Machine Learning course

Lecture 6: Ensembles

Harbour.Space
March 2021

Radoslav Neychev

Outline

1. Bootstrap and Bagging
2. Random Forest
3. Boosting intuitions

Bootstrap and Bagging

Bootstrap

Consider dataset X containing m objects.

Pick m objects with return from X and repeat in N times to get N datasets.

Error of model trained on Xj: $\varepsilon_j(x) = b_j(x) - y(x), \qquad j = 1, \dots, N,$

Then $\mathbb{E}_x (b_j(x) - y(x))^2 = \mathbb{E}_x \varepsilon_j^2(x).$

The mean error of N models: $$E_1 = \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}_x \varepsilon_j^2(x).$$

Bootstrap

Consider the errors unbiased and uncorrelated:

$$\mathbb{E}_x \varepsilon_j(x) = 0;$$
$$\mathbb{E}_x \varepsilon_i(x) \varepsilon_j(x) = 0, \quad i \neq j.$$

$$E_N = \mathbb{E}_x \left(\frac{1}{N} \sum_{j=1}^{n} b_j(x) - y(x) \right)^2 =$$
$$= \mathbb{E}_x \left(\frac{1}{N} \sum_{j=1}^{N} \varepsilon_j(x) \right)^2 =$$
$$= \frac{1}{N^2} \mathbb{E}_x \left(\sum_{j=1}^{N} \varepsilon_j^2(x) + \underbrace{\sum_{i \neq j} \varepsilon_i(x) \varepsilon_j(x)}_{=0} \right) =$$
$$\boxed{= \frac{1}{N} E_1.}$$

The final model averages all predictions:

$$a(x) = \frac{1}{N} \sum_{j=1}^{N} b_j(x).$$

Error decreased by N times!

Bootstrap

Consider the errors ~~unbiased and uncorrelated~~:

This is a lie

$$\mathbb{E}_x \varepsilon_j(x) = 0;$$
$$\mathbb{E}_x \varepsilon_i(x)\varepsilon_j(x) = 0, \quad i \neq j.$$

The final model averages all predictions:

$$a(x) = \frac{1}{N}\sum_{j=1}^{N} b_j(x).$$

$$E_N = \mathbb{E}_x\left(\frac{1}{N}\sum_{j=1}^{n} b_j(x) - y(x)\right)^2 =$$
$$= \mathbb{E}_x\left(\frac{1}{N}\sum_{j=1}^{N} \varepsilon_j(x)\right)^2 =$$
$$= \frac{1}{N^2}\mathbb{E}_x\left(\sum_{j=1}^{N} \varepsilon_j^2(x) + \underbrace{\sum_{i\neq j} \varepsilon_i(x)\varepsilon_j(x)}_{=0}\right) =$$
$$\boxed{= \frac{1}{N}E_1.}$$

Error decreased by N times!

