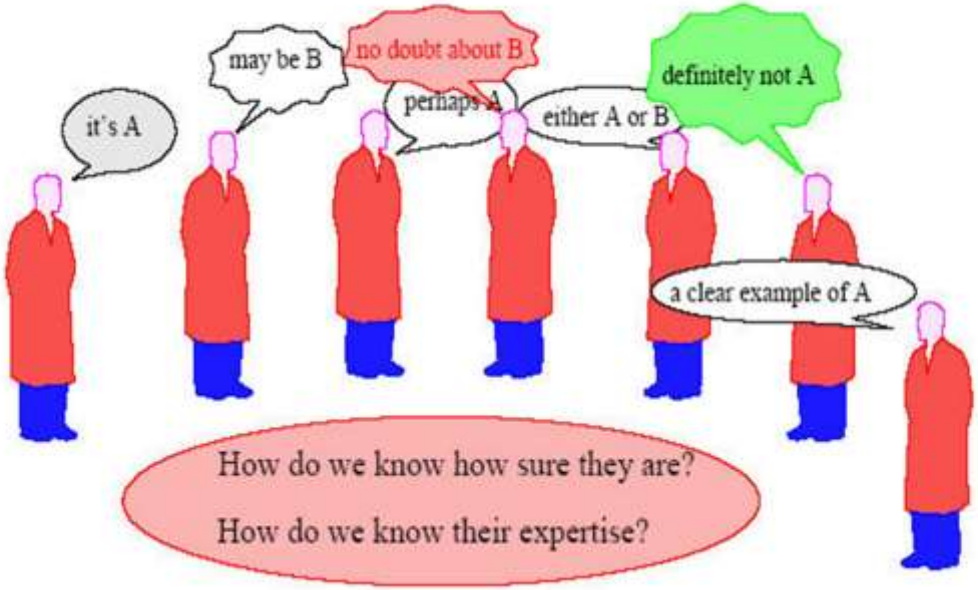
IN THE NAME OF ALLAH

Neural Networks

Classifier Combination

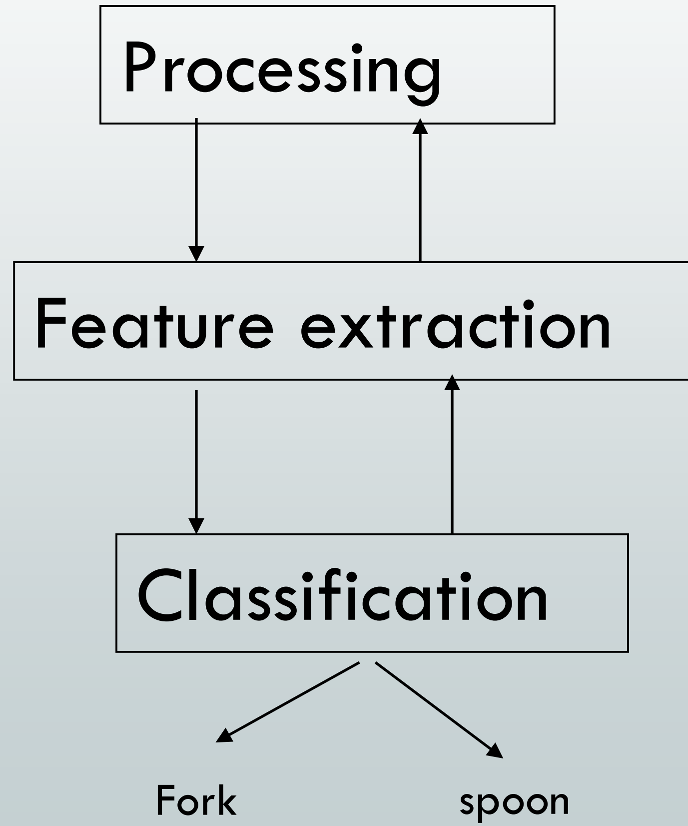


Shahrood University of Technology
Hossein Khosravi

Objectives

- An introductive tutorial on multiple classifier combination
- Motivation and basic concepts
- Main methods for creating multiple classifiers
- Main methods for fusing multiple classifiers
- Applications, achievement, open issues and conclusion

Pattern Classification



Traditional approach to pattern classification

- Unfortunately, no dominant classifier exists for all the data distributions, and the data distribution of the task at hand is usually unknown
- Not one classifier can be discriminative well enough if the number of classes are huge
- For applications where the objects/classes of content are numerous, unlimited, unpredictable, one specific classifier/detector cannot solve the problem.

Combine individual classifiers

- Fusion of multiple classifiers can improve the performance of the best individual classifiers
- This is possible if individual classifiers make different errors
- For linear combiners, Turner and Ghosh (1996) showed that averaging outputs of individual classifiers with unbiased and uncorrelated errors can improve the performance of the best individual classifier

Definitions

- A “classifier” is any mapping from the space of features(measurements) to a space of class labels (names, tags, distances, probabilities)
- A classifier is a hypothesis about the real relation between features and class labels
- A “learning algorithm” is a method to construct hypotheses
- A learning algorithm applied to a set of samples (training set) outputs a classifier

Definitions

- A multiple classifier system (MCS) is a structured way to combine (exploit) the outputs of individual classifiers
- MCS can be thought as:
  - Multiple expert systems
  - Committees of experts
  - Mixtures of experts
  - Classifier ensembles
  - Composite classifier systems

Basic concepts

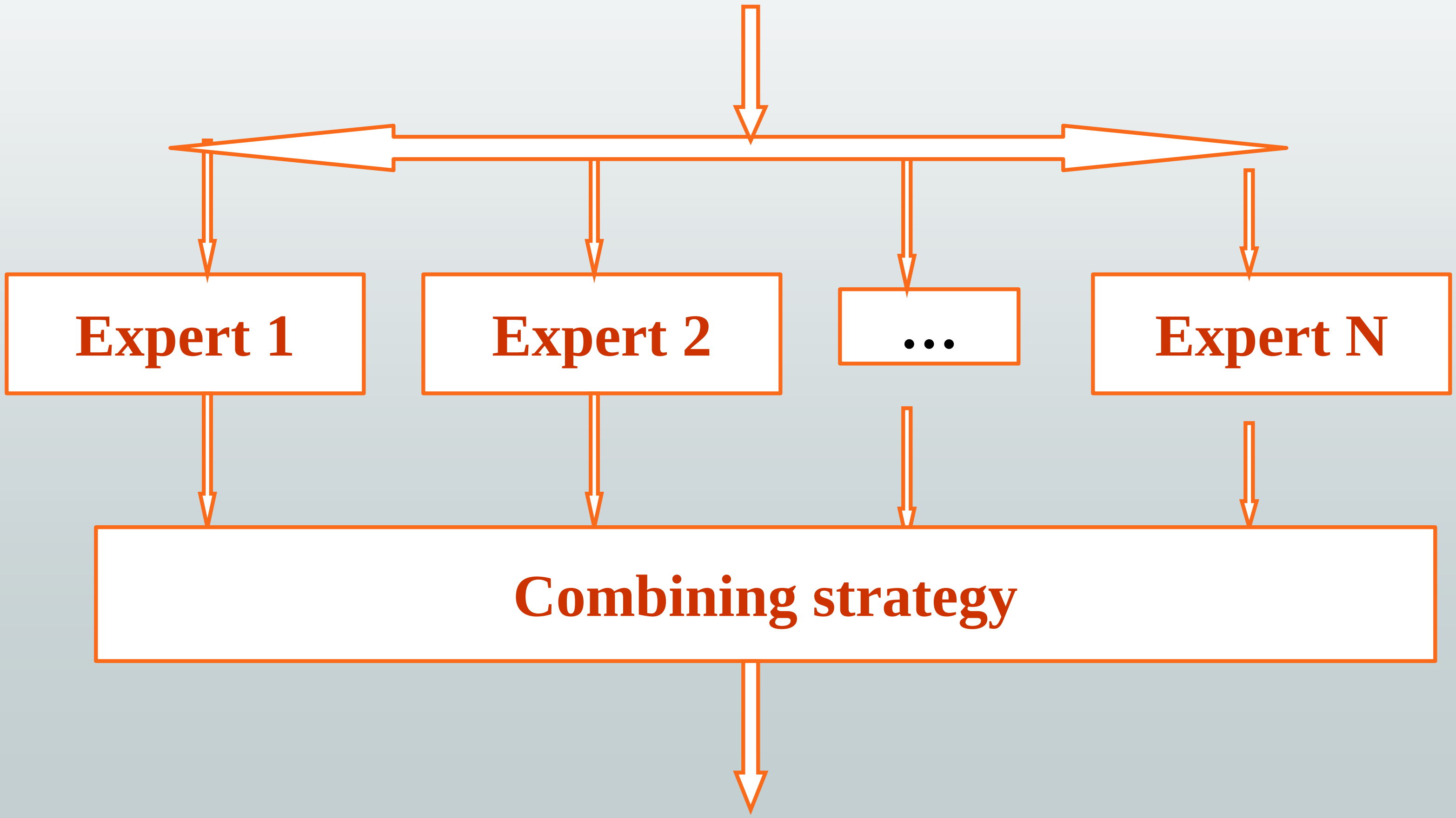
- Multiple Classifier Systems (MCS) can be characterized by:
  - The Architecture
  - Fixed/Trained Combination strategy
  - Others

