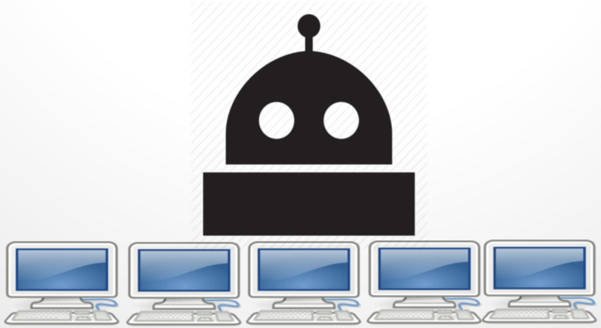
Boosting:

A weighted crowd of narrowminded experts

Aaron & Dylan



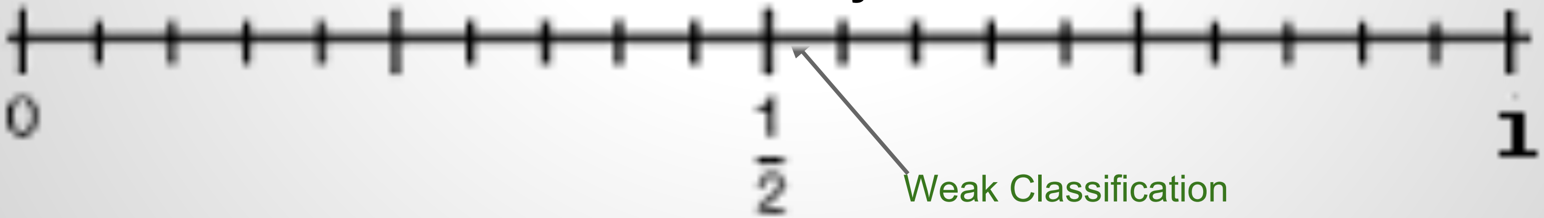
Boosting Hypothesis (Kearne, Valiant; 1988-89)

We can make a strong classifier (arbitrarily well at classification) from a collection of weak classifiers (somewhat better than random guess).

