Why does Bagging work?

Why would φ_B be any better than φ?

Why does Bagging work?

Illustration on the regression case:

Suppose (X, Y) drawn from distribution $P_{X,Y}$.
φ predictor trained on $\mathcal{T}$ or any bootstrap sample of $\mathcal{T}$
$\hat{P}_{\mathcal{T}}$ empirical distribution of $\mathcal{T}$
$P_{\mathcal{T}}$ true distribution of $\mathcal{T}$
To simplify notation: $\mathbb{E}_{P_{X,Y}} = \mathbb{E}_{X,Y}$, $\mathbb{E}_{P_{\mathcal{T}}} = \mathbb{E}_{\mathcal{T}}$ and $\mathbb{E}_{\hat{P}_{\mathcal{T}}} = \mathbb{E}_{\hat{\mathcal{T}}}$.
$\varphi_B(\cdot) = \mathbb{E}_{\hat{\mathcal{T}}}(\varphi(\cdot))$ Bagging predictor
$\varphi_A(\cdot) = \mathbb{E}_{\mathcal{T}}(\varphi(\cdot))$ aggregated predictor

Why does Bagging work?

Average prediction error of φ: $e = \mathbb{E}_{\mathcal{T}}\left(\mathbb{E}_{X,Y}\left(\left[Y - \varphi\left(X\right)\right]^2\right)\right)$.

Why does Bagging work?

Average prediction error of φ: $e = \mathbb{E}_{\mathcal{T}}\left(\mathbb{E}_{X,Y}\left(\left[Y - \varphi\left(X\right)\right]^2\right)\right)$.

Average prediction error of φ_A: $e_A = \mathbb{E}_{X,Y}\left(\left[Y - \varphi_A\left(X\right)\right]^2\right)$.

Why does Bagging work?

Average prediction error of φ: $e = \mathbb{E}_{\mathcal{T}}\left(\mathbb{E}_{X,Y}\left([Y - \varphi(X)]^2\right)\right)$.

Average prediction error of φ_A: $e_A = \mathbb{E}_{X,Y}\left([Y - \varphi_A(X)]^2\right)$.

$$e = \mathbb{E}_{X,Y}\left(Y^2\right) - 2\mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left(Y\varphi(X)\right)\right) + \mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left([\varphi(X)]^2\right)\right)$$

Why does Bagging work?

Average prediction error of φ: $e = \mathbb{E}_{\mathcal{T}}\left(\mathbb{E}_{X,Y}\left([Y - \varphi(X)]^2\right)\right)$.

Average prediction error of φ_A: $e_A = \mathbb{E}_{X,Y}\left([Y - \varphi_A(X)]^2\right)$.

$$e = \mathbb{E}_{X,Y}\left(Y^2\right) - 2\mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left(Y\varphi(X)\right)\right) + \mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left([\varphi(X)]^2\right)\right)$$

Why does Bagging work?

Average prediction error of φ: $e = \mathbb{E}_{\mathcal{T}}\left(\mathbb{E}_{X,Y}\left(\left[Y - \varphi\left(X\right)\right]^2\right)\right)$.

Average prediction error of φ_A: $e_A = \mathbb{E}_{X,Y}\left(\left[Y - \varphi_A\left(X\right)\right]^2\right)$.

$e = \mathbb{E}_{X,Y}\left(Y^2\right) - 2\mathbb{E}_{X,Y}\left(Y\varphi_A(X)\right) + \mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left(\left[\varphi(X)\right]^2\right)\right)$

Why does Bagging work?

Average prediction error of φ: $e = \mathbb{E}_{\mathcal{T}}\left(\mathbb{E}_{X,Y}\left([Y - \varphi(X)]^2\right)\right)$.

Average prediction error of φ_A: $e_A = \mathbb{E}_{X,Y}\left([Y - \varphi_A(X)]^2\right)$.

