Boosting

Μηχανική Μάθηση

Μάθηση Ensemble

- **Επιβλεπόμενη μάθηση**
  - Αναζήτηση σε *χώρο υποθέσεων* κατάλληλου μοντέλου, το οποίο θα παράγει *ορθές* προβλέψεις για δοσμένο πρόβλημα
  - Σε πολλά προβλήματα, δεν είναι *εύκολο* να βρούμε τέτοια *μοντέλα*
  - Ωστόσο είναι *εύκολο* να βρούμε *απλούστερα μοντέλα* με σχετικά *καλή ακρίβεια*
- Μπορούμε να συνδυάσουμε απλούστερα μοντέλα για να κατασκευάσουμε ένα πιο ισχυρό;
  - **Ναι ⇒ Μάθηση Ensemble** (*Ensemble learning*)
    - *Συνδυασμός*, υπό προϋποθέσεις, *απλούστερων μοντέλων*
    - Πλεονέκτημα: υψηλότερη ακρίβεια
    - Μειονέκτημα: μεγαλύτερος χρόνος εκπαίδευσης

Τεχνικές μάθησης Ensemble

1. **Boosting**
   - *Σταδιακή* κατασκευή ensemble
   - Σε κάθε μοντέλο που προστίθεται, η *έμφαση* δίνεται στα *δείγματα* εκείνα που έχουν ταξινομηθεί λάθος από τα προηγούμενα
   - Παράδειγμα: AdaBoost
2. **Bootstrap Averaging (Bagging)**
   - Κάθε μοντέλο έχει *ίδιο βάρος* στη λήψη της απόφασης ταξινόμησης
   - Για την αύξηση της διακύμανσης, τα μοντέλα εκπαιδεύονται σε *διαφορετικά* υποσύνολα δεδομένων
   - Παράδειγμα: Random Forests
3. **Μπεϋζιανά Μοντέλα**
   1. Bayesian Model Averaging: Μέσος όρος *πολλών μοντέλων* με τα βάρη να προκύπτουν από την *εκ των υστέρων πιθανότητα* κάθε μοντέλου
   2. Bayesian Model Combination: Συνδυασμός *πολλών ensembles*
4. **Stacking**
   - Εκπαίδευση *μοντέλου* που έχει ως *είσοδο* τις προβλέπεις *άλλων μοντέλων*

Boosting

- **Κεντρική ιδέα**
  - Χρήση **ασθενών μοντέλων μάθησης** (*weak learners*) για την κατασκευή ενός πιο ισχυρού
- **Διαδικασία**
  - *Επαναληπτική* διαδικασία μάθησης *ασθενών μοντέλων* και *προσθήκη* τους σε έναν τελικό *ισχυρό ταξινομητή*
    - Το **βάρος** κάθε ασθενούς μοντέλου εξαρτάται από την **ακρίβειά** του
  - Μετά την προσθήκη ενός μοντέλου, τα **βάρη** των **δειγμάτων εκπαίδευσης μεταβάλλονται**
    - Στα ορθά ταξινομημένα μειώνονται
    - Στα εσφαλμένα ταξινομημένα αυξάνονται
  - Έτσι, τα *επόμενα μοντέλα* δίνουν **μεγαλύτερη έμφαση** στα δείγματα που τα προηγούμενα μοντέλα *ταξινόμησαν λάθος*
- **AdaBoost** (*Adaptive Boosting*)
  - Προτάθηκε από τους *Freund* και *Schapire* το *1997*
  - Από τις *πρώτες* (και πιο *γνωστές*) προσεγγίσεις

Ασθενής Μάθηση

- **Ορισμός**
  - Μια *κλάση εννοιών* C «μαθαίνεται» **ασθενώς** (*weakly PAC-learnable*) αν υπάρχει *ασθενής αλγόριθμος* L και $\gamma > 0$ τέτοια ώστε $\forall \delta > 0, c \in C, D$: $\mathbb{P}_{S\sim D}\left[R(h_S) \leq \frac{1}{2} - \gamma\right] \geq 1 - \delta$
    - D: όλες οι πιθανές κατανομές
    - R *σφάλμα γενίκευσης*
    - S: *δείγμα* μεγέθους $m = poly(\frac{1}{\delta})$
      - για κάποιο συγκεκριμένο πολυώνυμο
- **Ισχυρή μάθηση**
  - $\mathbb{P}_{S\sim D}[R(h_S) \leq \epsilon] \geq 1 - \delta$
  - $m = poly\left(\frac{1}{\epsilon}, \frac{1}{\delta}\right)$
  - ϵ σφάλμα γενίκευσης
- **Διαφορά ασθενούς** με **ισχυρή** μάθηση
  - Για ίδια **εμπιστοσύνη** (*confidence*) 1 – δ, μικρότερη **ορθότητα** (*accuracy*)
    - $\frac{1}{2} + \gamma$ αντί $1 - \epsilon$

