Ensemble Learning: Principles

Victor Araujo Ferraz
COPPE/UFRJ
email: victor.ecomp@gmail.com

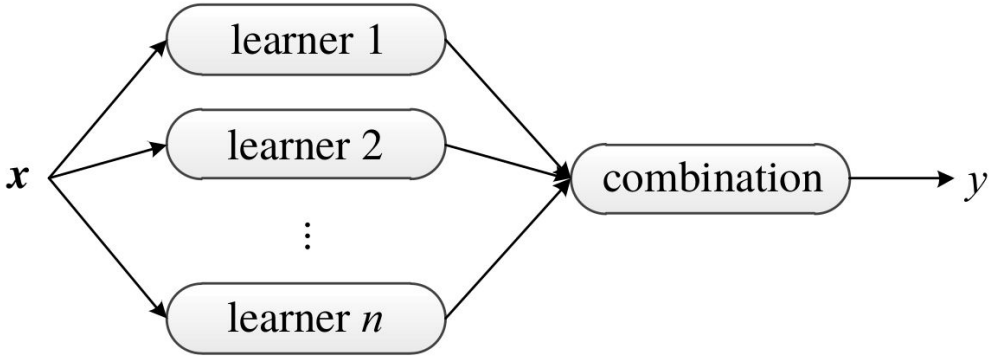
Summary

- Introduction
- Combination
  - averaging
  - voting
  - combining by learning
- Boosting
- Bagging
- References

Introduction

- Ensemble methods train multiple learners to solve the same problem
- Try to construct a set of learners and combine them
- Also called:
  - **committee-based learning**
  - **learning multiple classifier systems**
  - **combining pattern classifiers**

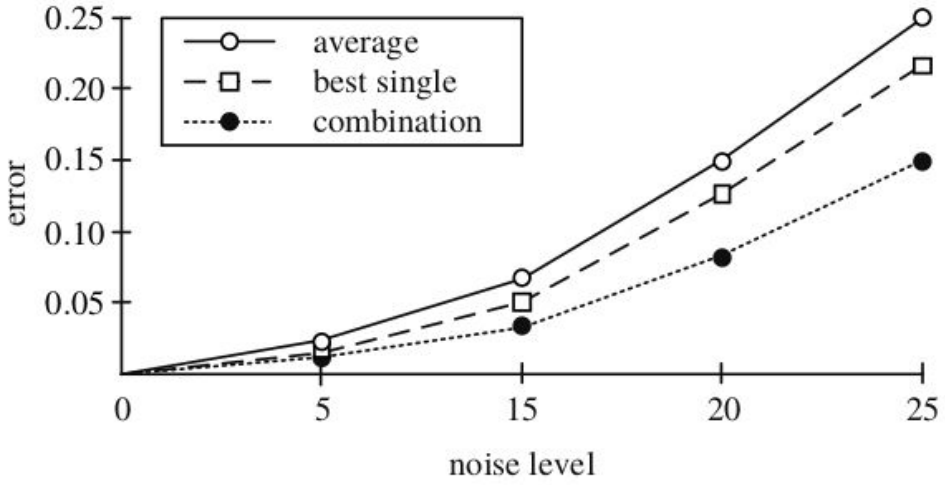
- Contains a number of learners called:
  - **base learners or**
  - **individual learners or**
  - **component learner**



Introduction

- Ensemble methods have become a major learning paradigm since the 1990s
- Two pieces of pioneering work
  - Empirical [Hansen e Salamon, 1990]
  - Theoretical [Schapire, 1990]

- The computational cost is not much larger than creating a single learner



Source: [Hansen e Salamon, 1990]

Why Ensemble?

- Best motivation for us
  - CheXpert is a large dataset of chest X-rays and competition for automated chest x-ray interpretation, which features uncertainty labels and radiologist-labeled reference standard evaluation sets
  - https://stanfordmlgroup.github.io/competitions/chexpert/

Leaderboard

Will your model perform as well as radiologists in detecting different pathologies in chest X-rays?

