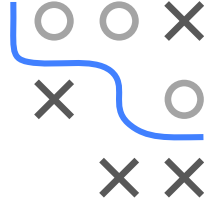
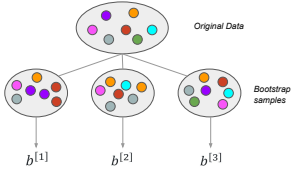
Introduction to Machine Learning



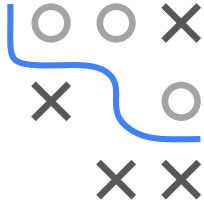
Regularization
Bagging as Regularization (Deep-Dive)



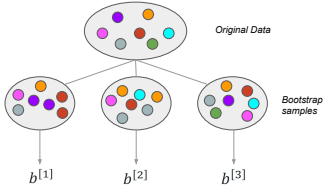
Learning goals

- Understand that bagging can be seen as a form of regularization
- Know which factors influence the effectiveness of bagging

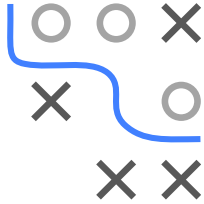
RECAP: WHAT IS BAGGING?



- Bagging is short for **B**ootstrap **Agg**regation.
- It's an **ensemble method**, i.e., it combines many models into one big "meta-model". Ensembles often work much better than their members alone would.
- The components of an ensemble are called **base learners** (BLs)
- In a **bagging** ensemble, all base learners are of the same type. The only difference between the models is the data they are trained on.



RECAP: WHAT IS BAGGING? / 2

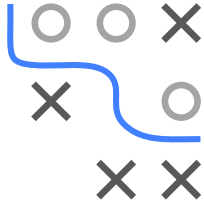


Specifically, we train base learners $b^{[m]}, m = 1, \ldots, M$ on M **bootstrap** samples of training data $\mathcal{D}$:

- Draw n observations from $\mathcal{D}$ with replacement
- Fit the base learner on each of the M bootstrap samples to get models $\hat{f}(x) = \hat{b}^{[m]}, m = 1, \ldots, M$
- Aggregate predictions of the M fitted base learners to get ensemble model $\hat{f}^{[M]}(\mathbf{x})$ via **averaging** (regression) or **majority voting** (classification)

Bagging helps because variability of the averaged prediction over many base learners is smaller than variability of the predictions from one such model. If error of BL is mostly due to (random) variability and not structural reasons bagging helps reducing this variability.

WHY/WHEN DOES BAGGING HELP?



Assume we use quadratic loss and measure instability of the ensemble with $\Delta\left(f^{[M]}(\mathbf{x})\right) = \frac{1}{M}\sum_m^M \left(b^{[m]} - f^{[M]}(\mathbf{x})\right)^2$:

$$\begin{aligned}
\Delta\left(f^{[M]}(\mathbf{x})\right) &= \frac{1}{M}\sum_m^M \left(b^{[m]} - f^{[M]}(\mathbf{x})\right)^2 \\
&= \frac{1}{M}\sum_m^M \left(\left(b^{[m]} - y\right) + \left(y - f^{[M]}(\mathbf{x})\right)\right)^2 \\
&= \frac{1}{M}\sum_m^M L(y, b^{[m]}) + L(y, f^{[M]}(\mathbf{x})) \underbrace{-2\left(y - \frac{1}{M}\sum_{m=1}^M b^{[m]}\right)\left(y - f^{[M]}(\mathbf{x})\right)}_{-2L\left(y, f^{[M]}(\mathbf{x})\right)}
\end{aligned}$$

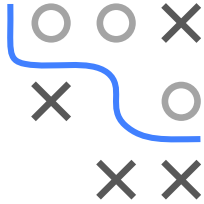
So, if we take the expected value over the data's distribution:

$$\mathbb{E}_{xy}\left[L\left(y, f^{[M]}(\mathbf{x})\right)\right] = \frac{1}{M}\sum_m^M \mathbb{E}_{xy}\left[L\left(y, b^{[m]}\right)\right] - \mathbb{E}_{xy}\left[\Delta\left(f^{[M]}(\mathbf{x})\right)\right]$$

$\Rightarrow$ Expected loss of the ensemble is lower than the average loss of single BL by the amount of instability in the ensemble's BLs. The more accurate and diverse the BLs, the better.

DETERMINANTS OF BAGGING EFFECTIVENESS

How to make $\mathbb{E}_{xy}\left[\Delta\left(f^{[M]}(\mathbf{x})\right)\right]$ as large as possible?