AdaBoost

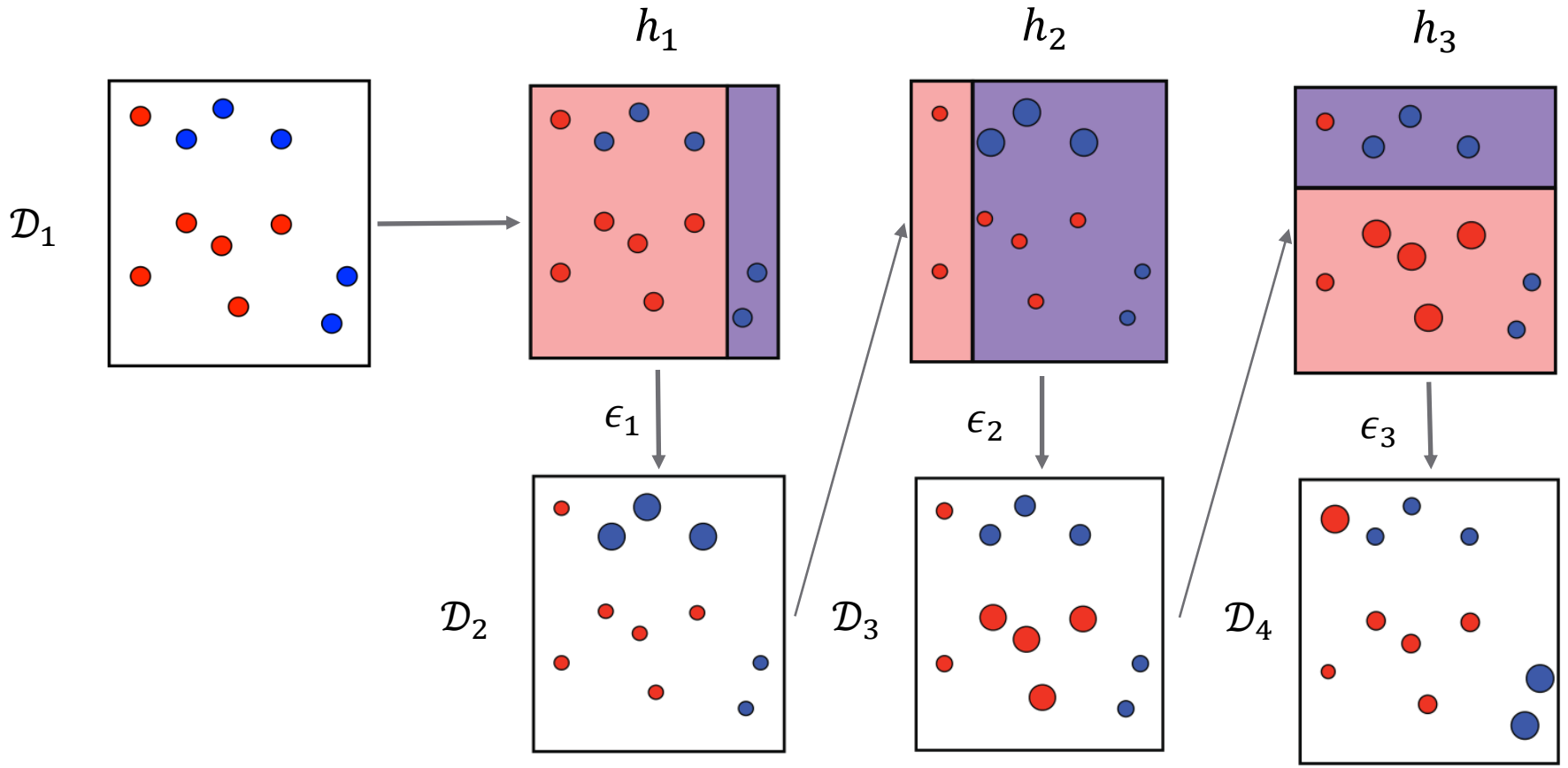
- *Χώρος υποθέσεων* $\mathcal{H} \subseteq \{-1, +1\}^{\mathcal{X}}$
- *Είσοδος* $S = \big((x_1, y_1), \dots, (x_m, y_m)\big)$
- **AdaBoost**
  1. for $i \leftarrow 1$ to m
     i. $\mathcal{D}_1(i) \leftarrow \frac{1}{m}$
  2. for $t \leftarrow 1$ to T
     i. Επιλογή **ασθενούς ταξινομητή** $h_t \in \mathcal{H}$ τέτοιου ώστε $h_t \in \operatorname{argmin}_{h \in \mathcal{H}} \epsilon_t$
        - **Σφάλμα ταξινόμησης** $\epsilon_t = \mathbb{P}_{i \sim \mathcal{D}_t}[h(x_i) \neq y_i] = \sum_{i=1}^{m} \mathcal{D}_t(i) \mathbf{1}_{h(x_i) \neq y_i}$
     ii. Συντελεστής (βάρος) $\alpha_t \leftarrow \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t}$
     iii. Όρος **κανονικοποίησης** $Z_t \leftarrow 2\sqrt{\epsilon_t(1-\epsilon_t)}$
     iv. for $i \leftarrow 1$ to m
        i. $\mathcal{D}_{t+1}(i) \leftarrow \mathcal{D}_t(i) \frac{1}{Z_t} e^{-\alpha_t y_i h(x_i)}$
  3. return $f = f_T = \sum_{t=1}^{T} \alpha_t h_t$