Rank	Date	Model	AUC	Num Rads Below Curve
1	Aug 31, 2020	SuperCNN *ensemble*	0.930	2.8
2	Sep 01, 2019	Hierarchical-Learning-V1 (ensemble) *Vingroup Big Data Institute* https://arxiv.org/abs/1911.06475	0.930	2.6
3	Oct 15, 2019	Conditional-Training-LSR *ensemble*	0.929	2.6
4	Dec 04, 2019	Hierarchical-Learning-V4 (ensemble) *Vingroup Big Data Institute* https://arxiv.org/abs/1911.06475	0.929	2.6

Combination: Averaging

- It is the most popular and fundamental combination method for numeric outputs
- **Simple Averaging**

$$H(\boldsymbol{x}) = \frac{1}{T}\sum_{i=1}^{T} h_i(\boldsymbol{x})$$

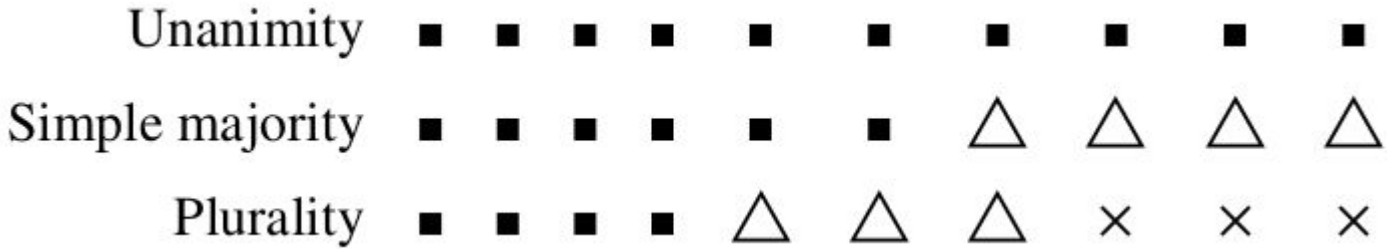
- **Weighted Averaging**

$$H(\boldsymbol{x}) = \sum_{i=1}^{T} w_i h_i(\boldsymbol{x})$$

- Given a set of T individual learners $\{h_1, \ldots, h_T\}$ and the output of h_i for the instance x is $h_i(x) \in \mathbb{R}$

Voting

- It is the most popular and fundamental combination method for nominal outputs
- Every classifier votes for one class label, and the final output class label is decided accordingly the voting method
- **Majority Voting**: three consensus patterns



- If none of the class labels receives the enough votes, a rejection option will be given and the combined classifier makes no prediction

Voting

- **Simple Majority**

$$H(\boldsymbol{x}) = \begin{cases} c_j & \text{if } \sum_{i=1}^{T} h_i^j(\boldsymbol{x}) > \frac{1}{2} \sum_{k=1}^{l} \sum_{i=1}^{T} h_i^k(\boldsymbol{x}) \\ \text{rejection} & \text{otherwise}. \end{cases}$$

