Sapienza University of Rome

Master in Artificial Intelligence and Robotics

Machine Learning

A.Y. 2025/2026

Prof. Luca Iocchi

Sapienza University of Rome, Italy - Machine Learning (2024/2025)

13. Multiple learners

Luca Iocchi

with contributions from Valsamis Ntouskos

Overview

- Combining multiple learners
- Voting
- Bagging
- Boosting
- AdaBoost

Reference

E. Alpaydin. Introduction to Machine Learning. Chapter 17.
C. Bishop. Pattern Recognition and Machine Learning. Chapter 14.

Multiple learners / Ensemble learning

General idea: instead of training a complex learner/model, train many different learners/models and then combine their results.

Committees: set of models trained on a dataset.

Models can be trained in parallel (*voting* or *bagging*) or in sequence (*boosting*).

Voting

Given a dataset D

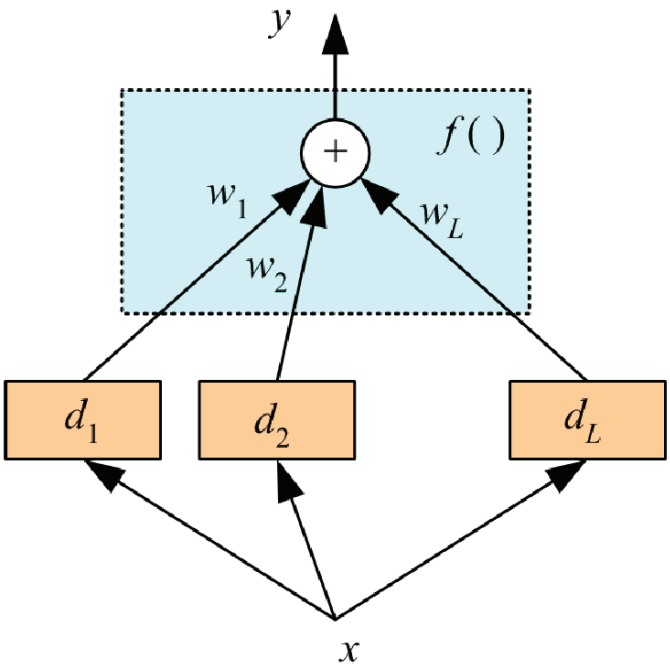
1. use D to train a set of models $y_m(\mathrm{x})$, for $m = 1, \ldots, M$
2. make predictions with

$$y_{voting}(\mathrm{x}) = \sum_{m=1}^{M} w_m y_m(\mathrm{x}) \qquad \text{(regression)}$$

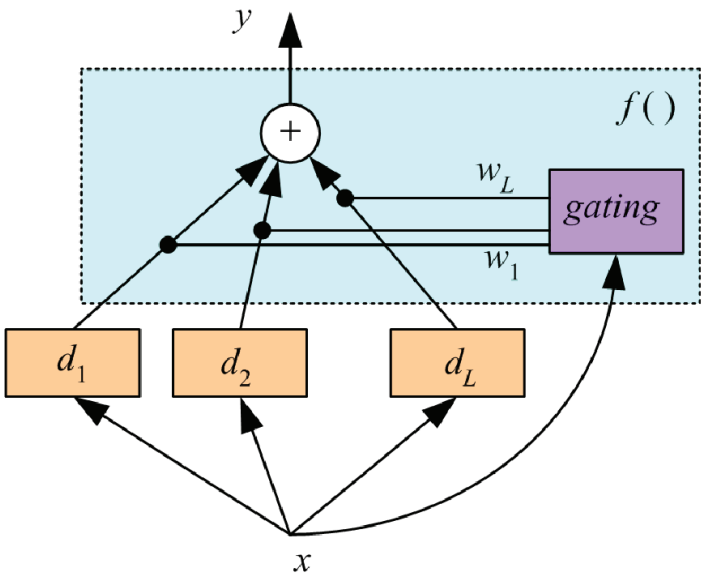
$$y_{voting}(\mathrm{x}) = \underset{c}{\operatorname{argmax}} \sum_{m=1}^{M} w_m I(y_m(\mathrm{x}) = c) \qquad \text{weighted majority (classification)}$$

with $w_m \geq 0$, $\sum_m w_m = 1$ (prior probability of each model),
$I(e) = 1$ if e is true, 0 otherwise.