AdaBoost – Παρατηρήσεις

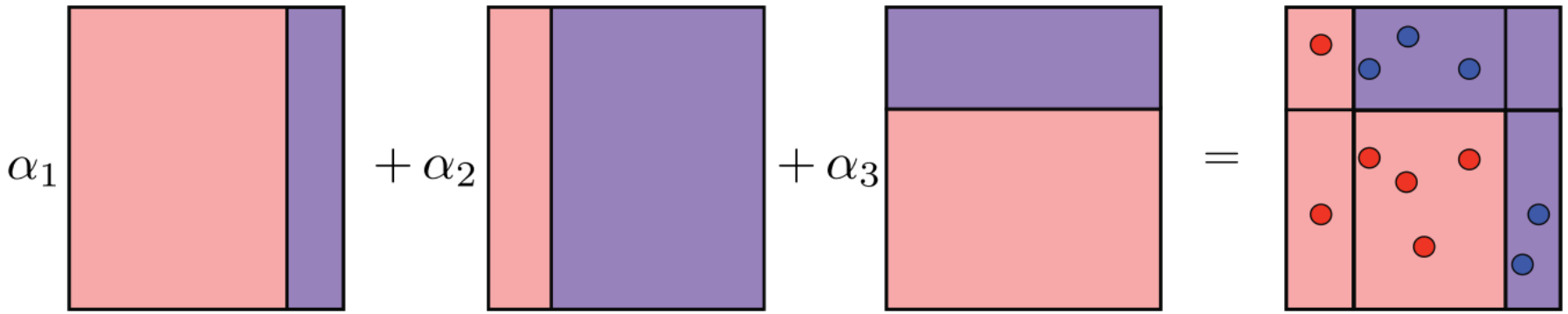
- *Δημιουργία* **κατανομών δειγμάτων** $\mathcal{D}_t$ από τα *δεδομένα εκπαίδευσης*
  - *Αρχικά* **ομοιόμορφη κατανομή**
  - Έστω ότι βρισκόμαστε στο γύρο t. Ο ασθενής ταξινομητής h_t έχει σφάλμα $\epsilon_t \leq \frac{1}{2} - \gamma, \gamma > 0$ (συνθήκη ασθενούς μάθησης). Τότε:
    1. το **βάρος** των εσφαλμένα ταξινομημένων δειγμάτων αυξάνει
       - $y_i h(x_i) = -1$ οπότε $\mathcal{D}_{t+1}(i) \leftarrow \mathcal{D}_t(i) \frac{1}{Z_t} e^{\alpha_t} = \mathcal{D}_t(i) \frac{1}{2\epsilon_t}$
    2. το **βάρος** των ορθά ταξινομημένων δειγμάτων μειώνεται
       - $y_i h(x_i) = 1$ οπότε $\mathcal{D}_{t+1}(i) \leftarrow \mathcal{D}_t(i) \frac{1}{Z_t} e^{-\alpha_t} = \mathcal{D}_t(i) \frac{1}{2(1-\epsilon_t)}$