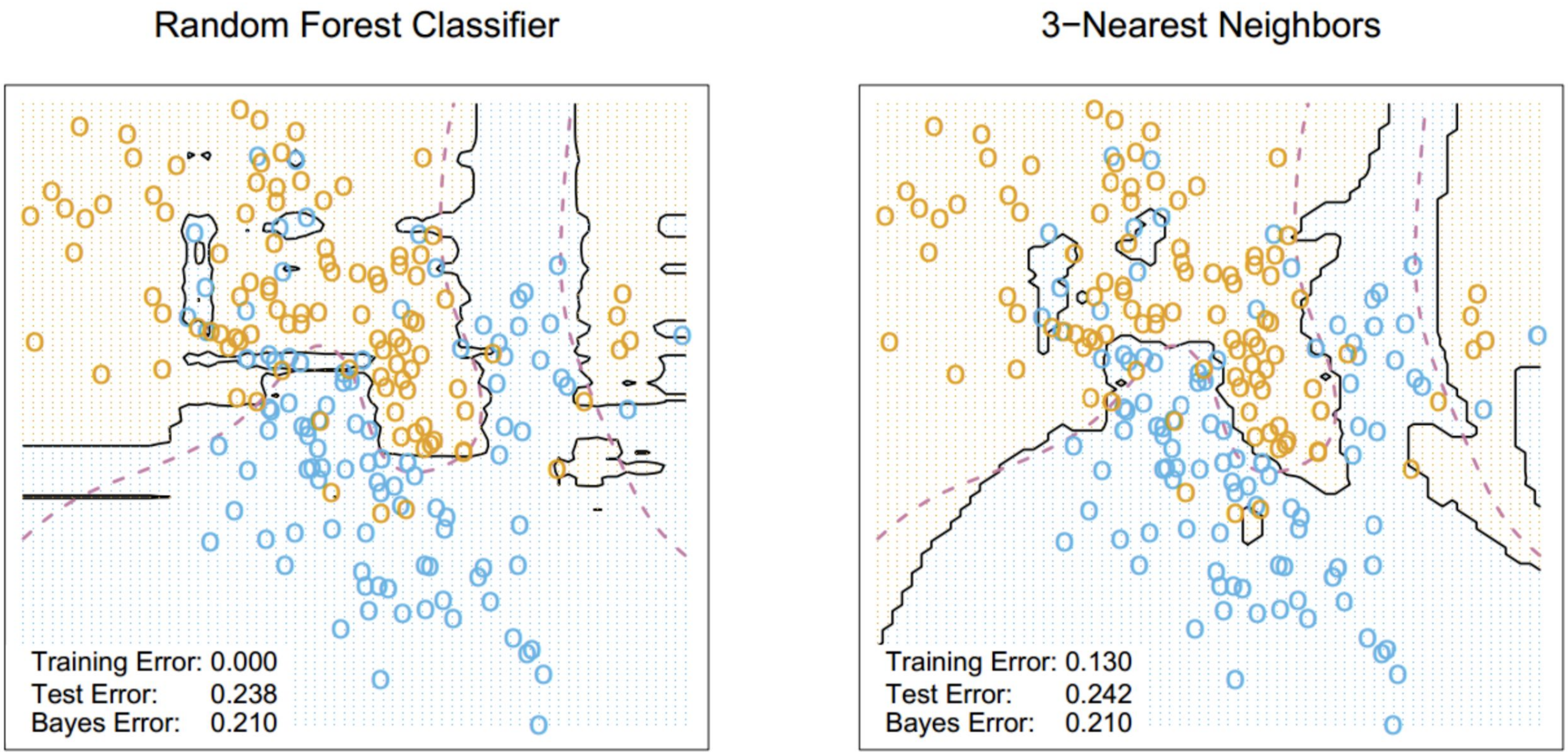
Bagging = Bootstrap aggregating

Decreases the variance if the basic algorithms are not correlated.