- **Plurality**

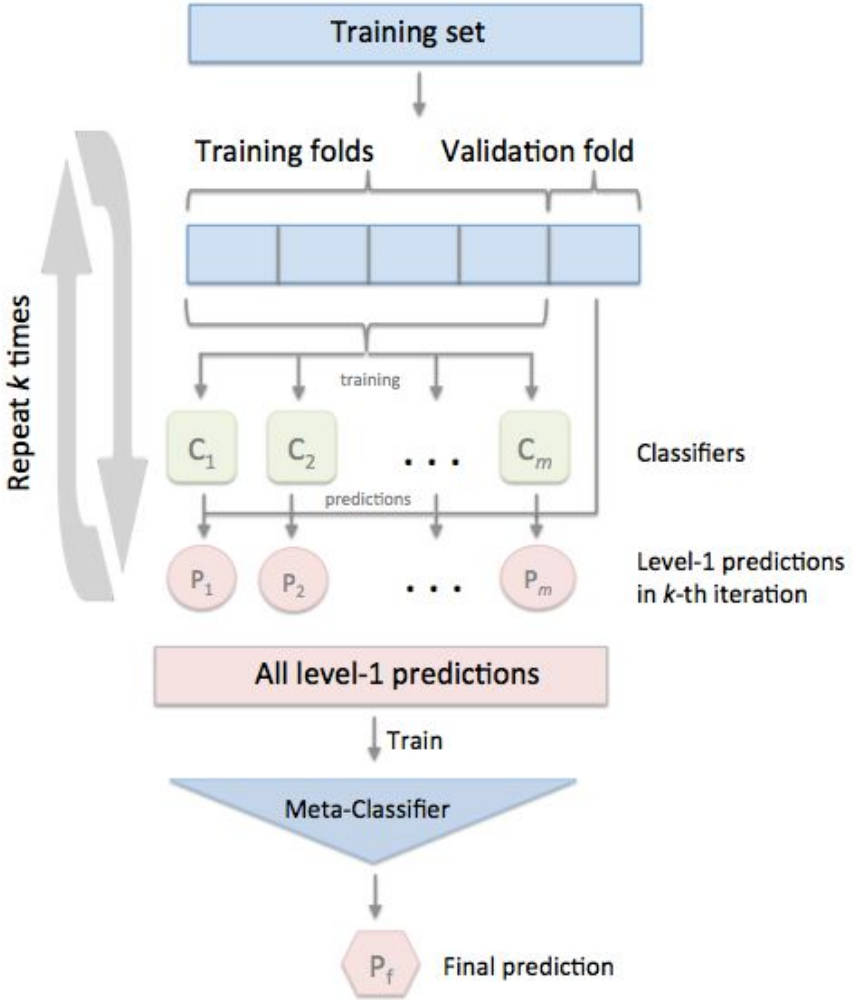
$$H(\boldsymbol{x}) = c_{\arg\max_j \sum_{i=1}^{T} h_i^j(\boldsymbol{x})}$$