- Βάρος *βασικού ταξινομητή* h_t
  - $\alpha_t \leftarrow \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t}$
    - Εξαρτάται από την **ορθότητα** $1 - \epsilon_t$ του ασθενούς ταξινομητή h_t στην επανάληψη t
  - Αν $\epsilon_t < \epsilon_t{}'$ τότε , $\alpha_t > \alpha_t{}'$, οπότε οι *πιο ορθοί* ταξινομητές έχουν *μεγαλύτερη* συνεισφορά στο τελικό άθροισμα

Παράδειγμα


h_1
h_2
h_3
$\mathcal{D}_1$
ϵ_1
ϵ_2
ϵ_3
$\mathcal{D}_2$
$\mathcal{D}_3$
$\mathcal{D}_4$

Παράδειγμα (συνέχεια)



Όριο εμπειρικού σφάλματος AdaBoost

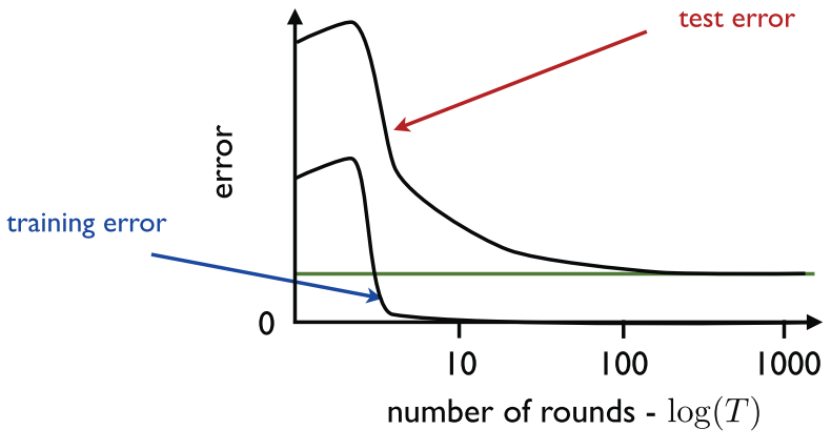
- Για το **εμπειρικό σφάλμα** $\hat{R}(h)$ του AdaBoost ισχύει:
  - $\hat{R}(h) \leq exp\left[-2\sum_{t=1}^{T}\left(\frac{1}{2}-\epsilon_t\right)^2\right]$
- Αν επιπρόσθετα ισχύει $\forall t \in [1,T], \gamma \leq \frac{1}{2}-\epsilon_t$, τότε
  - $\hat{R}(h) \leq e^{-2\gamma^2 T}$
- γ: «**πλεονέκτημα**» (*edge*)
  - *Σχετίζεται* με την *ακρίβεια* των *ασθενών* (βασικών) μοντέλων ταξινόμησης (5η διαφάνεια)
  - Ο αλγόριθμος *δεν απαιτεί* να είναι *γνωστό εκ των προτέρων*
    - «*Προσαρμόζεται*» στην ακρίβεια των βασικών ταξινομητών ⇒ **adaptive boosting**

Χρήση AdaBoost

- **Βασικοί ταξινομητές**
  - Συνήθως **δέντρα απόφασης** βάθους 1 (*stumps*)
- **Stumps**
  - **Συναρτήσεις κατωφλίου** (*threshold functions*) σχετιζόμενες *με ένα χαρακτηριστικό*
    - N *χαρακτηριστικά* $\Rightarrow$ N *stumps* σε <u>κάθε γύρο</u> boosting
  - Ταξινόμηση τιμών κάθε χαρακτηριστικού: $O(m \log m)$
    - N χαρακτηριστικά: $O(mN \log m)$
  - <u>Για κάθε</u> χαρακτηριστικό:
    - $m + 1$ *πιθανά* βέλτιστα κατώφλια $\Rightarrow$ $m + 1$ συγκρίσεις
    - οπότε $O(m)$ σε έναν γύρο και $O(mNT)$ συνολικά
  - <u>Συνολική πολυπλοκότητα</u>: $O(mN \log m + mNT)$

Υπερπροσαρμογή

- Οικογένεια συναρτήσεων $\mathcal{F}_T$ από τις οποίες ο AdaBoost διαλέγει την έξοδο μετά από T γύρους
  - $\mathcal{F}_T = \{\text{sgn}(\sum_{t=1}^{T} \alpha_t h_t) : \alpha_t \geq 0, h_t \in \mathcal{H}, t \in [1,T]\}$
- Αν $VCdim(\mathcal{H}) = d$ τότε αποδεικνύεται πως
  - $VCdim(\mathcal{F}_T) \leq 2(d+1)(T+1)\log_2\big((T+1)e\big)$
- Το *άνω όριο* μεγαλώνει ως προς $O(dT\log T)$ συνεπώς μπορεί να εμφανιστεί **υπερπροσαρμογή** για *μεγάλες τιμές* του T
- Ωστόσο, η **εμπειρική ανάλυση** δείχνει πως το σφάλμα γενίκευσης πέφτει όσο το T *μεγαλώνει*