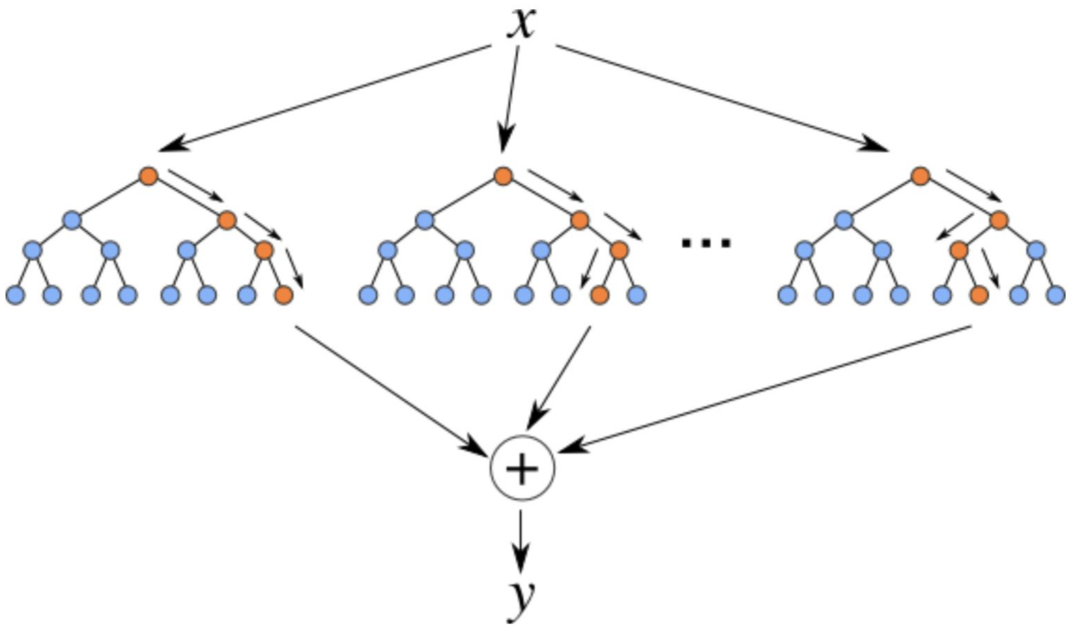
Random Forest

RSM - Random Subspace Method

Same approach, but with features.


Random Forest Classifier
Training Error: 0.000
Test Error: 0.238
Bayes Error: 0.210
3−Nearest Neighbors
Training Error: 0.130
Test Error: 0.242
Bayes Error: 0.210