MCS Architecture/Topology

- Serial

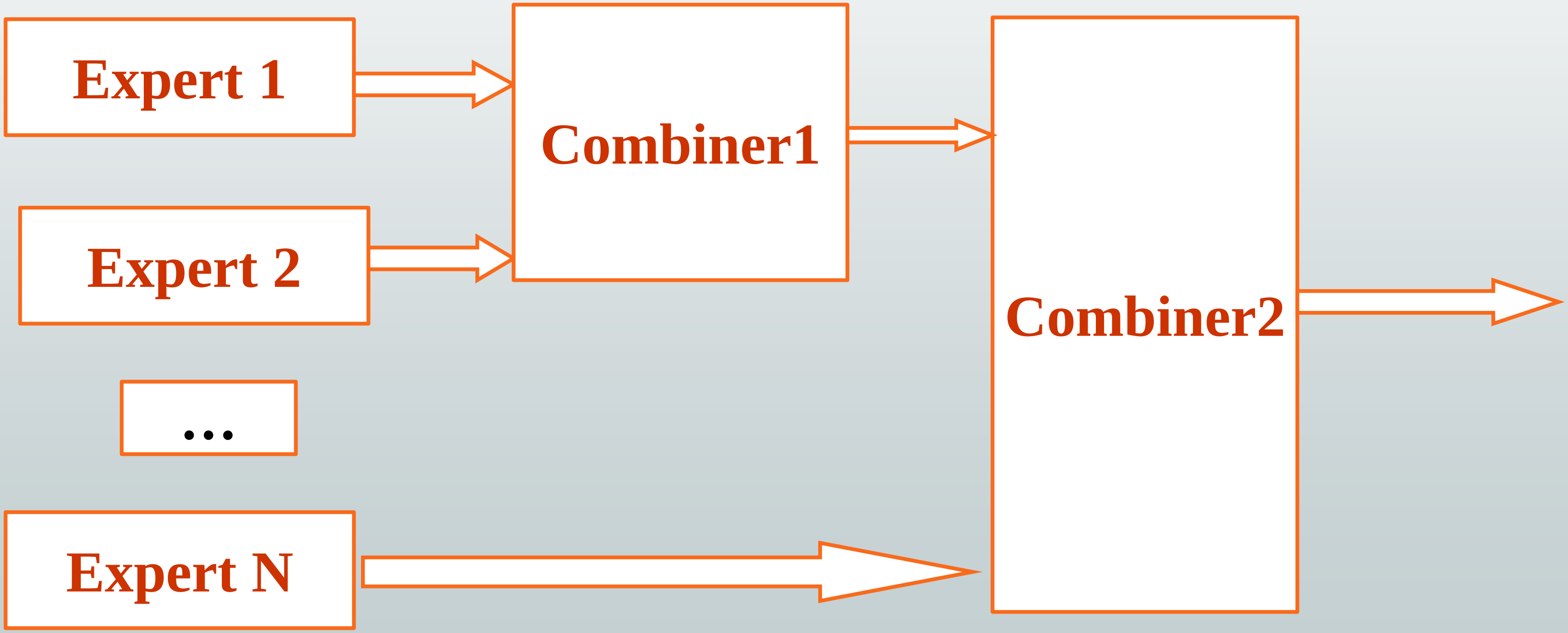
MCS Architecture/Topology

- Parallel



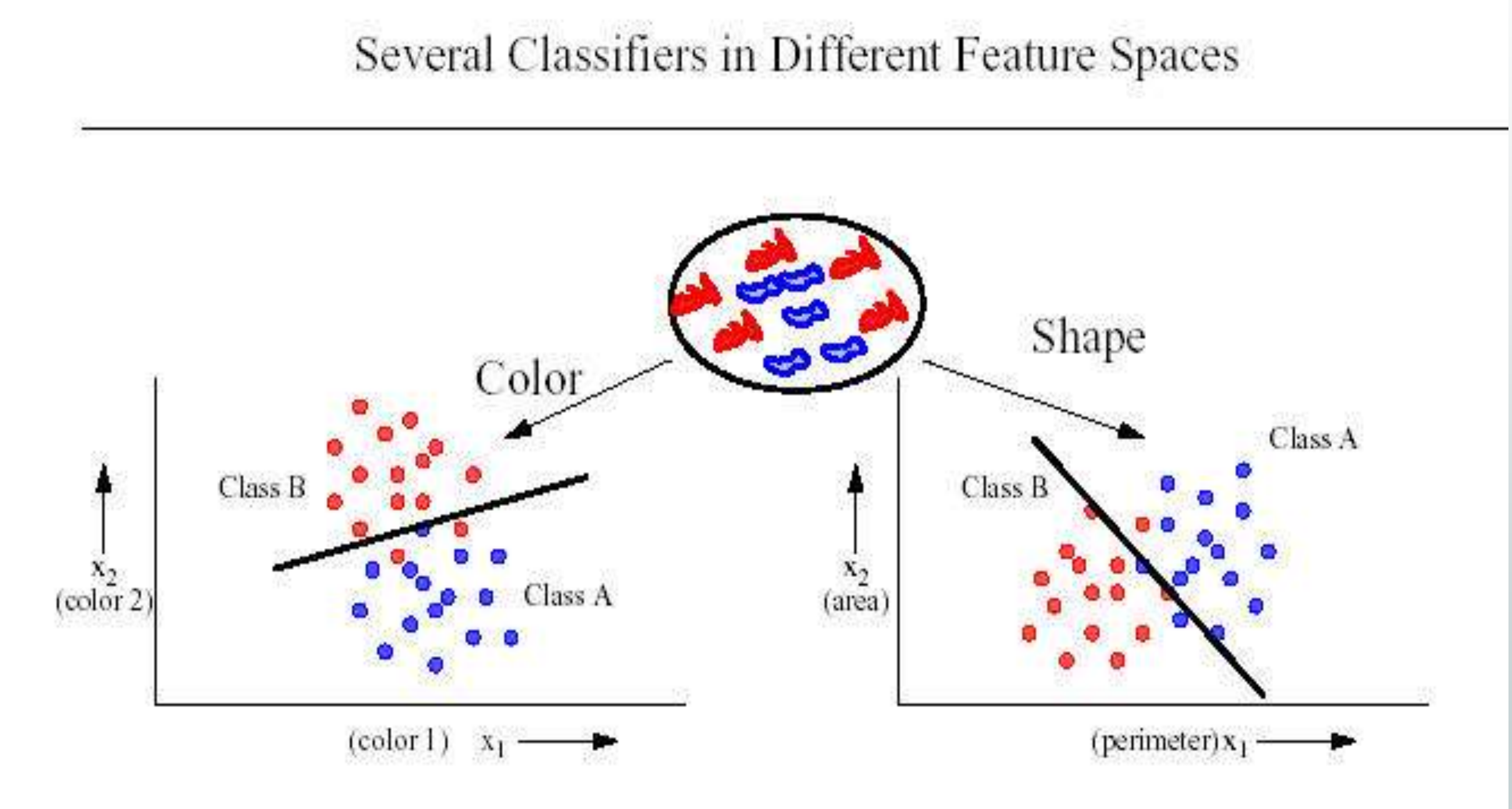
MCS Architecture/Topology

- Hybrid