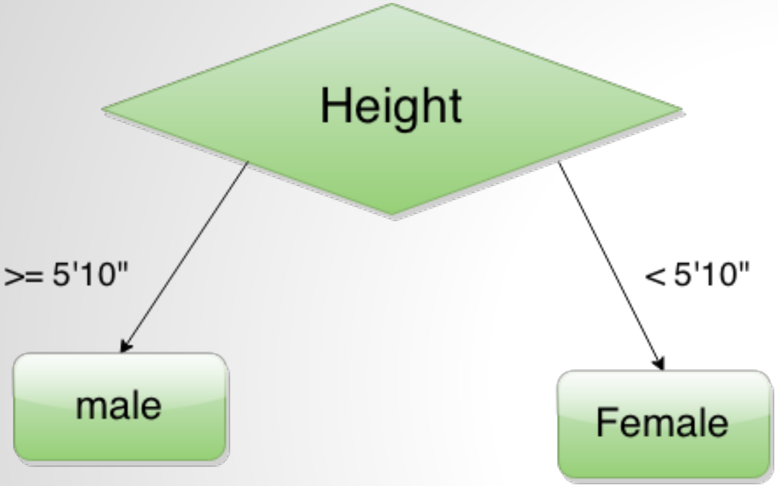
Weak Classifiers

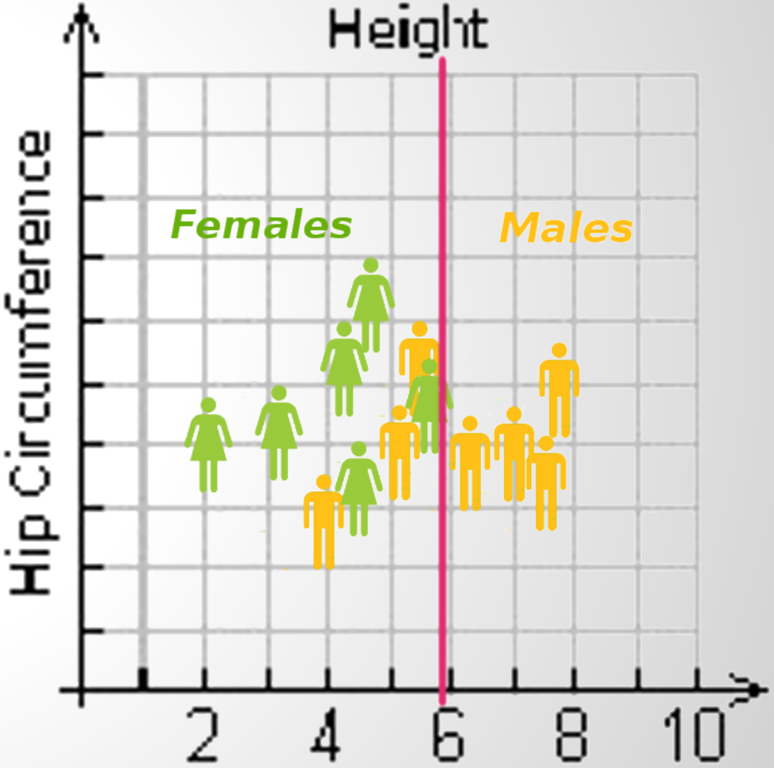
- Classifier which may be only slightly correlated with true classification (accuracy > 50%)
- Examples: Naïve Bayes, logistic regression, decision stumps

Accuracy



Decision Stumps

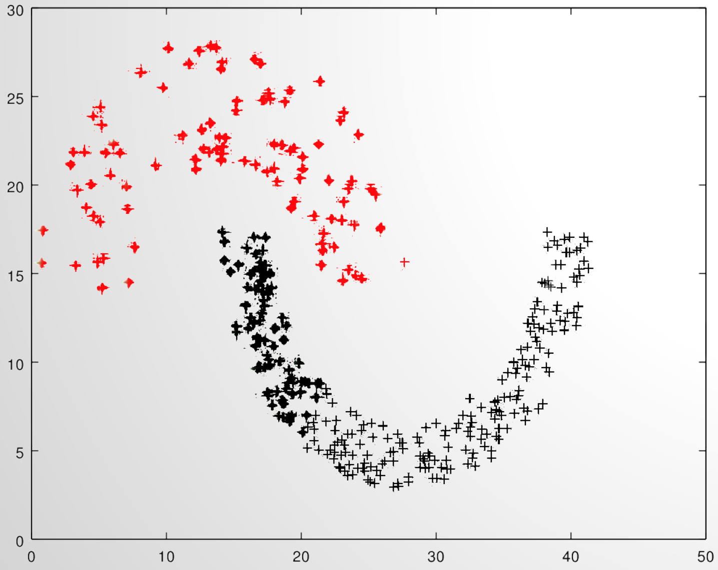


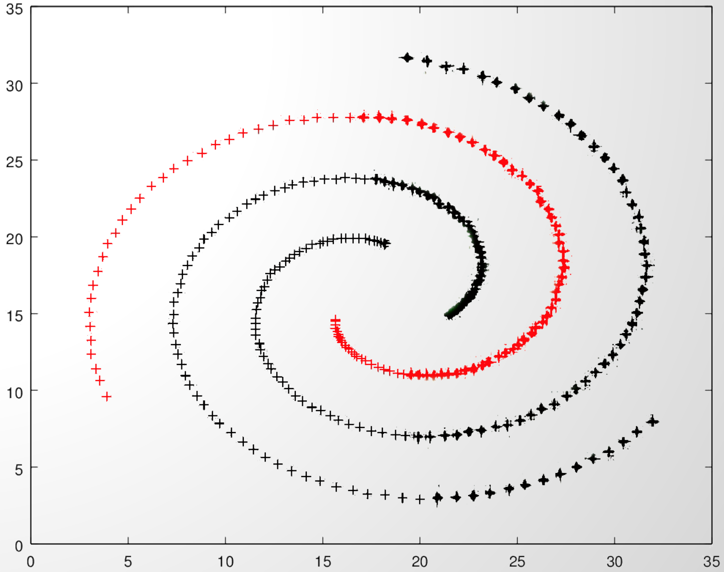


- Single Level Decision Tree
- Focus on a single feature dimension
- Create a decision boundary along that dimension