Bagging + RSM = Random Forest

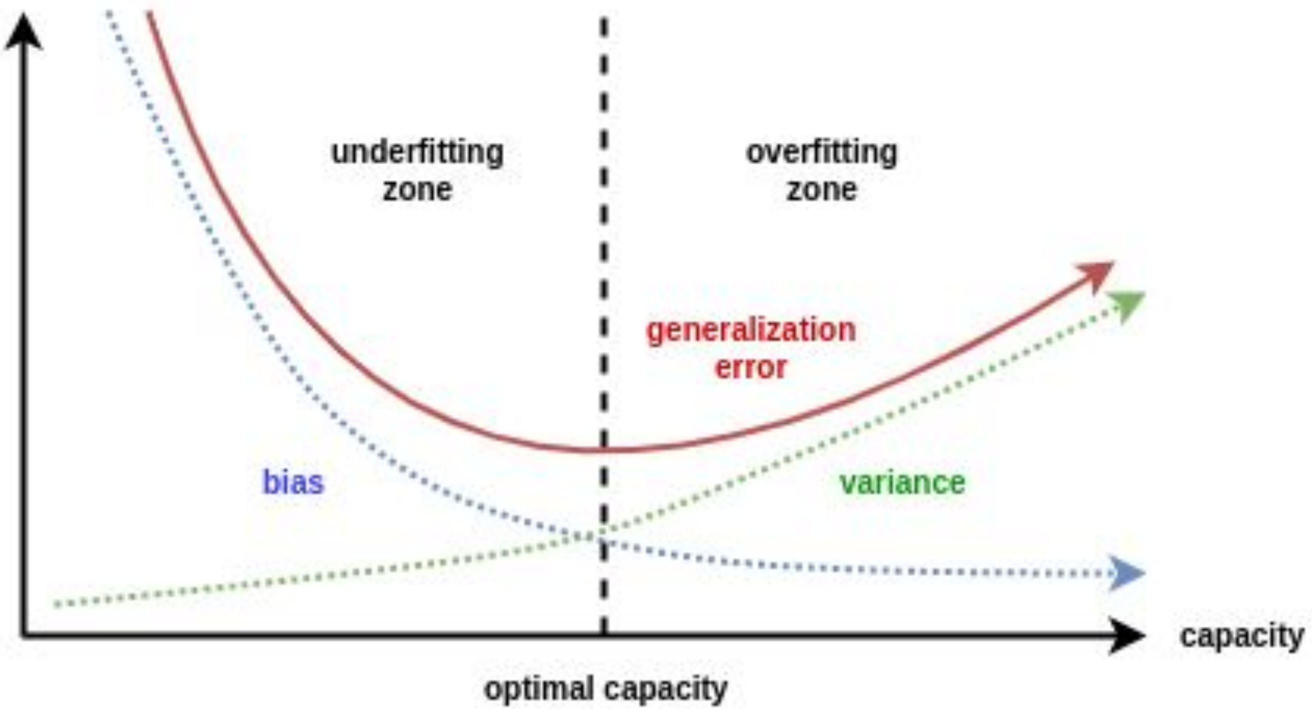


Random Forest

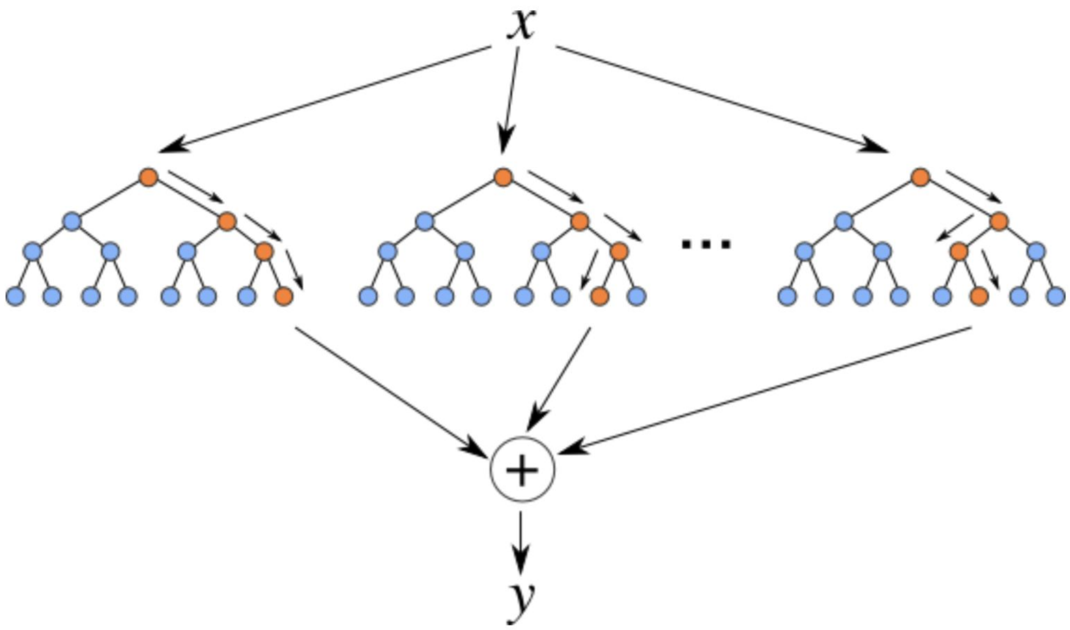
- One of the greatest “universal” models.
- There are some modifications: Extremely Randomized Trees, Isolation Forest, etc.
- Allows to use train data for validation: OOB

$$\text{OOB} = \sum_{i=1}^{\ell} L\left(y_i, \frac{1}{\sum_{n=1}^{N}[x_i \notin X_n]} \sum_{n=1}^{N}[x_i \notin X_n] b_n(x_i)\right)$$