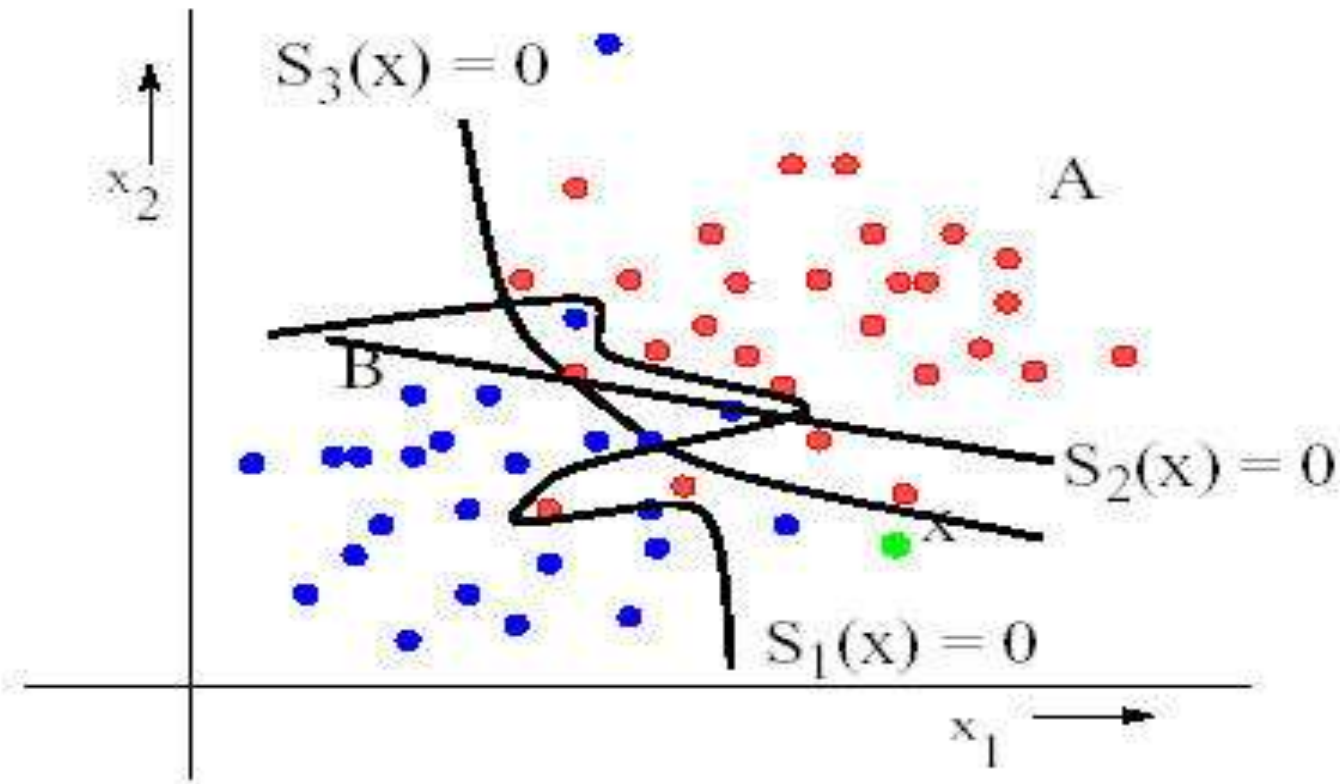


Multiple Classifiers Sources?

- Different feature spaces: face, voice fingerprint;

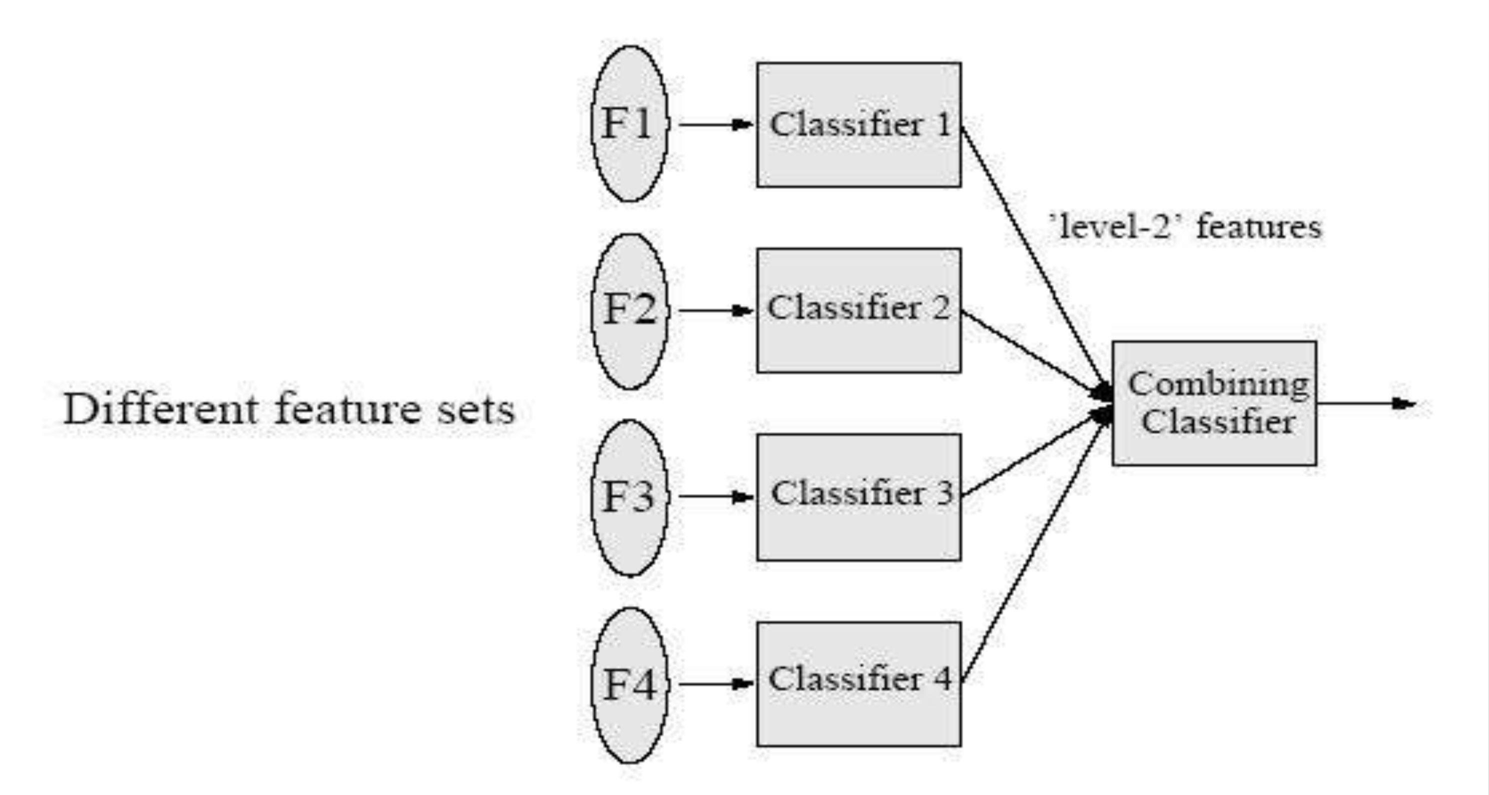


Multiple Classifiers Sources?