Συνθήκη ασθενούς μάθησης για τον AdaBoost

- **Ορισμός**
  - Το «**πλεονέκτημα**» (*edge*) ενός **βασικού ταξινομητή** (*base classifier*) h_t για κατανομή $\mathcal{D}$ των δεδομένων εκπαίδευσης ορίζεται ως εξής
    - $\gamma_t(\mathcal{D}) = \frac{1}{2} - \epsilon_t = \frac{1}{2}\sum_{i=1}^{m} y_i h_t(x_i)\mathcal{D}(i)$
- **AdaBoost Weak Learning Condition**
  - Υπάρχει $\gamma > 0$ τέτοιο ώστε για κάθε κατανομή $\mathcal{D}$ των δεδομένων εκπαίδευσης να ισχύει $\gamma_t(\mathcal{D}) > \gamma$

Παίγνια μηδενικού αθροίσματος 2 παικτών

- **Zero-Sum Games**
  - **Μήτρα** (*πίνακας*) απολαβής $M = (M_{ij}) \in \mathbb{R}^{m \times n}$
  - m πιθανές *ενέργειες* για τον *παίκτη-στήλη*
  - n πιθανές *ενέργειες* για τον *παίκτη-γραμμή*
  - M_{ij} **απολαβή** για τον παίκτη-στήλη όταν ο παίκτης-γραμμή *επιλέξει* την *στρατηγική* i και ο παίκτης-στήλη «*απαντήσει*» με την *στρατηγική* j

- Παράδειγμα: «*Πέτρα-Ψαλίδι-Χαρτί*»

	Πέτρα	**Ψαλίδι**	**Χαρτί**
Πέτρα	0	-1	1
Ψαλίδι	1	0	-1
Χαρτί	-1	1	0

Μεικτές Στρατηγικές

- **Mixed Strategies**
  - *Κατανομές* πάνω στο <u>πλήθος</u> των *δυνατών ενεργειών*
    - *Κατανομή* p για τις m πιθανές ενέργειες του *παίκτη-γραμμή* και q για τις n πιθανές ενέργειες του *παίκτη-στήλη*
  - **Αναμενόμενη τιμή** απολαβής
    - $\mathbb{E}_{i\sim p, j\sim q}[M_{ij}] = \sum_{i=1}^{m}\sum_{j=1}^{n} p_i M_{ij} q_j = \boldsymbol{p}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{q}$
- **Θεώρημα minimax** (John von Neumann, 1928)
  - $\min_{\boldsymbol{p}}\max_{\boldsymbol{q}} \boldsymbol{p}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{q} = \max_{\boldsymbol{q}}\min_{\boldsymbol{p}} \boldsymbol{p}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{q}$
    - <u>Για κάθε</u> παίγνιο *μηδενικού αθροίσματος* μεταξύ δύο παικτών, *υπάρχει* **μεικτή στρατηγική** <u>για κάθε παίκτη</u>, έτσι ώστε η *αναμενόμενη απώλεια* του *ενός* να είναι **ίση** με το *αναμενόμενο κέρδος* του *άλλου* ⇒ **αξία** του **παιγνίου**