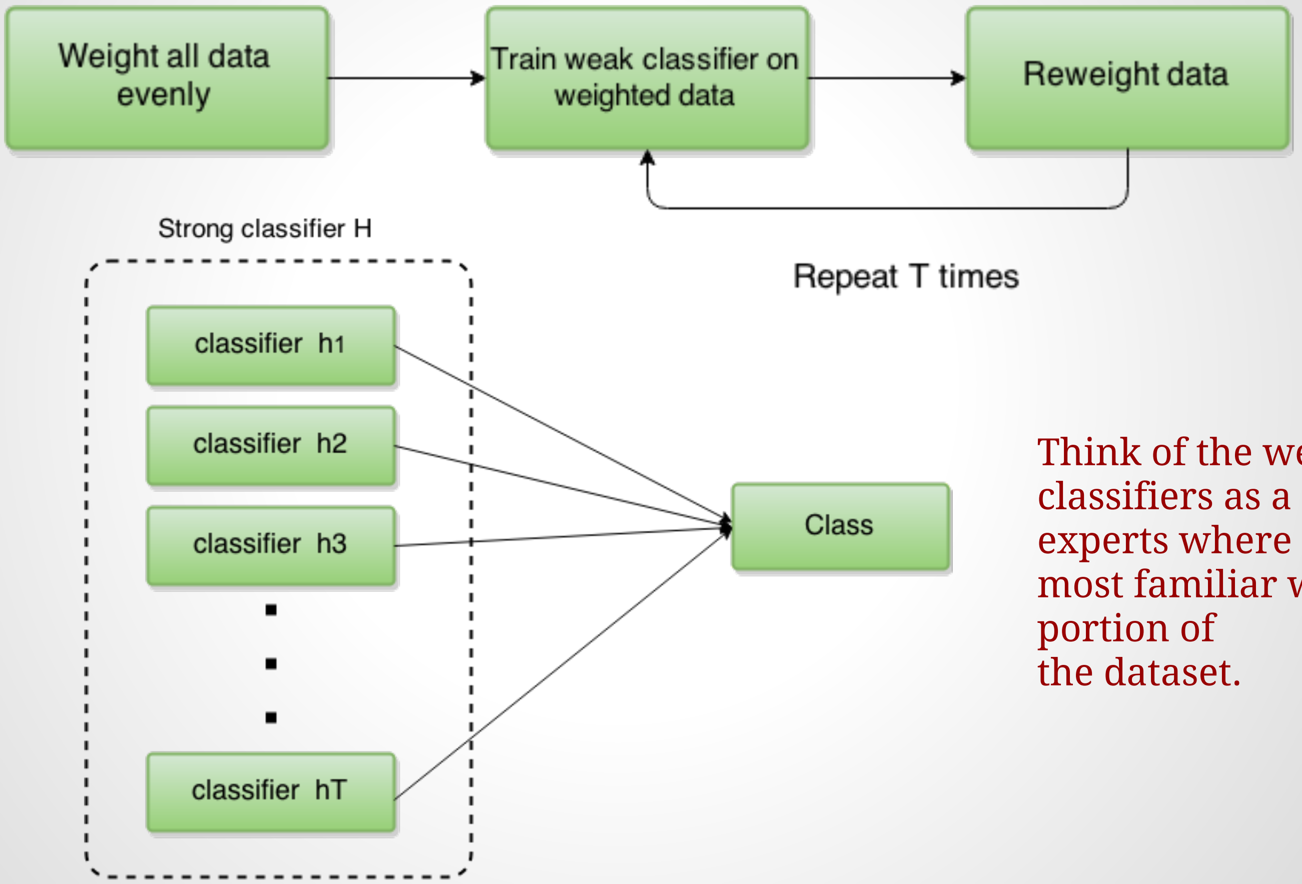
Advantages of Boosting:

- Easy and fast to train weak classifiers
- Simple models don't usually overfit
- Weak classifiers can not solve hard problems





Boosting: The Basic Idea



Think of the weak classifiers as a crowd of experts where each is most familiar with some portion of the dataset.

AdaBoost: Boosting for Binary Classification

Suppose dataset: $(x_1, y_1), ..., (x_N, y_N)$

where $x_i \in \mathbb{R}^n, y_i \in Y = \{-1, 1\}$

Let $D_t(i) =$ **weight of point** x_i

Goal: Build classifer $H(x) = \textbf{sign}(\alpha_1 h_1(x)+, ..., +\alpha_T h_T(x))$