$e = \mathbb{E}_{X,Y}\left(Y^2\right) - 2\mathbb{E}_{X,Y}\left(Y\varphi_A(X)\right) + \mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left([\varphi(X)]^2\right)\right)$

But $\mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left([\varphi(X)]^2\right)\right) \geq \mathbb{E}_{X,Y}\left([\mathbb{E}_{\mathcal{T}}(\varphi(X))]^2\right)$

Why does Bagging work?

Average prediction error of φ: $e = \mathbb{E}_{\mathcal{T}}\left(\mathbb{E}_{X,Y}\left([Y - \varphi(X)]^2\right)\right)$.

Average prediction error of φ_A: $e_A = \mathbb{E}_{X,Y}\left([Y - \varphi_A(X)]^2\right)$.

$e = \mathbb{E}_{X,Y}\left(Y^2\right) - 2\mathbb{E}_{X,Y}\left(Y\varphi_A(X)\right) + \mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left([\varphi(X)]^2\right)\right)$

But $\mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left([\varphi(X)]^2\right)\right) \geq \mathbb{E}_{X,Y}\left([\mathbb{E}_{\mathcal{T}}(\varphi(X))]^2\right)$

Why does Bagging work?

Average prediction error of φ: $e = \mathbb{E}_{\mathcal{T}}\left(\mathbb{E}_{X,Y}\left(\left[Y - \varphi\left(X\right)\right]^2\right)\right)$.

Average prediction error of φ_A: $e_A = \mathbb{E}_{X,Y}\left(\left[Y - \varphi_A\left(X\right)\right]^2\right)$.

$e = \mathbb{E}_{X,Y}\left(Y^2\right) - 2\mathbb{E}_{X,Y}\left(Y\varphi_A(X)\right) + \mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left(\left[\varphi(X)\right]^2\right)\right)$

But $\mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left(\left[\varphi(X)\right]^2\right)\right) \geq \mathbb{E}_{X,Y}\left(\left[\varphi_A(X)\right]^2\right)$

Why does Bagging work?

Average prediction error of φ: $e = \mathbb{E}_{\mathcal{T}}\left(\mathbb{E}_{X,Y}\left([Y - \varphi(X)]^2\right)\right)$.

Average prediction error of φ_A: $e_A = \mathbb{E}_{X,Y}\left([Y - \varphi_A(X)]^2\right)$.

$e = \mathbb{E}_{X,Y}\left(Y^2\right) - 2\mathbb{E}_{X,Y}\left(Y\varphi_A(X)\right) + \mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left([\varphi(X)]^2\right)\right)$

But $\mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left([\varphi(X)]^2\right)\right) \geq \mathbb{E}_{X,Y}\left([\varphi_A(X)]^2\right)$

So $e \geq e_A$.

Why does Bagging work?

Average prediction error of φ: $e = \mathbb{E}_{\mathcal{T}}\left(\mathbb{E}_{X,Y}\left([Y - \varphi(X)]^2\right)\right)$.

Average prediction error of φ_A: $e_A = \mathbb{E}_{X,Y}\left([Y - \varphi_A(X)]^2\right)$.

$e = \mathbb{E}_{X,Y}\left(Y^2\right) - 2\mathbb{E}_{X,Y}\left(Y\varphi_A(X)\right) + \mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left([\varphi(X)]^2\right)\right)$

But $\mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left([\varphi(X)]^2\right)\right) \geq \mathbb{E}_{X,Y}\left([\varphi_A(X)]^2\right)$

So $e \geq e_A$.

Moreover:

$e - e_A = \mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left([\varphi(X)]^2\right) - [\mathbb{E}_{\mathcal{T}}(\varphi(X))]^2\right)$

$e - e_A = \mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left([\varphi(X)]^2\right) - [\varphi_A(X)]^2\right)$

Why does Bagging work?

Average prediction error of φ: $e = \mathbb{E}_{\mathcal{T}}\left(\mathbb{E}_{X,Y}\left([Y - \varphi(X)]^2\right)\right)$.