AdaBoost και Παίγνια Μηδενικού Αθροίσματος

- **Παιχνίδι**
  - *Παίκτης-γραμμή* επιλέγει **δείγμα** δεδομένων $x_i, i \in [1, m]$
  - *Παίκτης-στήλη* επιλέγει **βασικό ταξινομητή** $h_t, t \in [1, T]$
  - Μήτρα απολαβής $M \in \{-1, +1\}^{m \times T}$: $M_{it} = y_i h_t(x_i)$
- **Εφαρμογή** *θεωρήματος von Neuman*
  - $\min_{\mathcal{D}} \max_{t} \sum_{i=1}^{m} \mathcal{D}(i) y_i h_t(x_i) = \max_{\boldsymbol{a}} \min_{i} \sum_{\boldsymbol{t=1}}^{\boldsymbol{T}} \frac{\alpha_t}{\|\boldsymbol{a}\|_1} y_i h_t(x_i)$
    - $\boldsymbol{\alpha}$: διάνυσμα που περιέχει τους *συντελεστές* όλων των **βασικών ταξινομητών**
      - $f = \sum_{t=1}^{T} a_t h_t$
  - Αν $\rho_{\boldsymbol{\alpha}}(x) = \frac{|\boldsymbol{\alpha h}(x)|}{\|\boldsymbol{a}\|_1}$ το **περιθώριο** (*margin*) που επιτυγχάνει για το σημείο x ο ταξινομητής f, τότε μπορούμε να γράψουμε ισοδύναμα
    - $2\gamma^* = \min_{\mathcal{D}} \max_{t} \gamma_t(\mathcal{D}) = \max_{\boldsymbol{a}} \min_{i} \rho_{\boldsymbol{\alpha}}(x_i) = \rho^*$
      - ρ^* **μέγιστο περιθώριο** ενός ταξινομητή
      - γ^* **βέλτιστο** δυνατό «**πλεονέκτημα**»

Συνέπειες

- Η **συνθήκη ασθενούς μάθησης** του ταξινομητή AdaBoost συνεπάγεται ότι επιτυγχάνει ένα **μη-μηδενικό περιθώριο**
  - Αφού ισχύει $\gamma^* > 0$, συνεπάγεται ότι $\rho^* > 0$
- Ωστόσο, παρότι ο AdaBoost αναζητά ένα μη-μηδενικό όριο μεταξύ των κλάσεων, αποδεικνύεται ότι αυτό δεν είναι το **βέλτιστο**
- Επίσης, η **συνθήκη ασθενούς μάθησης** είναι τελικά μια **ισχυρή συνθήκη**
  - *Υποθέτει* ότι τα *δείγματα* των κλάσεων είναι **γραμμικώς διαχωρίσιμα** με περιθώριο $2\gamma^* > 0$
  - Αυτό βέβαια *δεν ισχύει πάντοτε* στην πράξη…

Πλεονεκτήματα AdaBoost

- *Απλή*, *ξεκάθαρη* υλοποίηση
- *Παρότι* ψευδοπολυωνυμικός ($O(mNT)$) είναι αρκετά γρήγορος όταν το **πλήθος των χαρακτηριστικών** N και το πλήθος των **βασικών ταξινομητών** T δεν είναι μεγάλο
- Στην *πράξη*, δεν εμφανίζει *υπερπροσαρμογή*
- Υπάρχουν *θεωρητικές εγγυήσεις* για την **απόδοσή** του
  - Αναζήτηση θετικού περιθωρίου μεταξύ των κλάσεων

Μειονεκτήματα AdaBoost

- Προσδιορισμός **πλήθους** Τ βασικών **ταξινομητών**
  - Συνήθως επιλέγεται μέσω *διασταυρούμενης επικύρωσης*
- Εξάρτηση από **πολυπλοκότητα** *βασικών ταξινομητών*
  - Πιο «**απλοί**» βασικοί **ταξινομητές** οδηγούν σε μη-επαρκείς *υποθέσεις* και σε χαμηλότερα *περιθώρια*
- Δεν μπορεί να μοντελοποιήσει την ύπαρξη θορύβου και **έκτοπων τιμών** (*outliers*)
  - Θα τους δίνει **δυσανάλογα** μεγάλο βάρος
  - Ωστόσο μπορεί να χρησιμοποιηθεί για την *εύρεση* **θορύβου/έκτοπων τιμών**!
    - Δείγματα τα οποία είναι «*δύσκολο*» να ταξινομηθούν *ορθά*

Άλλες προσεγγίσεις

- **L1-regularized AdaBoost**
  - Προσθήκη *όρου ομαλοποίησης*
- **Deep Boosting**
  - Χρήση *βαθιών* δέντρων απόφασης
- **Gradient Boosting**
  - Μετατροπή σε *πρόβλημα βελτιστοποίησης* κατάλληλης συνάρτησης κόστους

Βιβλιογραφία

1. M. Mohri, A. Rostamizadeh, A. Talwalker – *Foundations of Machine Learning*
   - Κεφάλαιο 7
2. Rätsch, Gunnar, Takashi Onoda, and Klaus R. Müller. "Regularizing adaboost." *Advances in neural information processing systems*. 1999.
3. Cortes, Corinna, Mehryar Mohri, and Umar Syed. "*Deep boosting*." (2014).
4. Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." *Annals of statistics* (2001): 1189-1232.