Paper Reading: The Dynamics of AdaBoost: Cyclic Behavior and Convergence of Margins

Nguyen D. Pham

Department of Computer Science
University of Houston

November 30, 2015

Introduction

- AdaBoost: Overfitting?
- Interesting behavior:
  - After a few iteration, training error reaches zero.
  - Generalization error keeps reducing.

$\Longrightarrow$ Minimizing error is not the main factor.

Discussion

AdaBoost Model

- Training set $\{(\mathbf{x}_i, y_i)\}_{i=1,...,m}$. Labels $y_i \in \{-1, 1\}$. Data space $\mathcal{X}$.
- Let $\mathcal{H} = \{h_1, ..., h_n\}$ the set of possible weak classifiers. $h_j : \mathcal{X} \mapsto \{-1, 1\}$.
- Assumption: $\mathcal{H}$ is finite and $m \ll n$
- Matrix $M(m \times n)$, $M_{i,j} = y_i h_j(\mathbf{x}_i)$. Indicating if example $\mathbf{x}_i$ is correctly classified by h_j
- Weights (normalized) vectors: $\boldsymbol{\lambda} = [\lambda_1, ..., \lambda_n]$ and $\mathbf{d} = [d_1, ..., d_m]$.

AdaBoost Iteration (1)

- Iterations: $1, ..., t_{max}$.
- The combined classifier $f_{\boldsymbol{\lambda}}$:

$$f_{\boldsymbol{\lambda}} = \frac{\sum_{j=1}^{n} \lambda_j h_j}{\|\boldsymbol{\lambda}\|_1}$$

margin of example i: $y_i f_{\boldsymbol{\lambda}}(\mathbf{x}_i) = (\mathbf{M}\boldsymbol{\lambda})_i$

- Choose either h_j or $-h_j$, the $norm - 1$ is reduced to sum of weights.
- At iteration t: A classifier h_{j_t} is selected.
  - *probability of error* $d_- = \sum_{i: M_{i,j_t}=-1} d_{t,i}$ and $d_+ = 1 - d_-$
  - *edge* of classifier h_{j_t}: $(\mathbf{d}_t^T \mathbf{M})_{j_t} = d_+ - d_-$
  - choose classifier with lowest error:

$$j_t \in \arg\max_j (\mathbf{d}_t^T \mathbf{M})_j$$

corresponding edge: r_t and $d_+ = \frac{1+r_t}{2}$, $d_- = \frac{1-r_t}{2}$

AdaBoost Iteration (2)

- Each iteration: update $\mathbf{d}$ and $\boldsymbol{\lambda}$
- Goal: a combined classifier that maximizes the minimal margin
- Min-max theorem

$$\max_{\boldsymbol{\lambda}} \min_{i} (\mathbf{M}\boldsymbol{\lambda})_i = \min_{\mathbf{d}} \max_{j} (\mathbf{d}^T\mathbf{M})_j$$

- Reduced to iteration over $\mathbf{d}$ only

AdaBoost Original Algorithm

- Matrix $\mathbf{M}$, $\boldsymbol{\lambda}_1 = 0$
- Loop for $t = 1, ..., t_{max}$
  - $d_{t,i} = e^{-(\mathbf{M}\boldsymbol{\lambda}_t)_i} / \sum_{k=1}^{m} e^{-(\mathbf{M}\boldsymbol{\lambda}_t)_k}$
  - $j_t \in \arg\max_j (\mathbf{d}_t^T \mathbf{M})_j$
  - $r_t = (\mathbf{d}_t^T \mathbf{M})_{j_t}$
  - $\alpha_t = {}^1\!/_2 ln \left(\frac{1+r_t}{1-r_t} \right)$
  - $\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \alpha_t \mathbf{e}_{j_t}$ where $\mathbf{e}_{j_t}$ is 1 at j_t and 0 elsewhere.
- Output $\boldsymbol{\lambda}_{t_{max}} / \|\boldsymbol{\lambda}_{t_{max}}\|_1$

AdaBoost Reduced to Iterated Map

- Matrix $\mathbf{M}$, $_1 = random_values$
- Loop for $t = 1, ..., t_{max}$
  - $j_t \in \arg\max_j (\mathbf{d}_t^T \mathbf{M})_j$
  - $r_t = (\mathbf{d}_t^T \mathbf{M})_{j_t}$
  - $d_{t+1,i} = \dfrac{d_{t,i}}{1 + M_{i,j_t} r_t}$
- Output $\mathbf{d_{t_{max}}}$

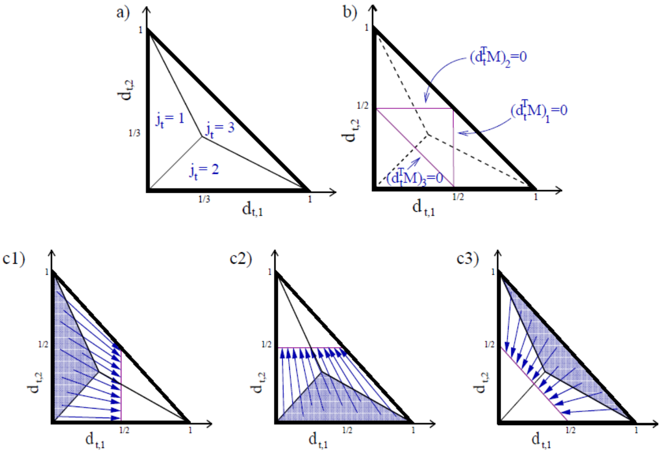
Dynamics of AdaBoost in 3×3 case

$$\mathbf{M} = \begin{pmatrix} -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}$$