- **Weighted**

$$H(\boldsymbol{x}) = c_{\arg\max_j \sum_{i=1}^{T} w_i h_i^j(\boldsymbol{x})}$$

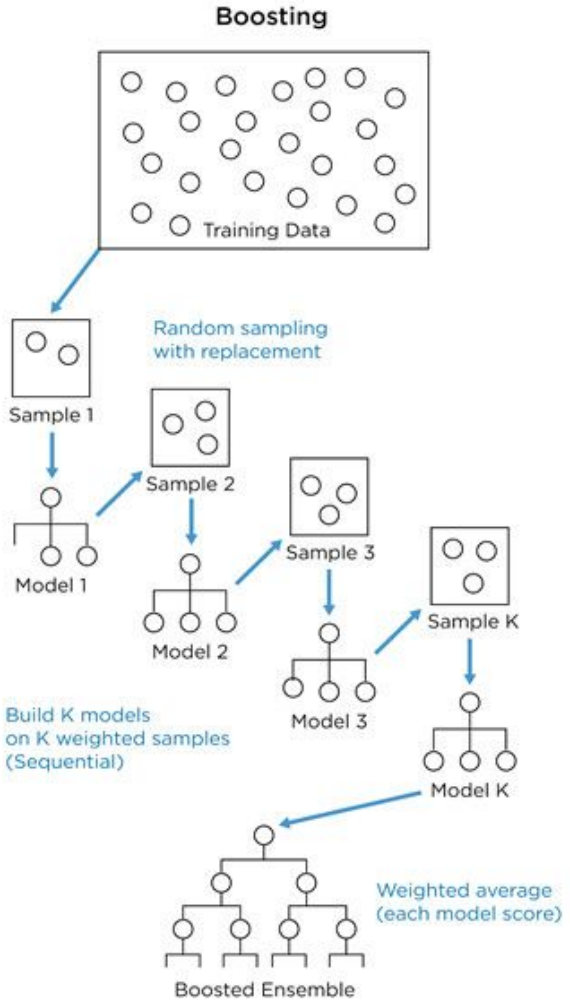
Combining by learning

- It is a general procedure where a learner is trained to combine the individual learners
- Main example: **Stacking**
  - the individual learners are called the first-level learners, while the combiner is called the second-level learner, or meta-learner



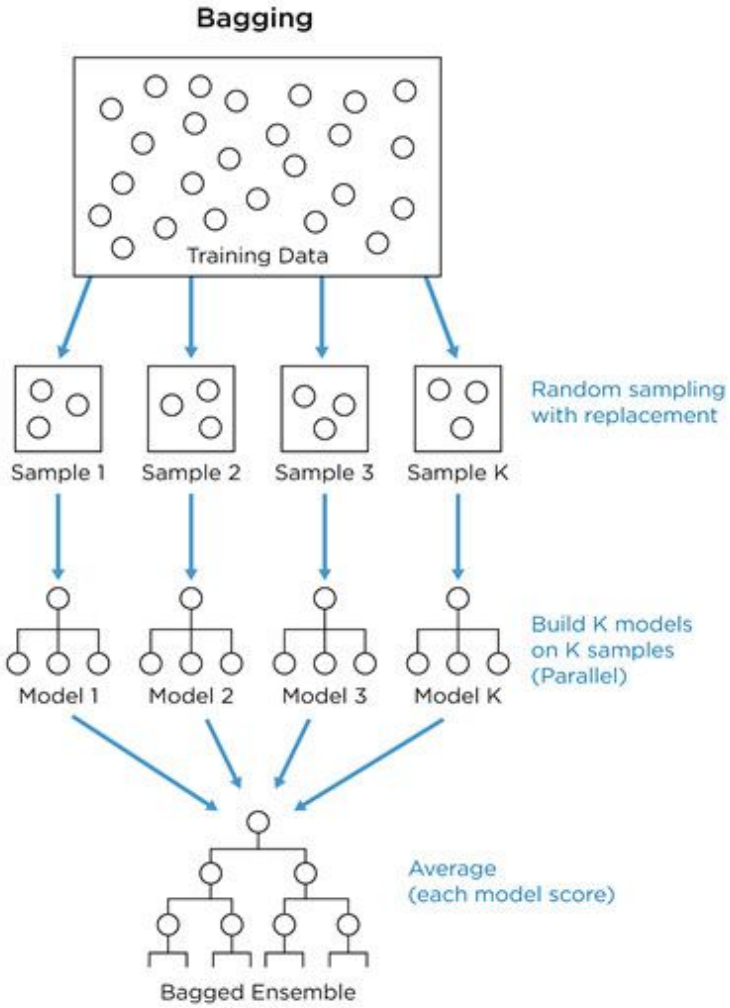
Boosting

- Refers to a family of algorithms that are able to convert weak learners to strong learners
- Boosting works by training a set of learners sequentially and combining them for prediction
- The later learners focus more on the mistakes of the earlier learners
- The most influential boosting algorithm: **AdaBoost**



Bagging

- Bootstrap AGGregatING
- It is a parallel ensemble method
- Bagging adopts the bootstrap distribution for generating different base learners
- For aggregating the outputs of the base learners, Bagging adopts the most popular strategies: voting for classification and averaging for regression