Average prediction error of φ_A: $e_A = \mathbb{E}_{X,Y}\left([Y - \varphi_A(X)]^2\right)$.

$e = \mathbb{E}_{X,Y}\left(Y^2\right) - 2\mathbb{E}_{X,Y}\left(Y\varphi_A(X)\right) + \mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left([\varphi(X)]^2\right)\right)$

But $\mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left([\varphi(X)]^2\right)\right) \geq \mathbb{E}_{X,Y}\left([\varphi_A(X)]^2\right)$

So $e \geq e_A$.

Moreover:

$e - e_A = \mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left([\varphi(X)]^2\right) - [\mathbb{E}_{\mathcal{T}}(\varphi(X))]^2\right)$

$e - e_A = \mathbb{E}_{X,Y}\left(\mathbb{E}_{\mathcal{T}}\left([\varphi(X)]^2\right) - [\varphi_A(X)]^2\right)$

Interpretation: if $\varphi_{\mathcal{T}}$ differs a lot from $\varphi_{\mathcal{T}'}$, then $e - e_A$ is large.
$\Rightarrow$ The highest the variance of φ across training sets $\mathcal{T}$, the more improvement φ_A produces.

Why does Bagging work?

Ok, so φ_A always improves on φ,
especially when φ is highly variable w.r.t. changes in $\mathcal{T}$.

Why does Bagging work?

Ok, so φ_A always improves on φ,
especially when φ is highly variable w.r.t. changes in $\mathcal{T}$.

But φ_A is not φ_B. Recall:
$\varphi_A(\cdot) = \mathbb{E}_{\mathcal{T}}\left(\varphi(\cdot)\right)$ aggregated predictor (over all N-size training sets)
$\varphi_B(\cdot) = \mathbb{E}_{\hat{\mathcal{T}}}\left(\varphi(\cdot)\right)$ Bagging predictor (over bootstrap samples)
φ_B approximates φ_A and thus $e_B \geq e_A$

Why does Bagging work?

Ok, so φ_A always improves on φ,
especially when φ is highly variable w.r.t. changes in $\mathcal{T}$.

But φ_A is not φ_B. Recall:
$\varphi_A(\cdot) = \mathbb{E}_{\mathcal{T}}\left(\varphi(\cdot)\right)$ aggregated predictor (over all N-size training sets)
$\varphi_B(\cdot) = \mathbb{E}_{\hat{\mathcal{T}}}\left(\varphi(\cdot)\right)$ Bagging predictor (over bootstrap samples)
φ_B approximates φ_A and thus $e_B \geq e_A$

- If φ highly variable w.r.t. $\mathcal{T}$, φ_B improves on φ through aggregation.
- But if φ is rather stable w.r.t. $\mathcal{T}$, $e_A \approx e$ and since φ_B approximates φ_A, e_B might be greater than e.

Why does Bagging work?

So it does not always work?

Why does Bagging work?

So it does not always work?
Actually, no, it does not always work.
Bagging should be used to transform highly variable predictors φ into a more accurate averaged commitee φ_B.

Examples of φ that Bagging improve:
$\rightarrow$ Trees, Neural Networks.
Examples of φ that Bagging does not improve much (or degrades):
$\rightarrow$ Support Vector Machines, Gaussian Processes.

Why does Bagging work?

And in the classification case?

Why does Bagging work?

And in the classification case?

Majority vote: $\varphi_B(x) = \arg\max_j \sum_{b=1}^{B} I(\varphi^b(x) = j)$

More drastic conclusions:

- φ unstable w.r.t. $\mathcal{T}$ and reasonable performance $\Rightarrow \varphi_B$ near optimal.
- φ stable w.r.t. $\mathcal{T} \Rightarrow \varphi_B$ worse than φ.
- φ poor performance $\Rightarrow \varphi_B$ worse than φ.