Voting

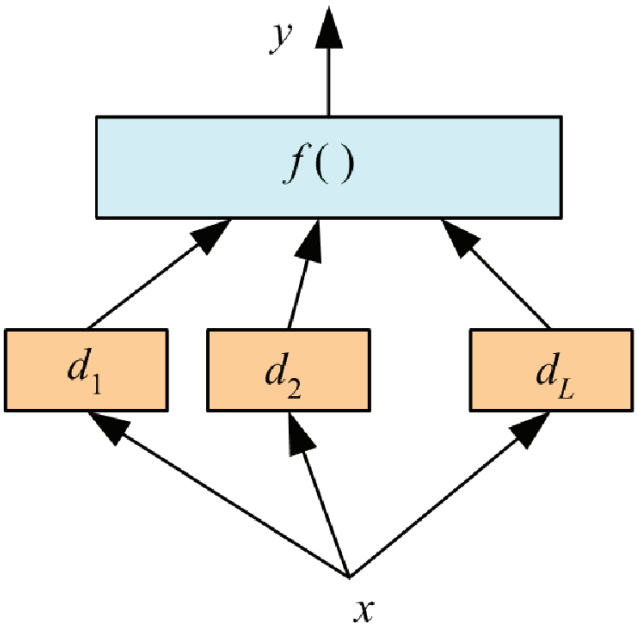


Mixture of experts



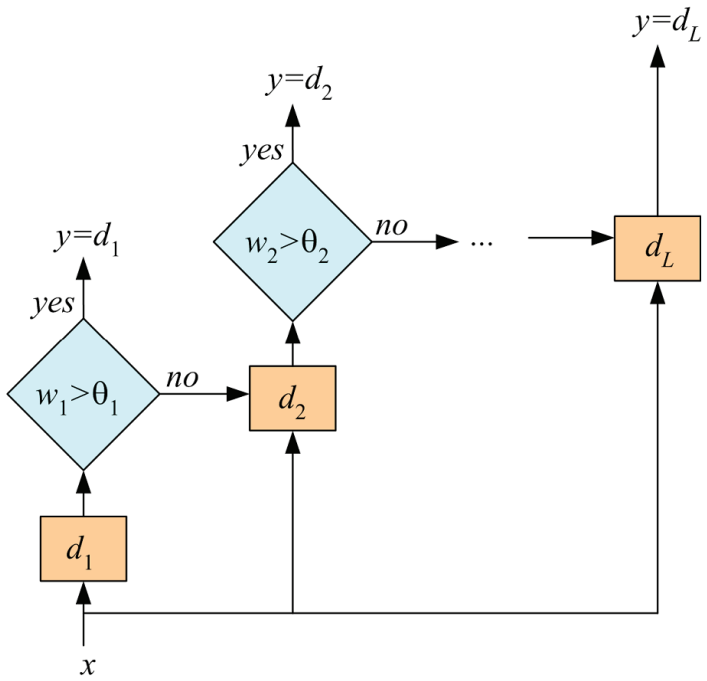
Non linear gating function f depending on input

Stacking



Combination function f is also learned

Cascading



Cascade learners based on confidence thresholds

Bagging

Given a dataset D,

1. generate M bootstrap data sets $D_1, \ldots, D_M$, with $D_i \subset D$
2. use each bootstrap data set D_m to train a model $y_m(\mathrm{x})$, for $m = 1, \ldots, M$
3. make predictions with a voting scheme

$$y_{bagging}(\mathrm{x}) = \frac{1}{M} \sum_{m=1}^{M} y_m(\mathrm{x})$$

In general, this is better than training any individual model.

Bootstrap data sets chosen with *random sampling with replacement*

Bagging

Remarks

- Bagging useful to train the very same model on different subsets of the training set
- Reduce overfitting
- Parallel training

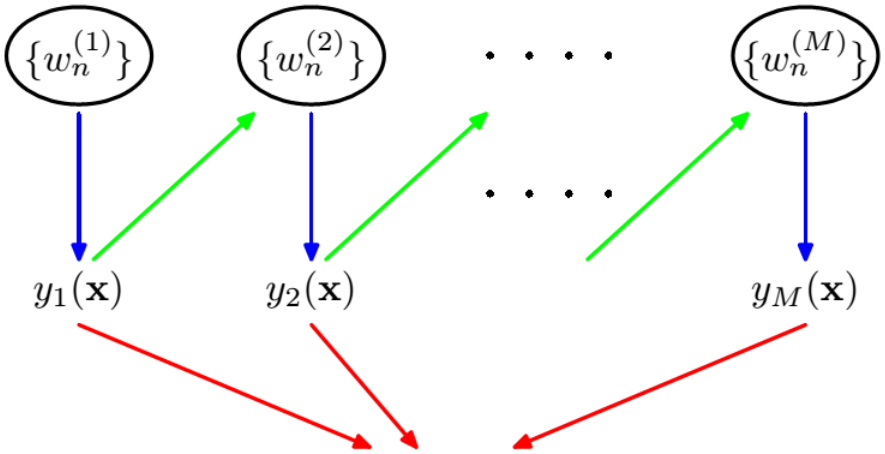
Boosting: general approach

Main points:

- Base classifiers (*weak learners*) trained sequentially
- Each classifier trained on weighted data
- Weights depend on performance of previous classifiers
- Points misclassified by previous classifiers are given greater weight
- Predictions based on weighted majority of votes
- Sequential training

Boosting: general approach

Base classifiers are trained in sequence using a weighted data set where weights are based on performance of previous classifiers.



$$Y_M(\mathrm{x}) = \mathrm{sign}\left(\sum_{m=1}^{M} \alpha_m y_m(\mathrm{x})\right)$$

AdaBoost

Given $D = \{(\mathrm{x}_1, t_1), \ldots, (\mathrm{x}_N, t_N)\}$, where $\mathrm{x}_n \in \mathrm{X}$, $t_n \in \{-1, +1\}$

1. Initialize $w_n^{(1)} = 1/N$, $n = 1, ..., N$.
2. For $m = 1, ..., M$:
   - Train a weak learner $y_m(\mathrm{x})$ by minimizing the weighted error function:

$$J_m = \sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathrm{x}_n) \neq t_n), \text{ with } I(e) = \begin{cases} 1 & \text{if } e \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

   - Evaluate: $\epsilon_m = \dfrac{\sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathrm{x}_n) \neq t_n)}{\sum_{n=1}^{N} w_n^{(m)}}$ and $\alpha_m = \ln\left[\dfrac{1-\epsilon_m}{\epsilon_m}\right]$
   - Update the data weighting coefficients:

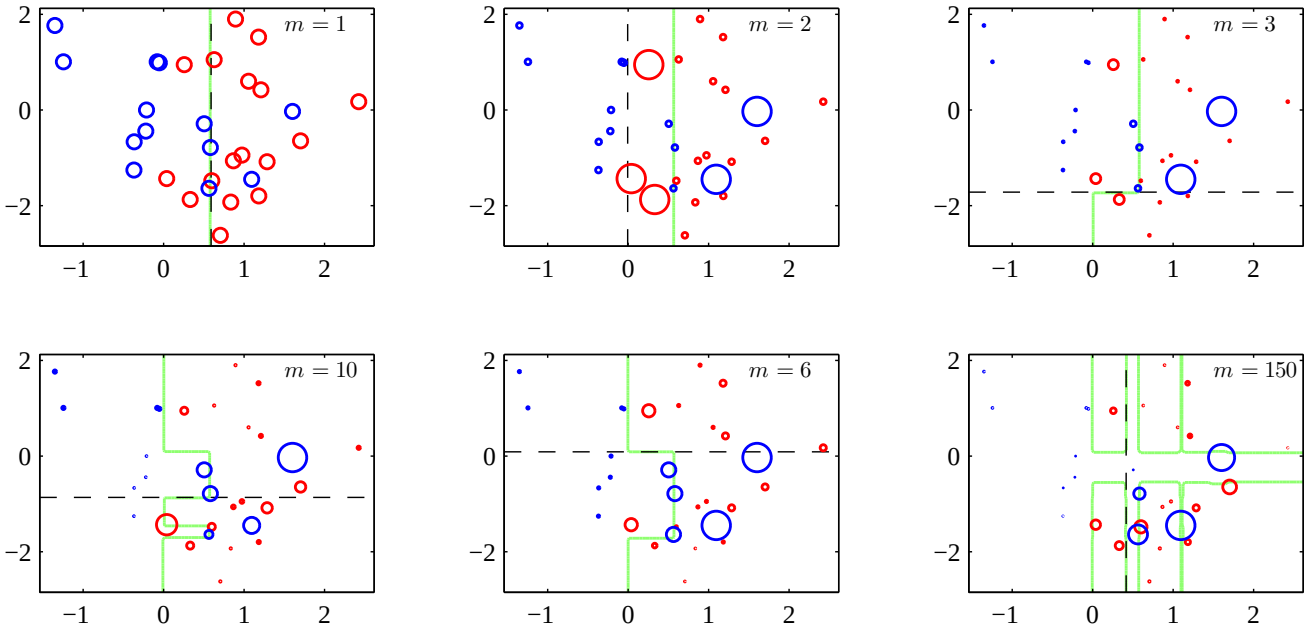
$$w_n^{(m+1)} = w_n^{(m)} \exp[\alpha_m I(y_m(\mathrm{x}_n) \neq t_n)]$$

AdaBoost

3. Output the final classifier

$$Y_M(\mathrm{x}) = \mathrm{sign}\left(\sum_{m=1}^{M} \alpha_m y_m(\mathrm{x})\right)$$

AdaBoost



Exponential error minimization

AdaBoost can be explained as the sequential minimization of an exponential error function.