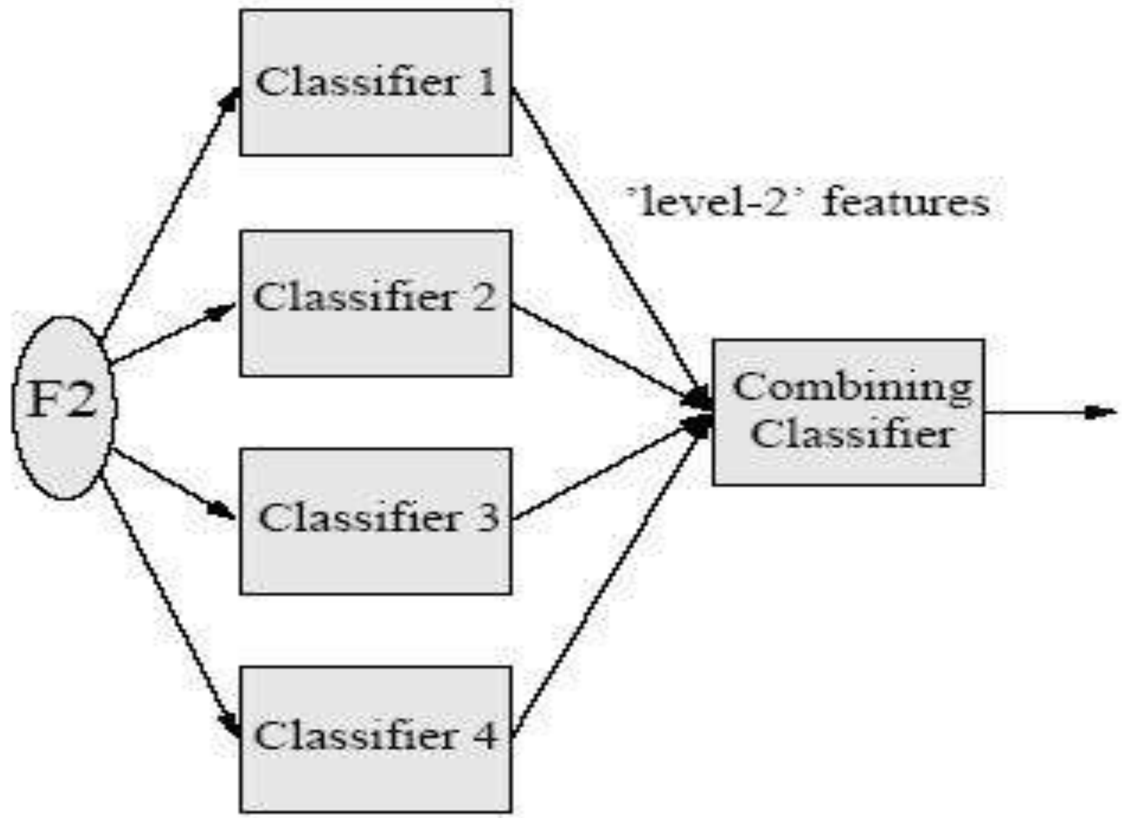


Same feature space, three classifiers demonstrate different performance

Combination based on different feature spaces

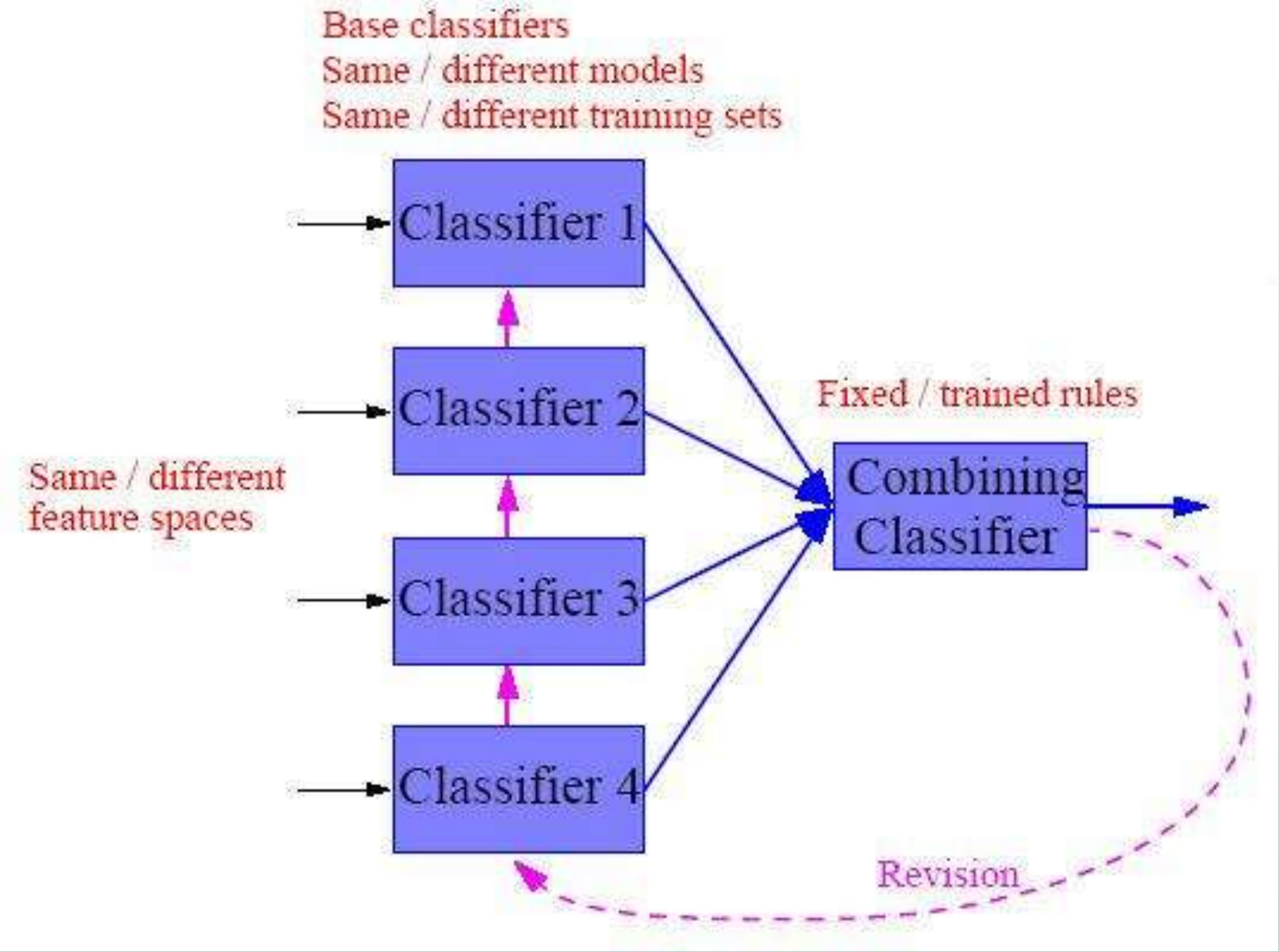


Combining based on a single space but different classifiers



Single feature set, different classifiers

Architecture of multiple classifier combination



Remarks on fixed and trained combination strategies

- Fixed rules
  - Simplicity
  - Low memory and time requirements
  - Well-suited for ensembles of classifiers with independent/low correlated errors and similar performances
- Trained rules
  - Flexibility: potentially better performances than fixed rules
  - Trained rules are claimed to be more suitable than fixed ones for classifiers correlated or exhibiting different performances
  - High memory and time requirements

Bagging (Bootstrap Aggregation)

- Training set $D=\{(x_1,y_1),\ldots,(x_N,y_N)\}$
- Sample S sets of N elements from D (with replacement): $D_1, D_2, \ldots, D_S$
- Train on each D_s, s=1,..,S and obtain a sequence of S outputs $f_1(X),..,f_S(X)$
- The final classifier is:

$$\bar{f}(X) = \sum_{s=1}^{S} f_s(X)$$

Regression

$$\bar{f}(X) = \theta\left(\sum_{s=1}^{S} \mathrm{sign}(f_s(X))\right)$$

Classification

AdaBoost.M1

Input: sequence of m examples $\langle(x_1, y_1), \ldots, (x_m, y_m)\rangle$ with labels $y_i \in Y = \{1, \ldots, k\}$
weak learning algorithm **WeakLearn**
integer T specifying number of iterations

Initialize $D_1(i) = 1/m$ for all i.
Do for $t = 1, 2, \ldots, T$

1. Call **WeakLearn**, providing it with the distribution D_t.
2. Get back a hypothesis $h_t : X \rightarrow Y$.
3. Calculate the error of h_t: $\epsilon_t = \sum_{i:h_t(x_i) \neq y_i} D_t(i)$. If $\epsilon_t > 1/2$, then set $T = t - 1$ and abort loop
4. Set $\beta_t = \epsilon_t/(1 - \epsilon_t)$.
5. Update distribution D_t: $D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} \beta_t & \text{if } h_t(x_i) = y_i \\ 1 & \text{otherwise} \end{cases}$
   where Z_t is a normalization constant (chosen so that D_{t+1} will be a distribution).

Output the final hypothesis: $h_{fin}(x) = \arg\max_{y \in Y} \sum_{t:h_t(x)=y} \log \frac{1}{\beta_t}$.

Results

نتایج اعمال الگوریتم AdaBoost.M1 روی چند ویژگی برای بازشناسی ارقام				
ویژگی	تعداد شبکه های ایجاد شده با AdaBoost.M1	نرخ بازشناسی داده های آموزش	نرخ بازشناسی داده های آزمایش	میزان بهبود بازشناسی داده های آزمایش
هیستوگرام گرادیان	۴	۹۹.۹۸	۹۸.۸۲	۰.۲۵
کریش	۳	۹۹.۷۱	۹۸.۱۲	۰.۳۷
کریش	۵	۹۹.۹۴	۹۸.۱۹	۰.۴۳
DCT81	۳	۹۹.۷۰	۹۷.۳۲	۰.۲۹
DCT81	۵	۹۹.۹۴	۹۷.۷۴	۰.۷۱

AdaBoost M2

Input: sequence of m examples $\langle(x_1, y_1), \ldots, (x_m, y_m)\rangle$ with labels $y_i \in Y = \{1, \ldots, k\}$
weak learning algorithm **WeakLearn**
integer T specifying number of iterations

Let $B = \{(i, y) : i \in \{1, \ldots, m\}, y \neq y_i\}$
Initialize $D_1(i, y) = 1/|B|$ for $(i, y) \in B$.
Do for $t = 1, 2, \ldots, T$

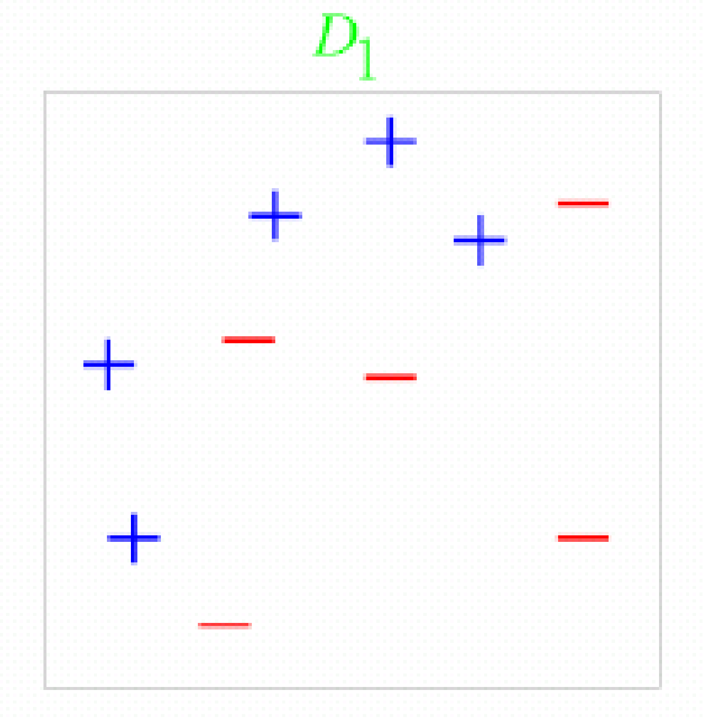
1. Call **WeakLearn**, providing it with mislabel distribution D_t.
2. Get back a hypothesis $h_t : X \times Y \to [0, 1]$.
3. Calculate the pseudo-loss of h_t: $\epsilon_t = \frac{1}{2} \sum_{(i,y)\in B} D_t(i, y)(1 - h_t(x_i, y_i) + h_t(x_i, y))$.
4. Set $\beta_t = \epsilon_t/(1 - \epsilon_t)$.
5. Update D_t: $D_{t+1}(i, y) = \frac{D_t(i, y)}{Z_t} \cdot \beta_t^{(1/2)(1+h_t(x_i,y_i)-h_t(x_i,y))}$
   where Z_t is a normalization constant (chosen so that D_{t+1} will be a distribution).