Bias-variance tradeoff



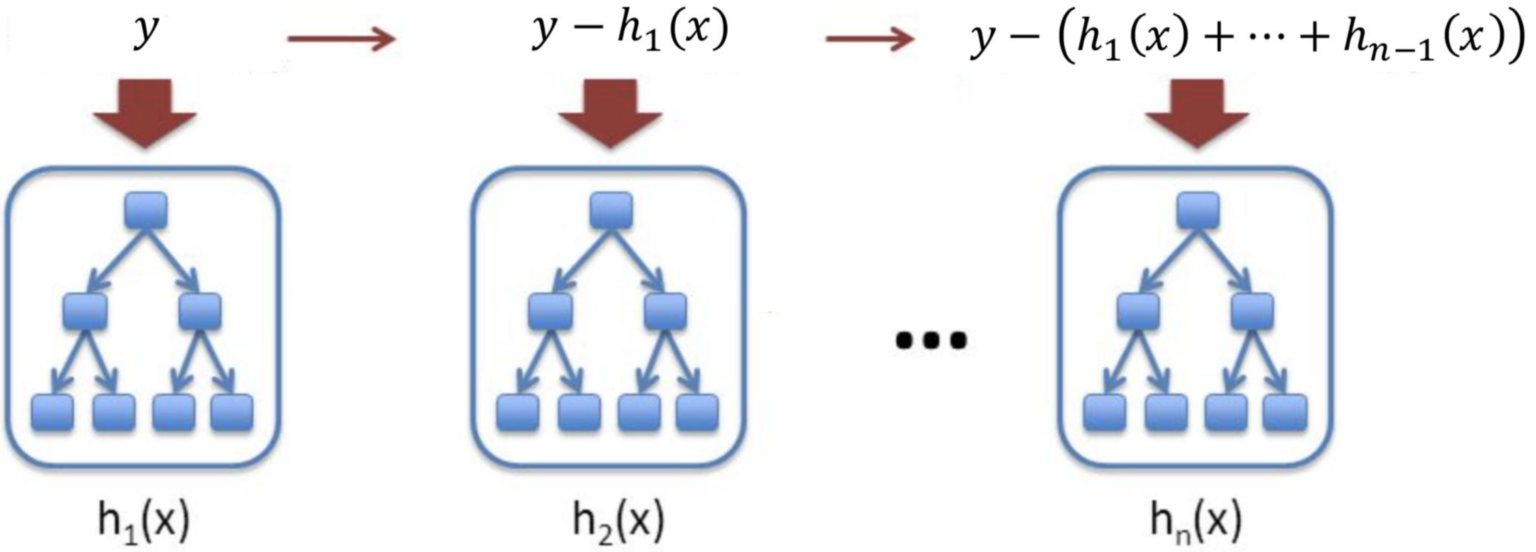
Is Random Forest decreasing bias or variance by building the trees ensemble?