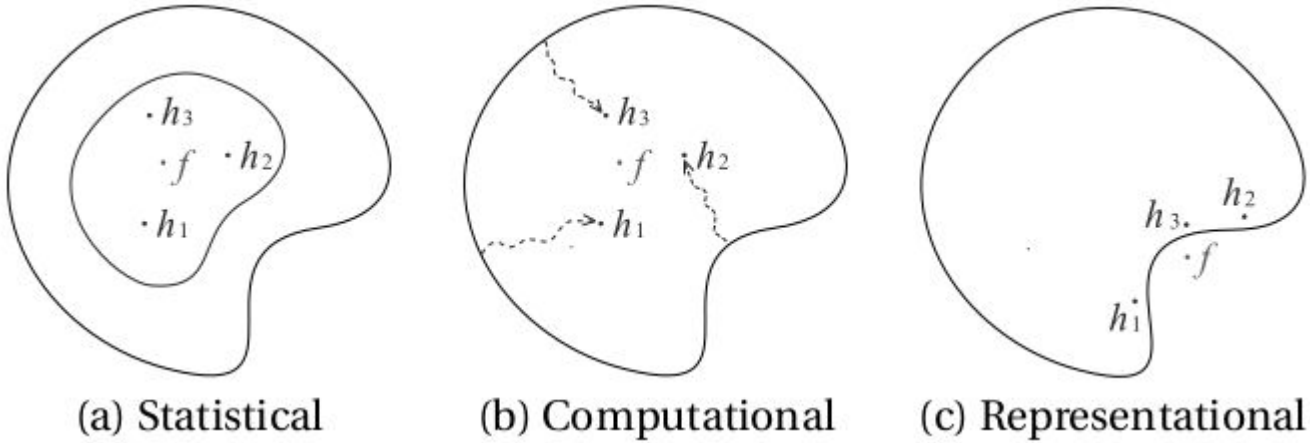
References

- Zhi-Hua Zhou. 2012. Ensemble Methods: Foundations and Algorithms (1st. ed.). Chapman & Hall/CRC.
- Kuncheva, L. I. (2004). Combining pattern classifiers : methods and algorithms. J. Wiley.
- L. K. Hansen and P. Salamon. Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(10):993–1001, 1990.
- R. E. Schapire. The strength of weak learnability. Machine Learning, 5(2):197–227, 1990.

Backup slides

Combination

- The combination method plays a crucial role
- Benefits on the following issues
  - Statistical: the risk of choosing a wrong hypothesis can be reduced
  - Computational: the risk of choosing a wrong local minimum can be reduced
  - Representational: it may be possible to expand the space of representable functions



General Boosting procedure

Input: Sample distribution $\mathcal{D}$;
Base learning algorithm $\mathfrak{L}$;
Number of learning rounds T.

Process:

1. $\mathcal{D}_1 = \mathcal{D}$. % Initialize distribution
2. **for** $t = 1, \ldots, T$:
3. $h_t = \mathfrak{L}(\mathcal{D}_t)$; % Train a weak learner from distribution $\mathcal{D}_t$
4. $\epsilon_t = P_{\boldsymbol{x} \sim D_t}(h_t(\boldsymbol{x}) \neq f(\boldsymbol{x}))$; % Evaluate the error of h_t
5. $\mathcal{D}_{t+1} = Adjust\_Distribution(\mathcal{D}_t, \epsilon_t)$
6. **end**

Output: $H(\boldsymbol{x}) = Combine\_Outputs(\{h_1(\boldsymbol{x}), \ldots, h_t(\boldsymbol{x})\})$

General Bagging Procedure

Input: Data set $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$;
Base learning algorithm $\mathfrak{L}$;
Number of base learners T.

Process:

1. **for** $t = 1, \ldots, T$:
2. $h_t = \mathfrak{L}(D, \mathcal{D}_{bs})$ % $\mathcal{D}_{bs}$ is the bootstrap distribution
3. **end**

Output: $H(\boldsymbol{x}) = \underset{y \in \mathcal{Y}}{\arg\max} \sum_{t=1}^{T} \mathbb{I}(h_t(\boldsymbol{x}) = y)$