- 3 training examples, 3 classifiers making one mistake each.
- Result: $\boldsymbol{\lambda} = [1/3, 1/3, 1/3]$.
- Our concern: the dynamics of $\mathbf{d}_t$.
- Topology method: $\mathbf{d}_t \in \triangle_3$, a 3-simplex, is projected onto 2-dimensional plane.

Projection

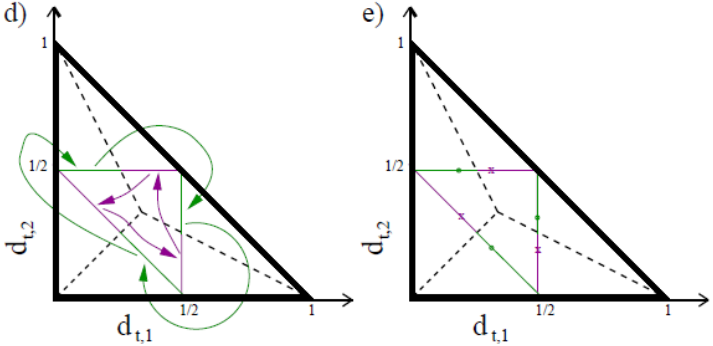
» The system at iteration-t represented by a point $(d_{t,1}, d_{t,2})$



Domains of $\mathbf{d}$

» $\mathbf{d}_t$ move along the edges of the inner triangle:

$$(\mathbf{d}_{t+1}^T \mathbf{M})_{j_t} = 0$$



Cyclic Behavior

- Cycle defined by the chosen classifier.
- Consider some cycle of $\mathbf{d}_t$, with length T: $d_1^{cyc}, d_2^{cyc}, ..., d_T^{cyc}$.
- From slide 8, the condition for a cycle is:

$$\prod_{t=1}^{T}(1 + M_{i,j_t} r_t^{cyc}) = 1$$

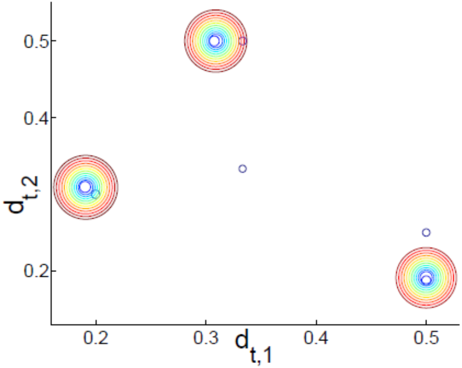
- Consider 3-cycles, noting the cyclic permutation, we might have 2 cycles (by index of weak classifiers): $(1, 2, 3)$ and $(1, 3, 2)$
- Some analysis will yield the solutions (only one is shown):

$$\mathbf{d}_1^{cyc} = \left({}^{3-\sqrt{5}}/{}_{4}, {}^{\sqrt{5}-1}/{}_{4}, {}^{1}/{}_{2}\right)^T$$
$$\mathbf{d}_2^{cyc} = \left({}^{\sqrt{5}-1}/{}_{4}, {}^{1}/{}_{2}, {}^{3-\sqrt{5}}/{}_{4}\right)^T$$
$$\mathbf{d}_3^{cyc} = \left({}^{1}/{}_{2}, {}^{3-\sqrt{5}}/{}_{4}, {}^{\sqrt{5}-1}/{}_{4}\right)^T$$

Theorem 1

For the 3×3 matrix $\mathbf{M}$

- Weight vectors converge to one of two cycles.
- The cycles correspond to combined classifier with maximal margin.



Theorem 2

$$\mathbf{M} = \begin{pmatrix} -1 & 1 & \cdots & 1 \\ 1 & -1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & -1 \end{pmatrix}$$

- AdaBoost converges to at least $(m-1)!$ stable cycles of lenght m.
- The convergence yields maximum margin.

Theorem 3 Manifolds of cycles

- Consider the matrix $\mathbf{M}$ with some sets of identical rows: examples that are identically classified by all weak classifiers.
- If there is a cycle composing identical examples, $\bar{I}$, with some weight components, then there is another cycle with a pertubation in weights of $\bar{I}$ as long as the sum does not change.

Other results

- General cases: None of the above results hold.
- Counter example: AdaBoost does not converge to stable cycle.
- Counter example: AdaBoost converges, but the margin is not maximized.
- Which lead to open questions...

Personal View

- A tough but interesting read.
- In the end, the results are not generalized: Proof for simple cases (which is complex enough), counter examples for general cases.
- The topology technique is nice, but might not be useful to generalize.
- Very nice idea: view AdaBoost as an optimization problem, insight to understanding the algorithm. The reduction to samples weights iteration is very useful.

Thank you!
Questions...