Output the hypothesis: $h_{fin}(x) = \arg\max_{y\in Y} \sum_{t=1}^{T} \left(\log \frac{1}{\beta_t}\right) h_t(x, y)$.

Results

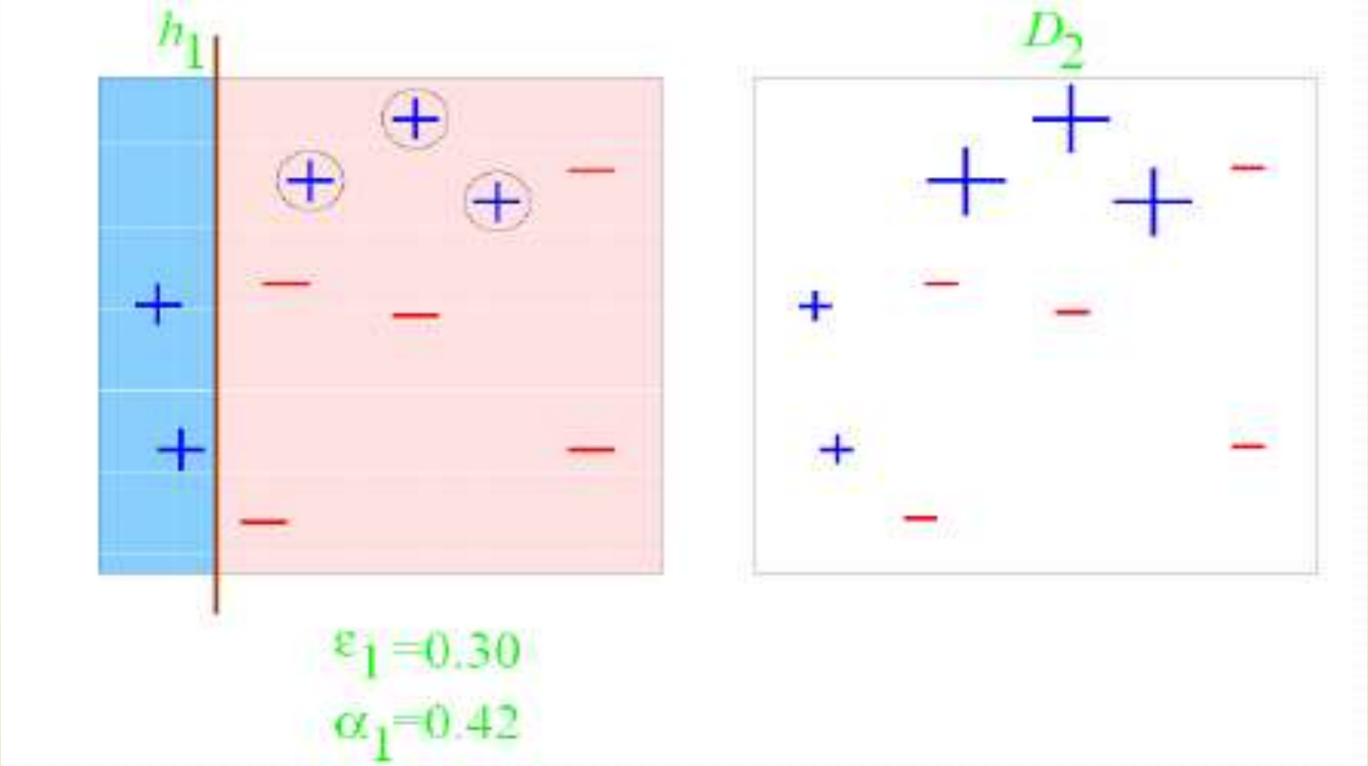
جدول ۴–۳۸. نتایج اعمال الگوریتم AdaBoost.M2 روی چند ویژگی برای بازشناسی ارقام					
ویژگی	تعداد شبکه های ایجاد شده با AdaBoost.M2	نورونهای لایه میانی	نرخ بازشناسی داده های آموزش	نرخ بازشناسی داده های آزمایش	میزان بهبود بازشناسی داده های آزمایش
هیستوگرام گرادیان	۴	۴۰	۹۹.۹۹۸	۹۸.۸۸	۰.۳۱
هیستوگرام گرادیان	۵	۴۰	۱۰۰	۹۸.۹۴	۰.۳۷
کریش	۳	۴۰	۹۹.۶۳	۹۸.۳۷	۰.۶۲
کریش	۵	۴۰	۹۹.۹۰	۹۸.۴۵	۰.۷۰
DCT81	۳	۴۰	۹۹.۷۳	۹۷.۸۰	۰.۷۷
DCT81	۵	۴۰	۹۹.۹۸	۹۸.۱۹	۱.۱۶
DCT225	۳	۶۰	۹۹.۹۳	۹۷.۹۸	۰.۷
DCT225	۵	۶۰	۱۰۰	۹۸.۱۸	۰.۹
بلوک بندی	۵	۴۰	۹۹.۸۶	۹۷.۵۳	۰.۹۵
گرادیان بهبود یافته	۳–۵	۶۰	۱۰۰	۹۹.۰۸	۰.۰۶

A toy example



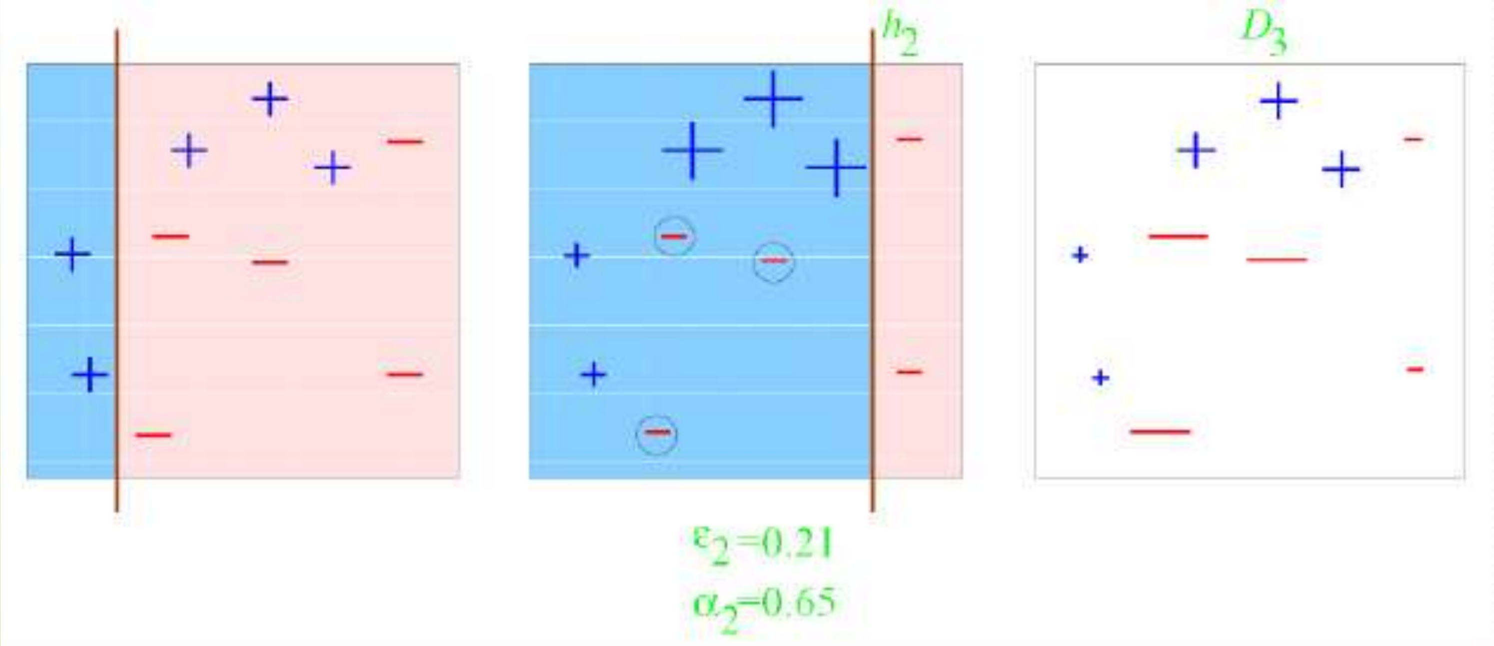
Training set: 10 points
(represented by plus or minus)
Original Status: Equal Weights for all training samples