Boosting

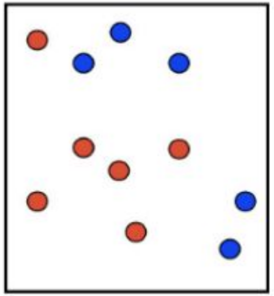
Boosting

$$a_n(x) = h_1(x) + \cdots + h_n(x)$$



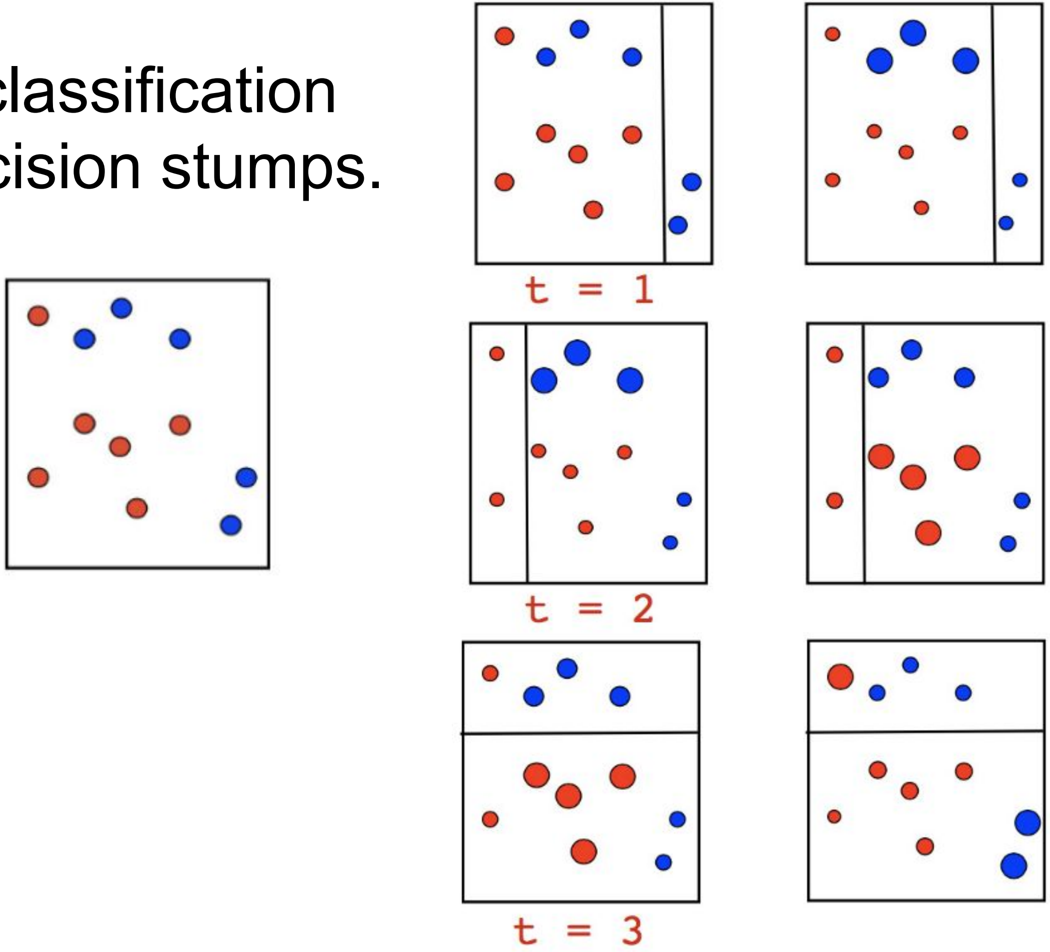
Boosting: intuition

Binary classification
Use decision stumps.



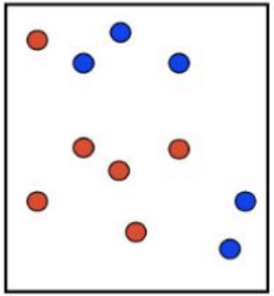
Boosting: intuition

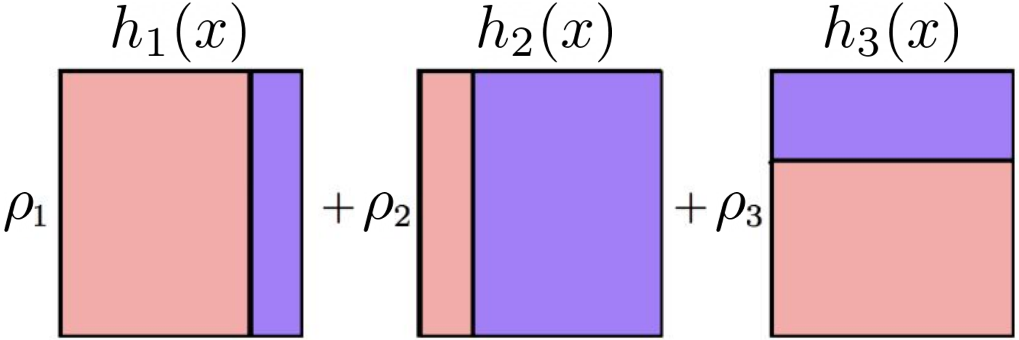
Binary classification
Use decision stumps.