Consider the error function

$$E = \sum_{n=1}^{N} \exp[-t_n f_M(\mathrm{x}_n)],$$

where

$$f_M(\mathrm{x}) = \frac{1}{2} \sum_{m=1}^{M} \alpha_m y_m(\mathrm{x}), \quad t_n \in \{-1, +1\}$$

Goal:

$$\text{minimize } E \text{ w.r.t. } \alpha_m, y_m(\mathrm{x}),\ m = 1, \ldots, M$$

Exponential error minimization

Sequential minimization. Instead of minimizing E globally

- assume $y_1(\mathrm{x}), \ldots, y_{M-1}(\mathrm{x})$ and $\alpha_1, \ldots, \alpha_{M-1}$ fixed;
- minimize w.r.t. $y_M(\mathrm{x})$ and α_M.

Making $y_M(\mathrm{x})$ and α_M explicit we have:

$$\begin{aligned} E &= \sum_{n=1}^{N} \exp\left[-t_n f_{M-1}(\mathrm{x}_n) - \frac{1}{2} t_n \alpha_M y_M(\mathrm{x}_n)\right] \\ &= \sum_{n=1}^{N} w_n^{(M)} \exp\left[-\frac{1}{2} t_n \alpha_M y_M(\mathrm{x}_n)\right], \end{aligned}$$

with $w_n^{(M)} = \exp[-t_n f_{M-1}(\mathrm{x}_n)]$ constant as we are optimizing w.r.t. α_M and $y_M(\mathrm{x})$.

Exponential error minimization

From sequential minimization of E, we obtain

$$w_n^{(m+1)} = w_n^{(m)} \exp[\alpha_m I(y_m(\mathrm{x}_n) \neq t_n)] \text{ and } \alpha_m = \ln \left[\frac{1 - \epsilon_m}{\epsilon_m} \right]$$

predictions are made with

$$\text{sign}(f_M(\mathrm{x})) = \text{sign} \left(\frac{1}{2} \sum_{m=1}^{M} \alpha_m y_m(\mathrm{x}) \right)$$

which is equivalent to

$$Y_M(\mathrm{x}) = \text{sign} \left(\sum_{m=1}^{M} \alpha_m y_m(\mathrm{x}) \right)$$

thus proving that AdaBoost minimizes such error function.

AdaBoost Remarks

Advantages:

- fast, simple and easy to program
- no prior knowledge about base learner is required
- no parameters to tune (except for M)
- can be combined with any method for finding base learners
- theoretical guarantees given sufficient data and base learners with moderate accuracy

Issues:

- Performance depends on data and the base learners
  (can fail with insufficient data or when base learners are too weak)
- Sensitive to noise

Bias-Variance-Noise Decomposition

$y(x) = f(x) + \epsilon$: observations
$\hat{f}(x; D)$: prediction of model trained with D
$S = \{(x_i, y_i)\}$: test set

$$MSE = Bias^2 + Variance + Noise$$

$$MSE = E_S[(y - \hat{f}(x; D))^2]$$

$$Bias = E_S[\hat{f}(x; D)] - f(x)$$

$$Variance = E_S[(\hat{f}(x; D) - E_S[\hat{f}(x; D)])^2] = Var_S\{\hat{f}(x; D)\}$$

$$Noise = E_S\{(y - f(x))^2\} = E[\epsilon^2]$$

Bias-Variance-Noise Decomposition

High Bias → high model error → underfitting
High Variance → high ability to adapt to datasets → overfitting

In general, we cannot minimize both Bias and Variance.

Model complexity

Low model complexity → high Bias (underfitting)
High model complexity → high Variance (overfitting)

Ensemble methods

Bagging methods reduce Variance
Boosting methods reduce Bias

Summary

- Instead of designing a learning algorithm that is accurate over the entire space one can focus on finding base learning algorithms that only need to be better than random (e.g., ensembles of small DNN may outperform very deep NN)
- Combined learners theoretically outperforms any individual learner
- AdaBoost practically outperforms many other base learners in many problems