A toy example(cont'd)



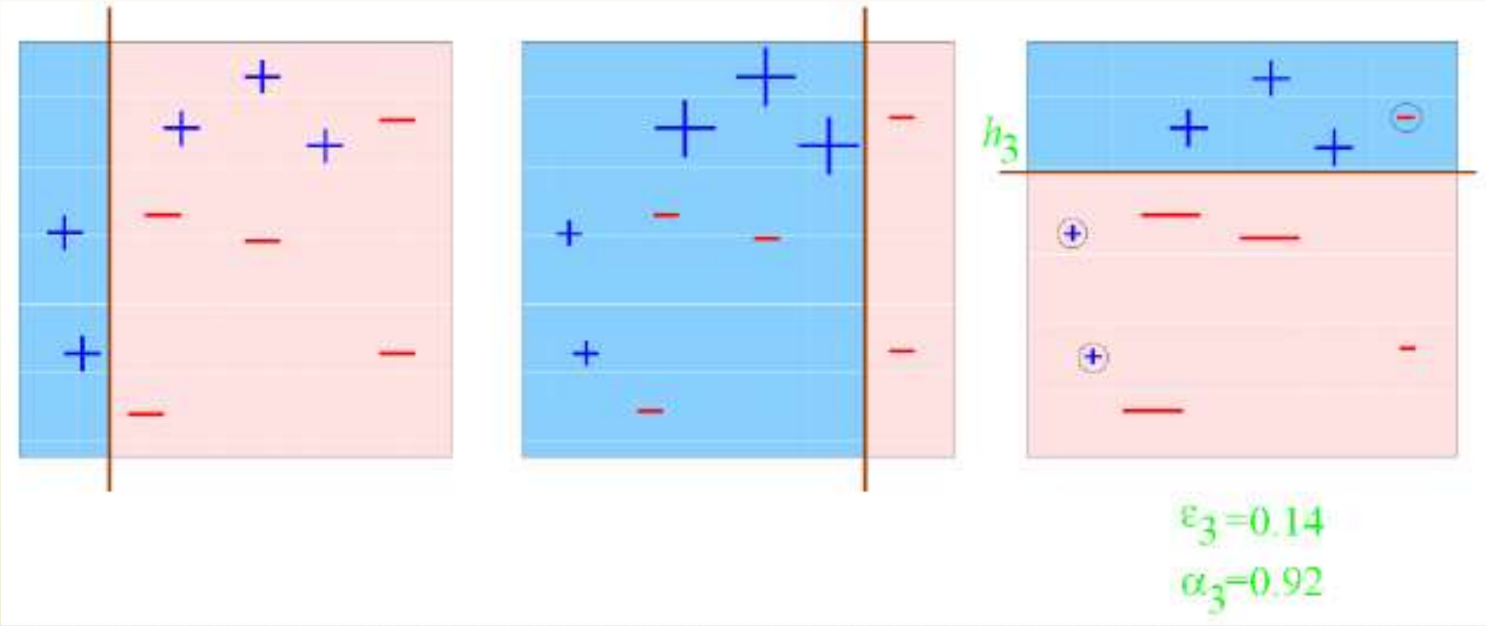
Round 1: Three "plus" points are not correctly classified;
They are given higher weights.

A toy example(cont'd)



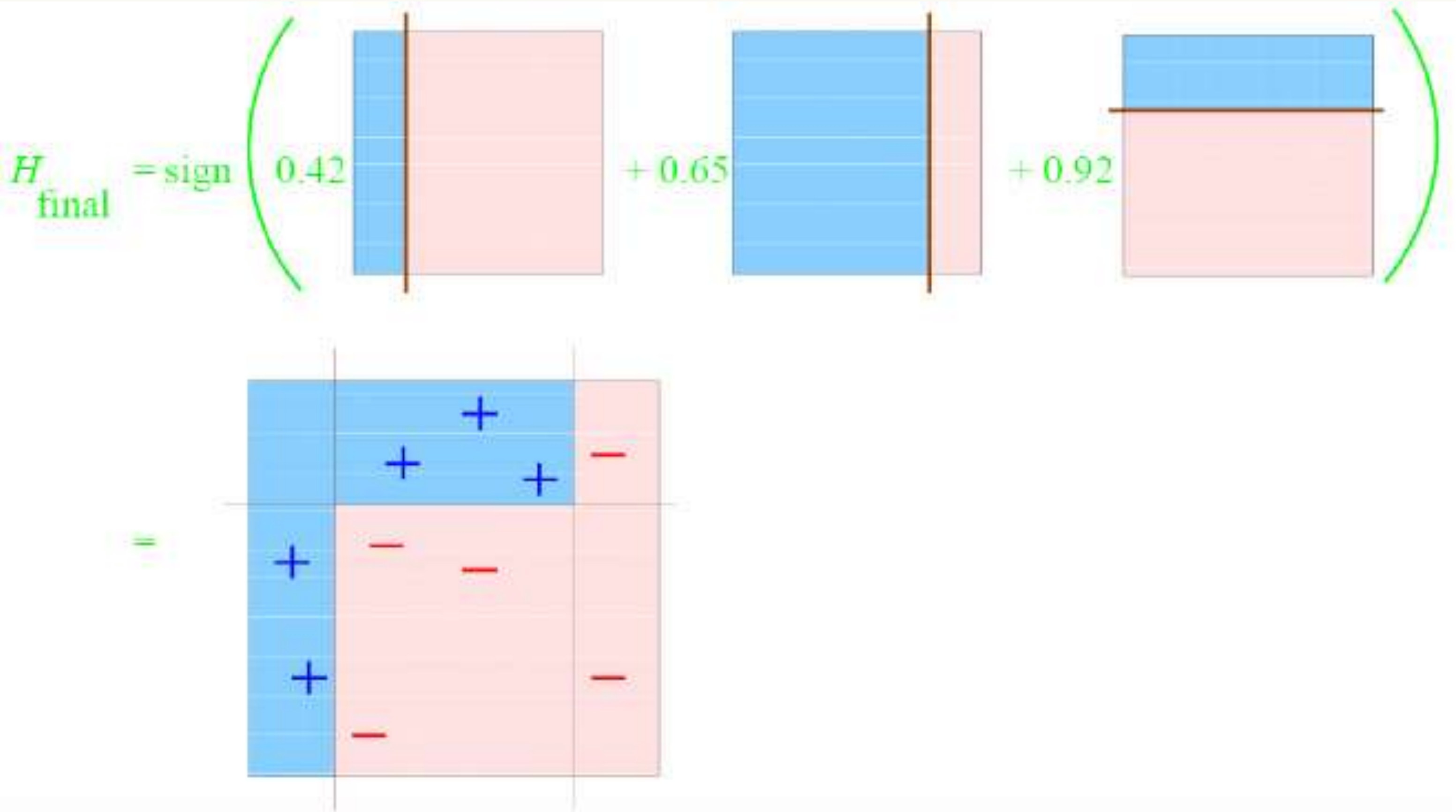
Round 2: Three "minus" points are not correctly classified;
They are given higher weights.

A toy example(cont'd)



Round 3: One "minus" and two "plus" points are not correctly classified;
They are given higher weights.

A toy example(cont'd)



Final Classifier: integrate the three "weak" classifiers and obtain a final strong classifier.

Example

Initialization...

