Predict the meta-features on the left-out fold and test data.
- Train the meta-algorithm on the meta-features representation of the train data.
- Use it on the meta-features representation of the test data.

Stacking

- Pros:
  - Powerful ensembling method, if you know how to use it
  - Quite popular in ML-competitions
  - One might perform stacking on the meta-features dataset as well
- Cons:
  - Meta-features on each fold are actually predicted by different models
    - However, regularization usually helps
  - Hard to explain your model behaviour

Stacking

Bonus:

Now you know how to stack XGBoost (or CatBoost/LightGBM)

Recap: ensembling methods

1. Bagging.
2. Random subspace method (RSM).
3. Bagging + RSM + Decision trees = Random Forest.
4. Gradient boosting.
5. Blending.
6. Stacking.

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