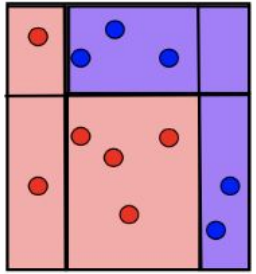
Boosting: intuition

Binary classification
Use decision stumps.

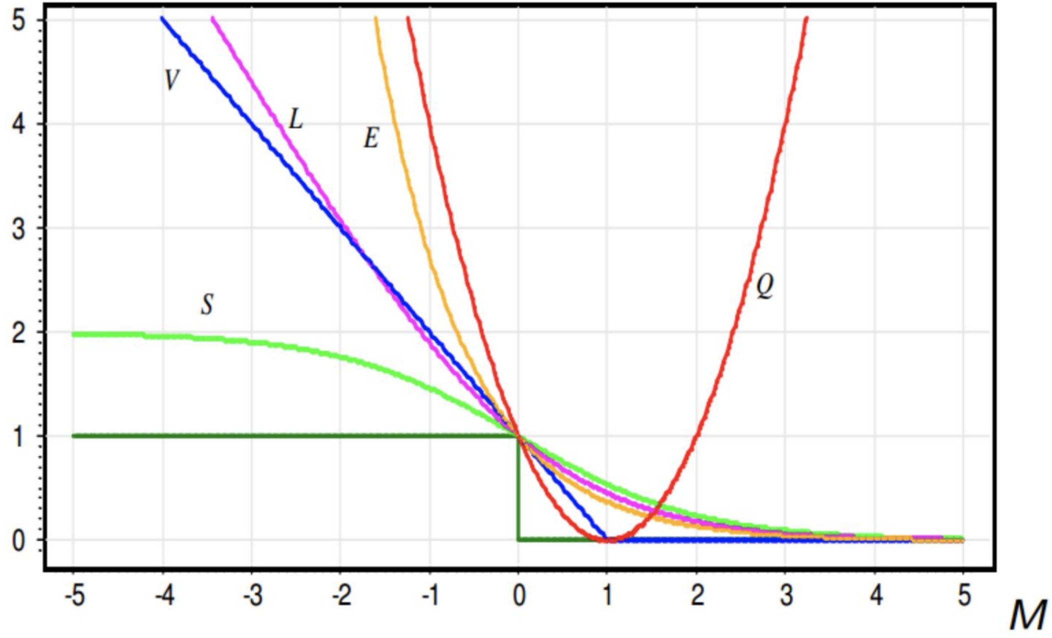




$$\hat{f}_T(x) = \sum_{t=1}^{T} \rho_t h_t(x) \quad =$$

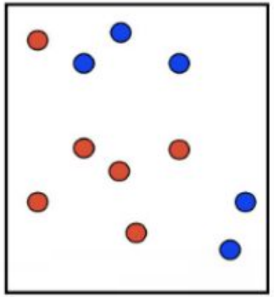


Recap: loss functions for classification



$$Q(M) = (1 - M)^2$$
$$V(M) = (1 - M)_+$$
$$S(M) = 2(1 + e^M)^{-1}$$
$$L(M) = \log_2(1 + e^{-M})$$
$$\boxed{E(M) = e^{-M}}$$

Boosting: AdaBoost



$$\hat{f}_T(x) = \sum_{t=1}^{T} \rho_t h_t(x)$$

$$L(y_i, \hat{f}_T(x_i)) = \exp(-y_i \hat{f}_T(x_i)) = \exp(-y_i \sum_{t=1}^{T} \rho_t h_t(x_i))$$

$$= \boxed{\exp\Big(- y_i \sum_{t=1}^{T-1} \rho_t h_t(x_i)\Big)} \cdot \exp(-y_i \rho_T h_T(x_i))$$

const on step T

$$= w_i \cdot \exp(-y_i \rho_T h_T(x_i))$$

Boosting: intuition

Binary classification
Use decision stumps.