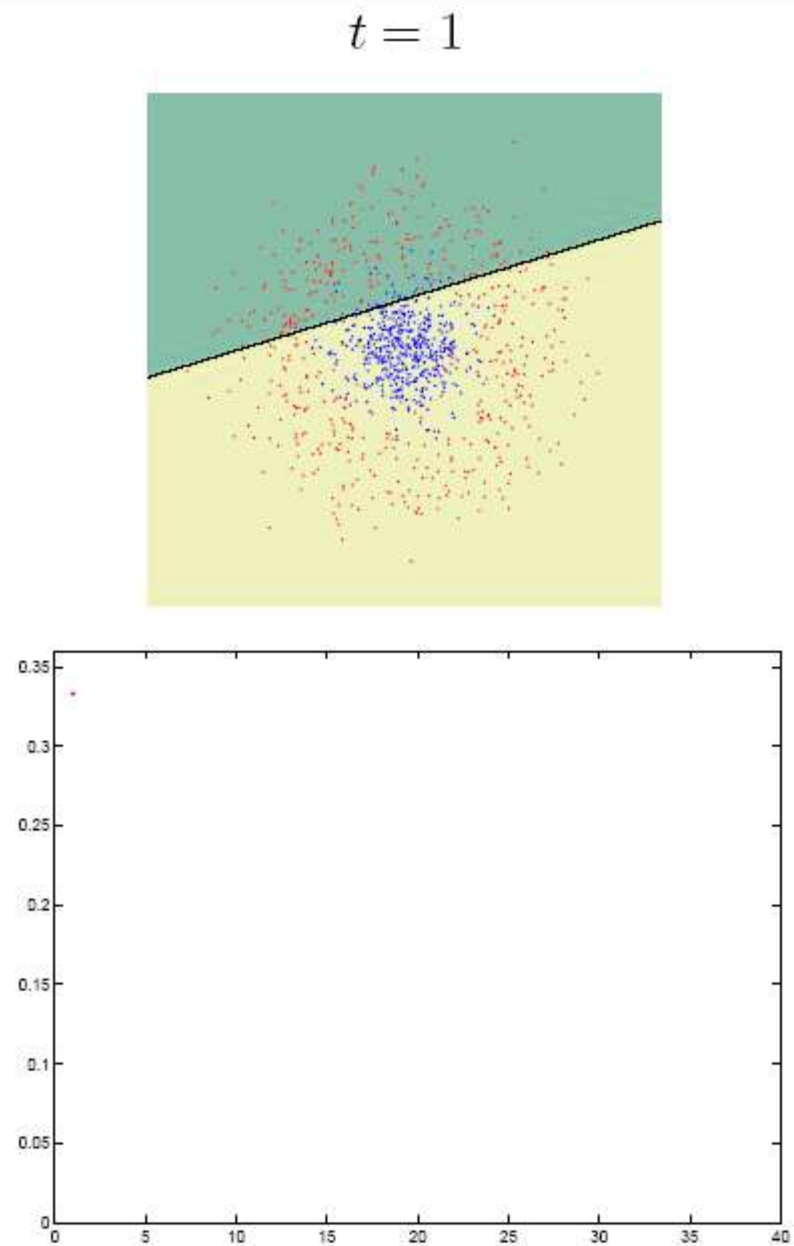
For $t = 1, \dots, T$:

- Find $h_t = \arg\min_{h_j \in \mathcal{H}} \epsilon_j = \sum_{i=1}^{m} D_t(i)[y_i \neq h_j(x_i)]$
- If $\epsilon_t \geq 1/2$ then stop
- Set $\alpha_t = \frac{1}{2}\log(\frac{1+r_t}{1-r_t})$
- Update

$$D_{t+1}(i) = \frac{D_t(i)exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

Output the final classifier:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$



Example

Initialization...

For $t = 1, ..., T$:

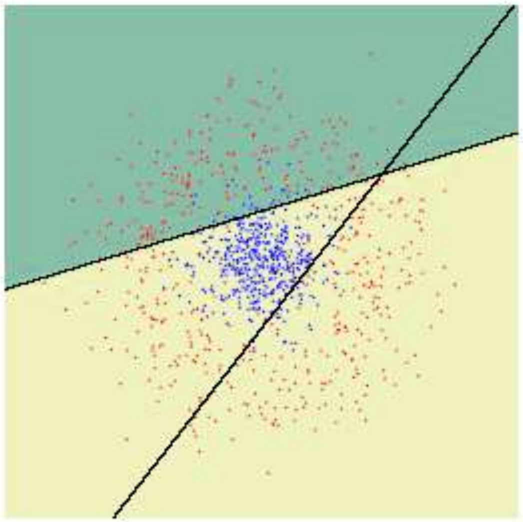
- Find $h_t = \arg\min_{h_j \in \mathcal{H}} \epsilon_j = \sum_{i=1}^{m} D_t(i)[y_i \neq h_j(x_i)]$
- If $\epsilon_t \geq 1/2$ then stop
- Set $\alpha_t = \frac{1}{2}\log(\frac{1+r_t}{1-r_t})$
- Update

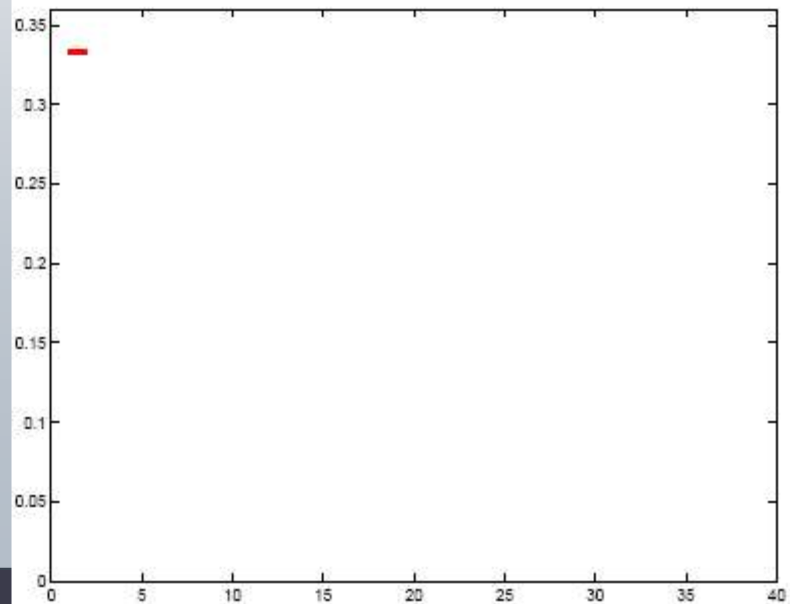
$$D_{t+1}(i) = \frac{D_t(i)exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

Output the final classifier:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

$t = 2$





Example

Initialization...

For $t = 1, \ldots, T$:

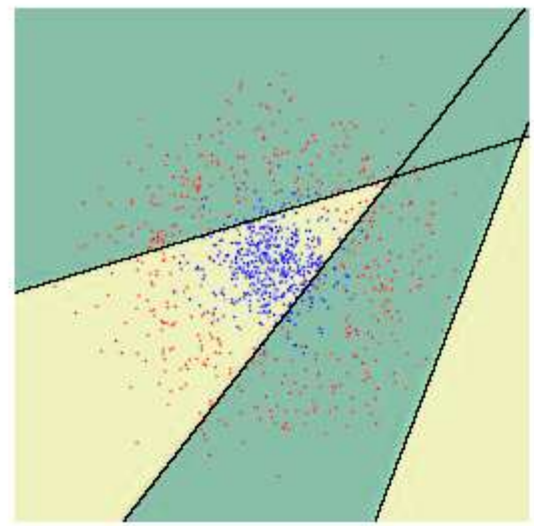
- Find $h_t = \arg\min_{h_j \in \mathcal{H}} \epsilon_j = \sum_{i=1}^{m} D_t(i)[y_i \neq h_j(x_i)]$
- If $\epsilon_t \geq 1/2$ then stop
- Set $\alpha_t = \frac{1}{2}\log(\frac{1+r_t}{1-r_t})$
- Update

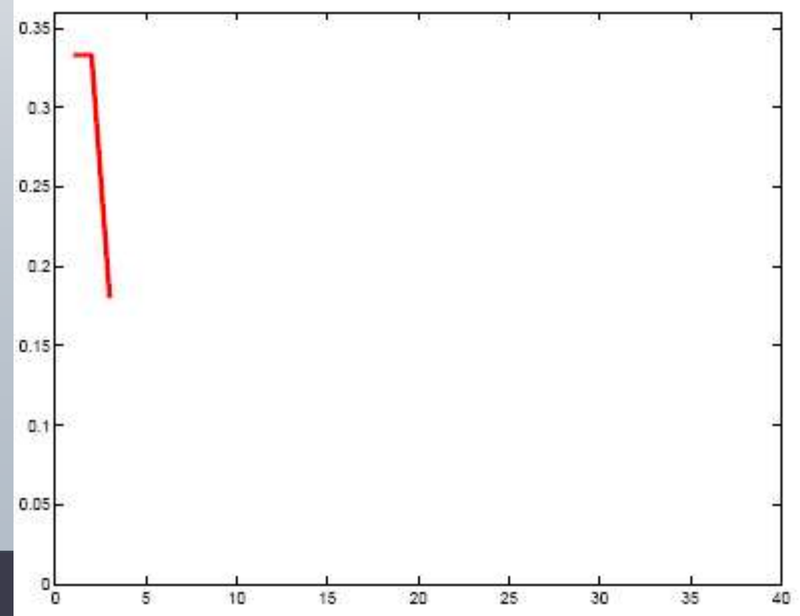
$$D_{t+1}(i) = \frac{D_t(i)exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

Output the final classifier:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

$t = 3$





Example

Initialization...