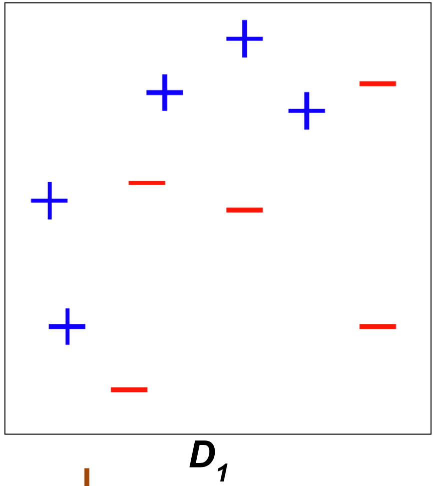
where $h_1(x), ..., h_T(x)$ **are binary classifiers, built on distributions** $D_1, ..., D_T$ **respectively.**

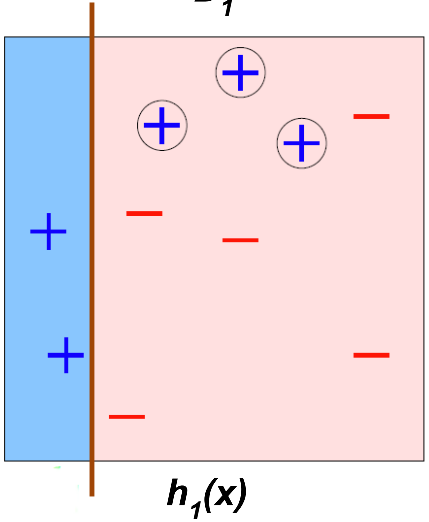
Issue: How to find the best α's and D's.

Answer: Iteratively minimize exponential loss:

If $F(x) = \alpha_1 h_1(x)+, ..., +\alpha_T h_T(x)$, then

$$L = \frac{1}{N} \sum_{i=1}^{T} \exp(-y_i F(x_i))$$





AdaBoost with Decision Stumps as Weak Classifiers

(Shapire, Freund. 1999)

Round One:

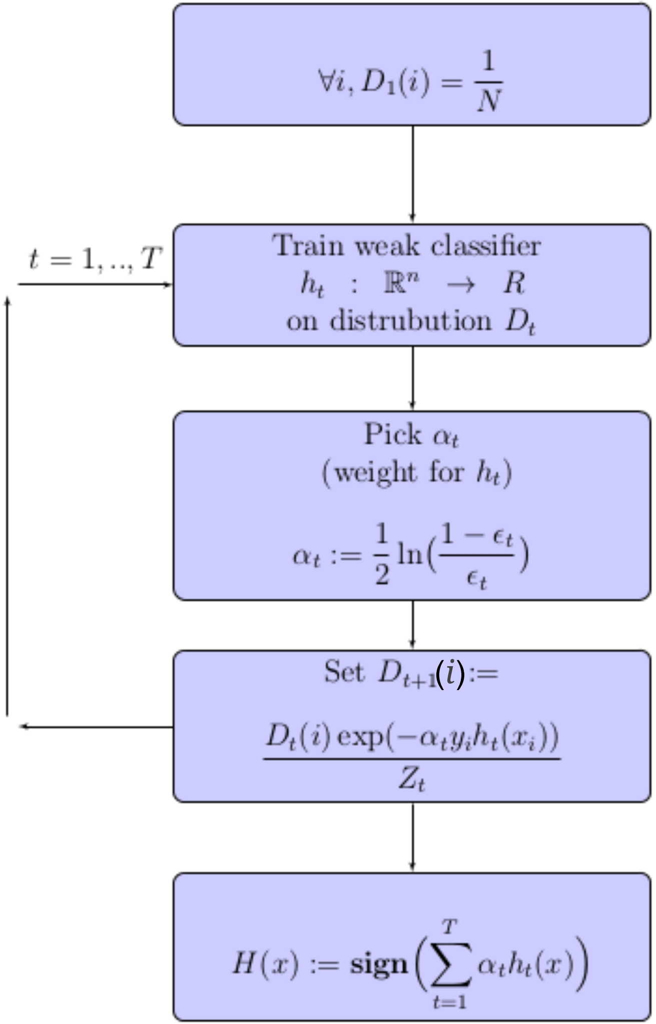
Build h_1 on distribution D_1

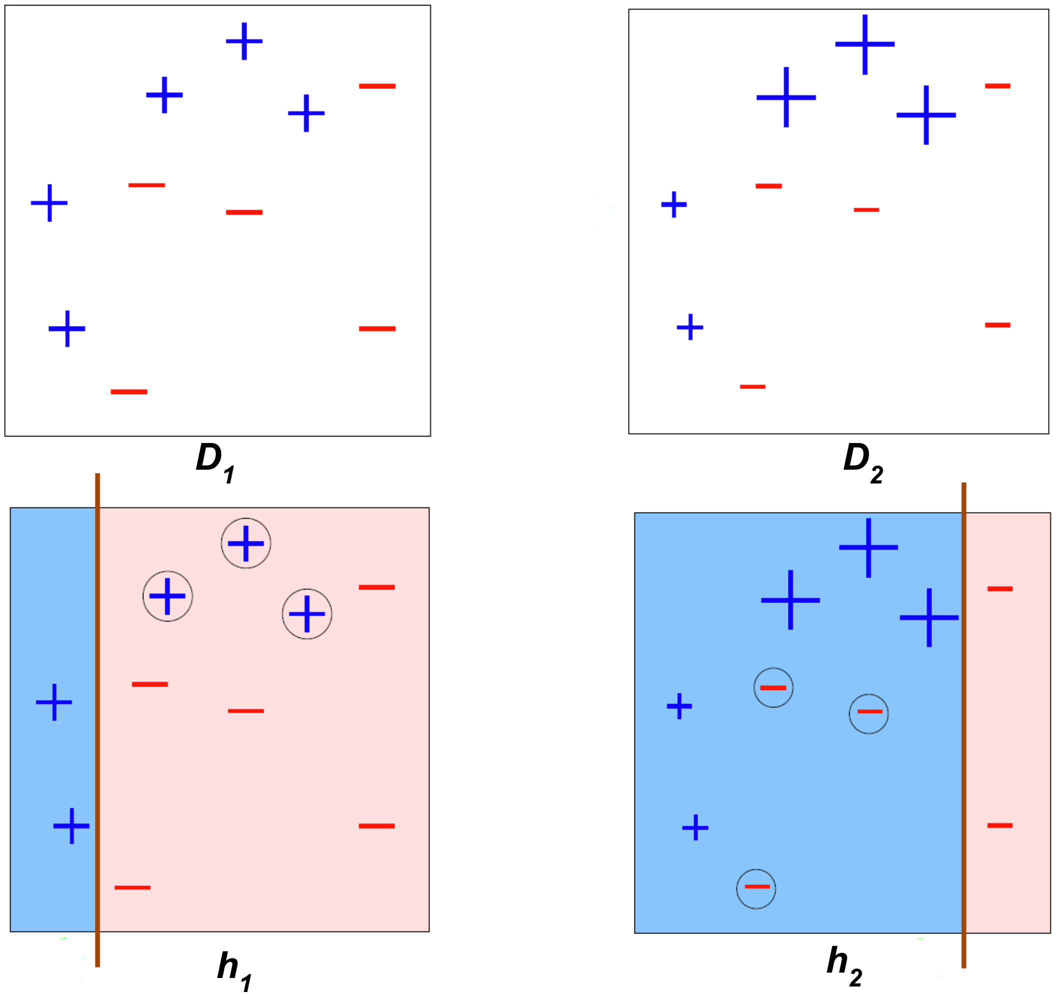
Then calculate:

$\varepsilon_1 = Pr_{i \sim Dt}(h_1(x) \neq y_i).$
(sum of misclassified point weights)

Next calculate α_1.

Then calculate D_2.





Round One:

Build h_1 on distribution D_1

$\varepsilon_1 = 3/10$

$\alpha_1 = 0.42$

$D_2(i) = 0.166$ for x_i that were misclassified

$D_2(i) = 0.072$ for x_i that were correctly classified

Round Two:

Build h_2 on distribution D_2

$\varepsilon_2 = 0.216$

$\alpha_2 = 0.65$

$\varepsilon_2 = 0.216$, $\alpha_2 = 0.65$

For each X_i where:

h_1 was wrong, h_2 was right:

$D_3(i) = 0.11$, $D_2(i) = 0.166$

h_1 was right, h_2 was wrong:

$D_3(i) = 0.175$, $D_2(i) = 0.072$

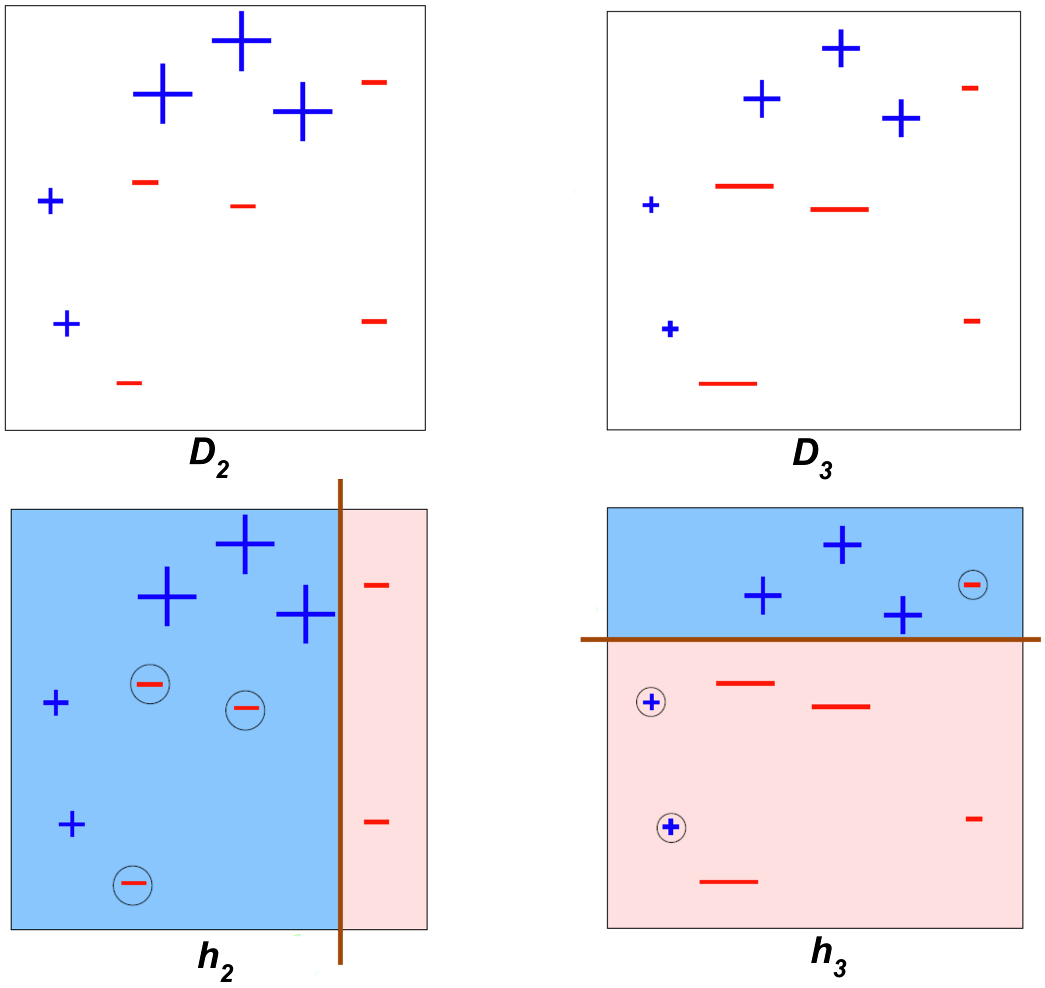
h_1 was right, h_2 was right:

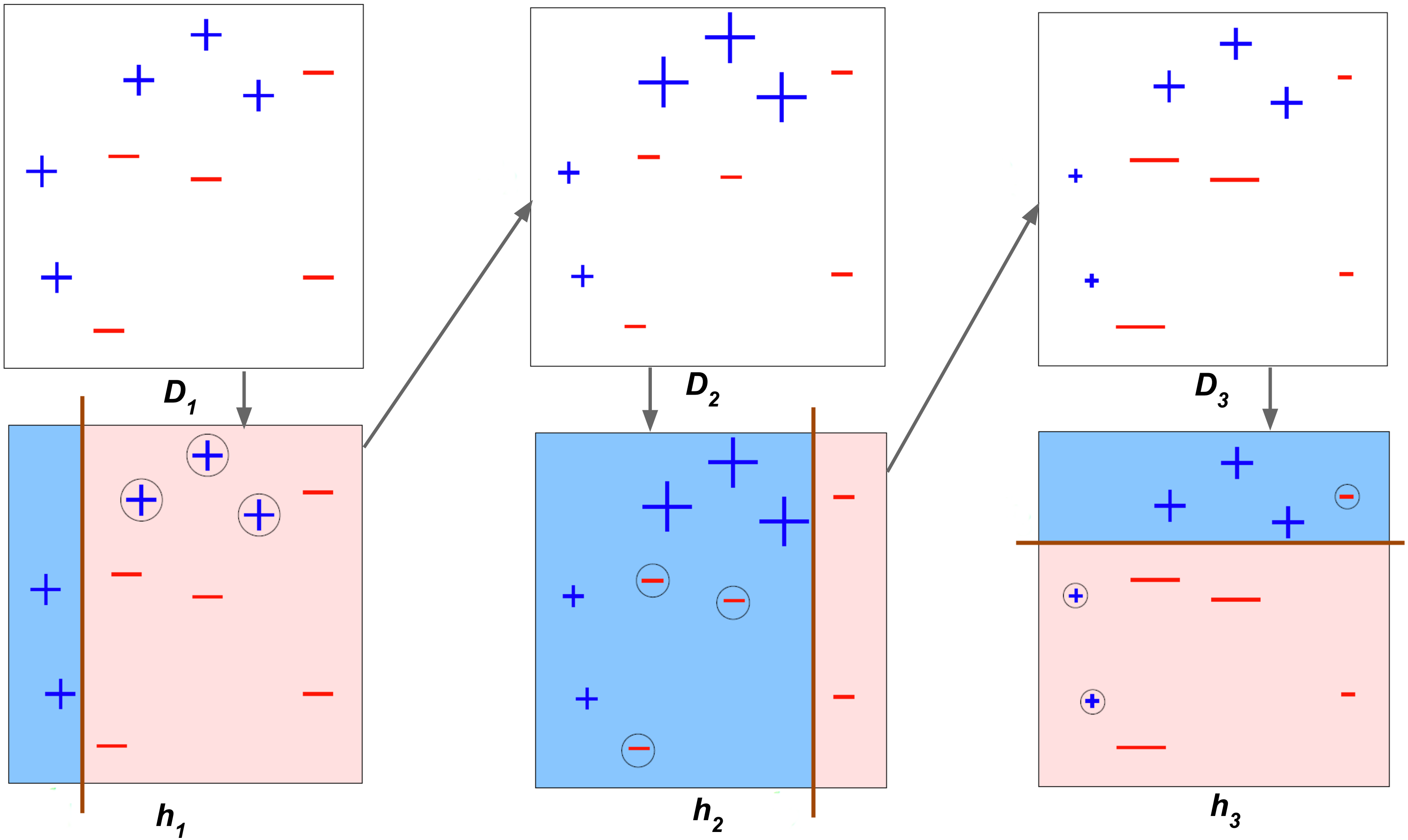
$D_3(i) = 0.047$ $D_2(i) = 0.072$

Round Three:

Train h_3 on D_3

$\varepsilon_3 = 0.144$, $\alpha_3 = 0.91$

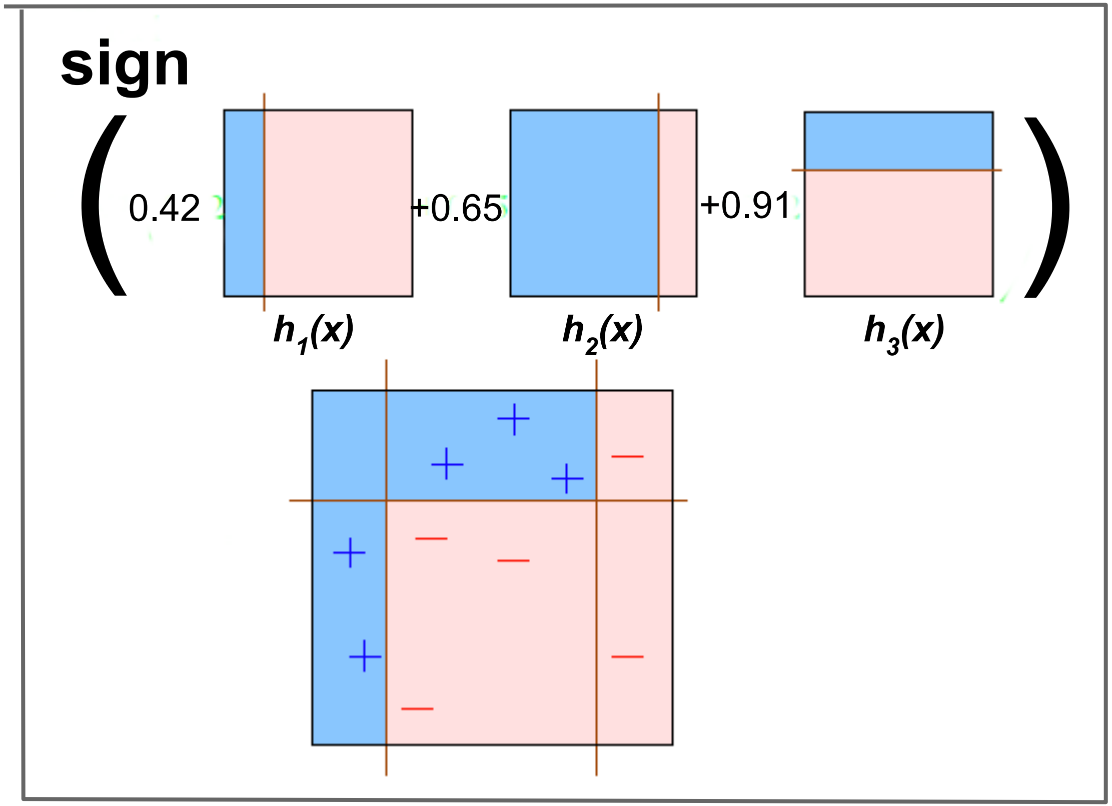


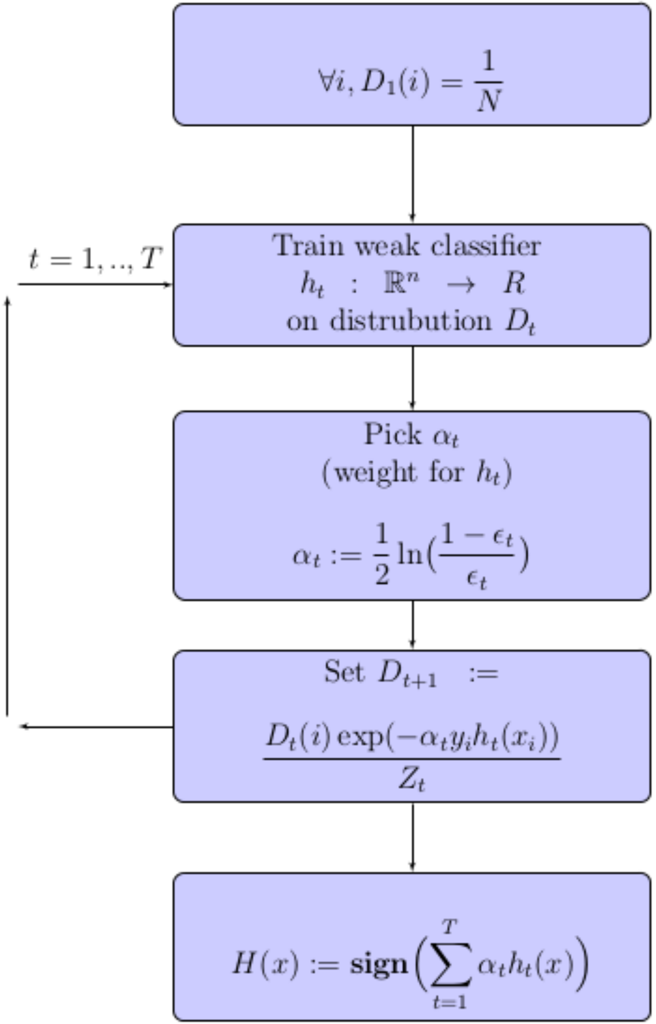


D_1
D_2
D_3
h_1
h_2
h_3

Strong Classifier

H(x)





Boosting Demos

Swirly boosting demo

More Swirly boosting demo

AdaBoost in acton

References:

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification* (pp. 149-171). Springer New York.

Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, *5*(2), 197-227.

Kearns, M. (1988). Thoughts on hypothesis boosting. *Unpublished manuscript*, *45*, 105.

Long, P. M., & Servedio, R. A. (2010). Random classification noise defeats all convex potential boosters. *Machine Learning*, *78*(3), 287-304.

Freund, Y., & Schapire, R. E. (1995, January). A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory* (pp. 23-37). Springer Berlin Heidelberg.

MIT Boosting Lecture

Software:

Wikipedia list from AdaBoost page

Boosting Song