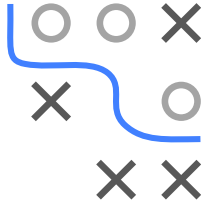


$$\mathbb{E}_{xy}\left[L\left(y, f^{[M]}(\mathbf{x})\right)\right] = \frac{1}{M}\sum_{m}^{M}\mathbb{E}_{xy}\left[L\left(y, b^{[m]}\right)\right] - \mathbb{E}_{xy}\left[\Delta\left(f^{[M]}(\mathbf{x})\right)\right]$$

For simplicity, assume $\mathbb{E}_{xy}\left[b^{[m]}\right] = 0$, $\mathrm{Var}_{xy}\left[b^{[m]}\right] = \mathbb{E}_{xy}\left[(b^{[m]})^2\right] = \sigma^2$, $\mathrm{Corr}_{xy}\left[b^{[m]}, b^{[m']}\right] = \rho$ for all m, m'.

$$\begin{aligned}
\implies \mathrm{Var}_{xy}\left[f^{[M]}(\mathbf{x})\right] &= \tfrac{1}{M}\sigma^2 + \tfrac{M-1}{M}\rho\sigma^2 \qquad \left(\ldots = \mathbb{E}_{xy}\left[(f^{[M]}(\mathbf{x}))^2\right]\right) \\
\mathbb{E}_{xy}\left[\Delta\left(f^{[M]}(\mathbf{x})\right)\right] &= \tfrac{1}{M}\sum_{m}^{M}\mathbb{E}_{xy}\left[\left(b^{[m]} - f^{[M]}(\mathbf{x})\right)^2\right] \\
&= \tfrac{1}{M}\left(M\mathbb{E}_{xy}\left[(b^{[m]})^2\right] + M\mathbb{E}_{xy}\left[(f^{[M]}(\mathbf{x}))^2\right] - 2M\mathbb{E}_{xy}\left[b^{[m]} f^{[M]}(\mathbf{x})\right]\right) \\
&= \sigma^2 + \mathbb{E}_{xy}\left[(f^{[M]}(\mathbf{x}))^2\right] - 2\tfrac{1}{M}\sum_{m'}^{M}\underbrace{\mathbb{E}_{xy}\left[b^{[m]}b^{[m']}\right]}_{=\mathrm{Cov}_{xy}\left[b^{[m]}, b^{[m']}\right] + \mathbb{E}_{xy}\left[b^{[m]}\right]\mathbb{E}_{xy}\left[b^{[m']}\right]} \\
&= \sigma^2 + \left(\tfrac{1}{M}\sigma^2 + \tfrac{M-1}{M}\rho\sigma^2\right) - 2\left(\tfrac{M-1}{M}\rho\sigma^2 + \tfrac{1}{M}\sigma^2 + 0\cdot 0\right) \\
&= \tfrac{M-1}{M}\sigma^2(1-\rho)
\end{aligned}$$

DETERMINANTS OF BAGGING EFFECTIVENESS / 2



$$\mathbb{E}_{xy}\left[L\left(y, f^{[M]}(\mathbf{x})\right)\right] = \frac{1}{M}\sum_{m}^{M}\mathbb{E}_{xy}\left[L\left(y, b^{[m]}\right)\right] - \mathbb{E}_{xy}\left[\Delta\left(f^{[M]}(\mathbf{x})\right)\right]$$

$$\mathbb{E}_{xy}\left[\Delta\left(f^{[M]}(\mathbf{x})\right)\right] \cong \frac{M-1}{M}\mathrm{Var}_{xy}\left[b^{[m]}\right]\left(1 - \mathrm{Corr}_{xy}\left[b^{[m]}, b^{[m']}\right]\right)$$

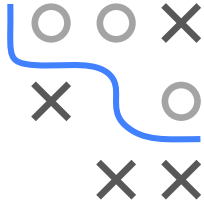
$\Rightarrow$ **better base learners** are better (... duh)

$\Rightarrow$ **more base learners** are better (theoretically, at least...)

$\Rightarrow$ **more variable base learners** are better (as long as their risk stays the same, of course!)

$\Rightarrow$ **less correlation between base learners** is better:
bagging helps more if base learners are wrong in different ways so that their errors “cancel” each other out.

BAGGING SUMMARY



- Basic idea: fit the same model repeatedly on many **bootstrap** replications of the training data set and **aggregate** the results
- Gains in performance by reducing variance of predictions, but (slightly) increases the bias: it reuses training data many times, so small mistakes can get amplified.
  Bagging is thus a **form of regularization**
- Works best for unstable/high-variance BLs, where small changes in training set can cause large changes in predictions:
  e.g., CART, neural networks, step-wise/forward/backward variable selection for regression
- Works best if BL predictions are only weakly correlated: they don't all make the same mistakes.
- Can degrade performance for stable methods like k-NN, LDA, Naive Bayes, linear regression