For $t = 1, ..., T$:

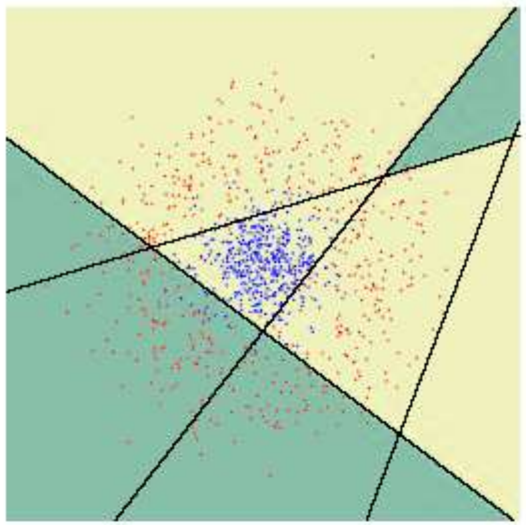
- Find $h_t = \arg\min_{h_j \in \mathcal{H}} \epsilon_j = \sum_{i=1}^{m} D_t(i)[y_i \neq h_j(x_i)]$
- If $\epsilon_t \geq 1/2$ then stop
- Set $\alpha_t = \frac{1}{2}\log(\frac{1+r_t}{1-r_t})$
- Update

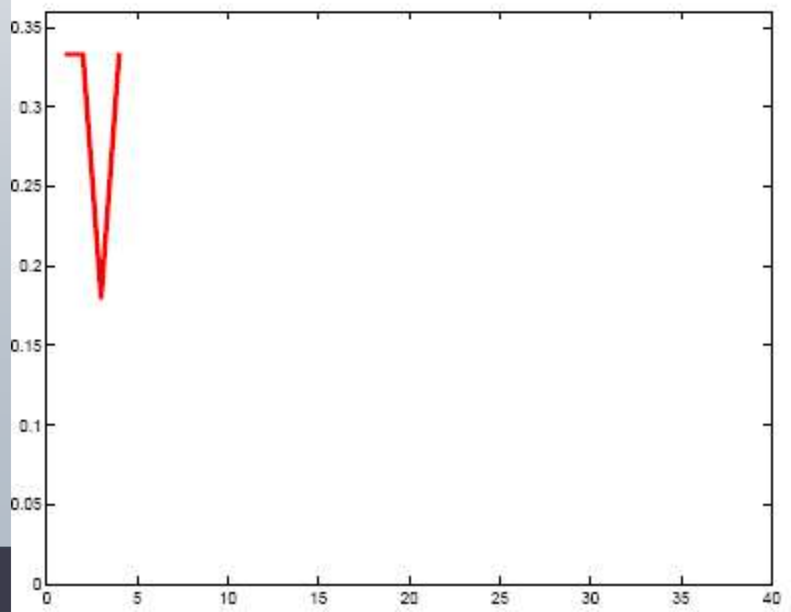
$$D_{t+1}(i) = \frac{D_t(i)exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

Output the final classifier:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

$t = 4$





Example

Initialization...

For $t = 1, ..., T$:

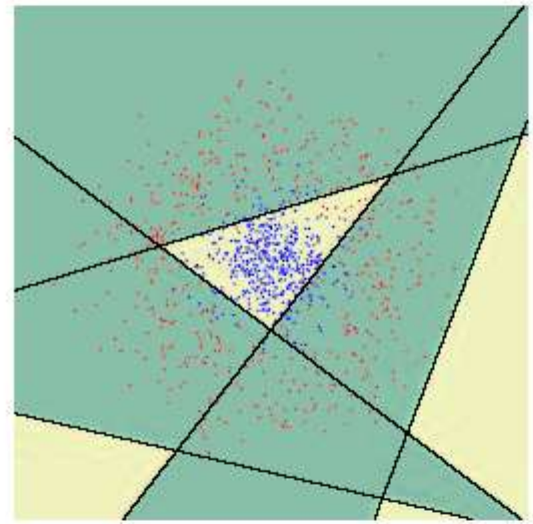
- Find $h_t = \arg\min_{h_j \in \mathcal{H}} \epsilon_j = \sum_{i=1}^{m} D_t(i)[y_i \neq h_j(x_i)]$
- If $\epsilon_t \geq 1/2$ then stop
- Set $\alpha_t = \frac{1}{2}\log(\frac{1+r_t}{1-r_t})$
- Update

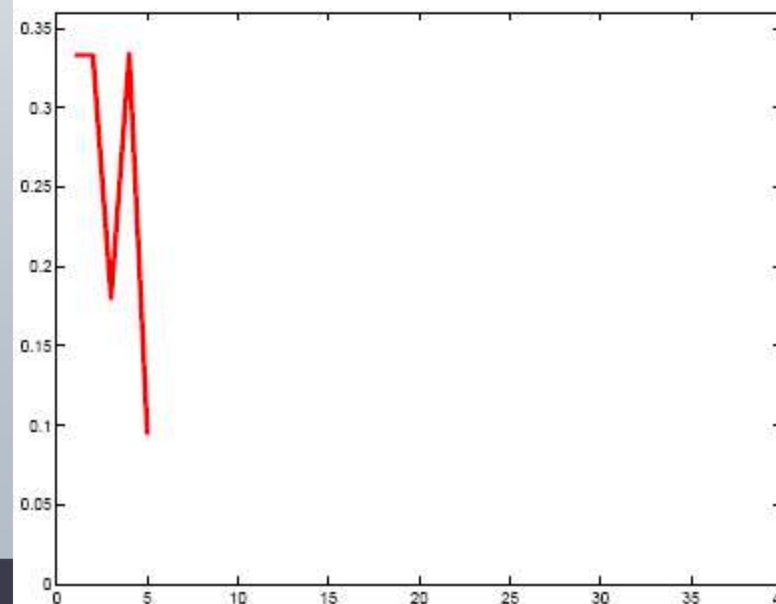
$$D_{t+1}(i) = \frac{D_t(i)exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

Output the final classifier:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

$t = 5$





Example

Initialization...

For $t = 1, \ldots, T$:

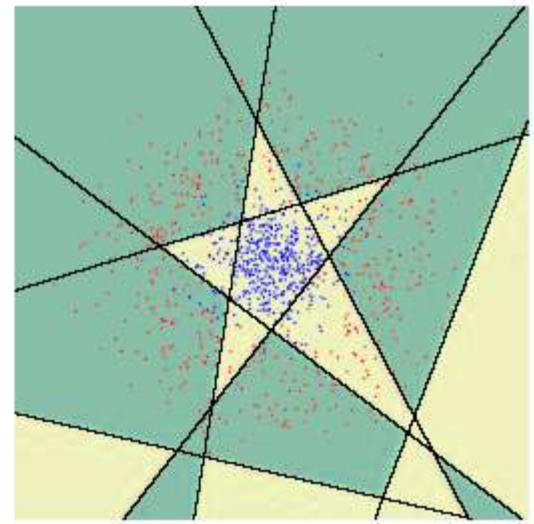
- Find $h_t = \arg\min_{h_j \in \mathcal{H}} \epsilon_j = \sum_{i=1}^{m} D_t(i)[y_i \neq h_j(x_i)]$
- If $\epsilon_t \geq 1/2$ then stop
- Set $\alpha_t = \frac{1}{2}\log(\frac{1+r_t}{1-r_t})$
- Update

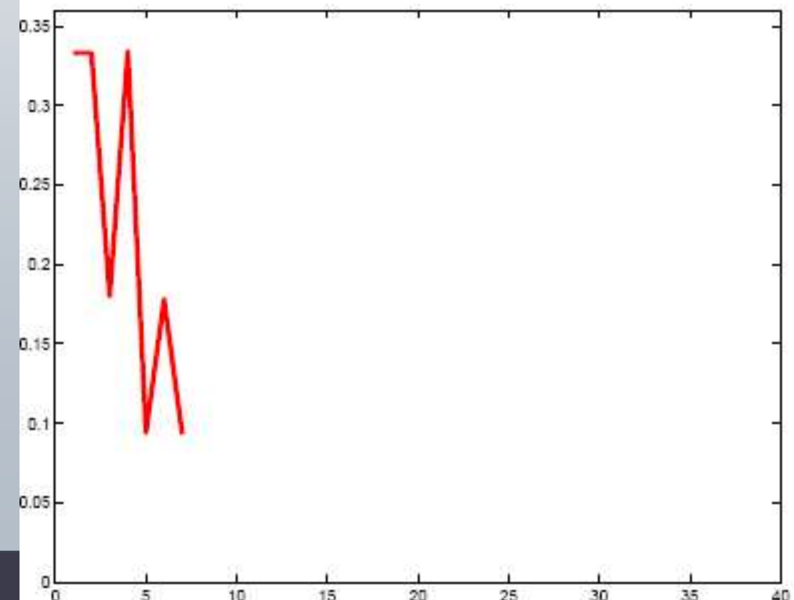
$$D_{t+1}(i) = \frac{D_t(i)exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

Output the final classifier:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

$t = 7$





Example

Initialization...

For $t = 1, ..., T$:

- Find $h_t = \arg\min_{h_j \in \mathcal{H}} \epsilon_j = \sum_{i=1}^{m} D_t(i)[y_i \neq h_j(x_i)]$
- If $\epsilon_t \geq 1/2$ then stop
- Set $\alpha_t = \frac{1}{2}\log(\frac{1+r_t}{1-r_t})$
- Update

$$D_{t+1}(i) = \frac{D_t(i)exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

Output the final classifier:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

$t = 40$