פרק 14: שילוב מסווגים: Bagging ו- Boosting

בפרק זה נציג מעט מהתיאוריה בנושא למידה והכללה על ידי שילוב מסווגים. המטרה הבסיסית היא שימוש במספר מסווגים הנלמדים על אותו data ומשולבים על מנת לקבל ביצועים משופרים ויציבות עדיפה.

הרעיונות המוצגים כאן מתבססים על הפרק הקודם בהשתמשותם בתיאוריה בעלת אופי סטטיסטי. אנו נסתפק בהצגת מספר תוצאות ומושגים יסודיים, וזאת עבור **בעיית הסיווג הבינארי בלבד**.

מקור בסיסי:

Freund and Schapire, A Short Introduction to Boosting,
http://www.cs.princeton.edu/~schapire/uncompress-papers.cgi/FreundSc99.ps

לקריאה נוספת בנושא:

Kearns and Vazirani, An introduction to computational learning theory,
MIT Press, 1994.

14.1 מודל הלומד החלש

נזכור כי בבעיית הלמידה המודרכת אנו נדרשים "ללמוד" פונקציה $f_0 : X \to Y$ בעזרת אוסף דוגמאות $\{x^{(k)}, y^{(k)}\}_{k=1}^{m}$. המודל הבסיסי בו נעסוק כולל את המרכיבים הבאים:

א. פונקצית המטרה: פונקציה $f_0 : X \to Y$ ממרחב הקלט X, למרחב היציאה Y, אותה ברצוננו ללמוד. בהרצאה זו נניח כי $Y = \{-1,+1\}$ (בעיית סיווג הבינארי).

ב. לומדים חלשים: אוסף H של פונקציות $H : X \to Y$, שמתוכו נבחר את הפונקציה $\hat{h}$ בכל שלב. H תכונה כאן מחלקת ההשערות החלשות.

ג. פונקציית השערוך: פונקציית השערוך של ההיפותיזה תהיה מהצורה:

$$h(x) = \operatorname{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

כאשר בחירת ההיפותיזות h_t והפרמטרים α_t תלויית אלגוריתם.

בהרצאה הקודמת התבוננו במודל למידת ה- PAC. אמרנו שאלגוריתם לומד (חזק) אם לכל $\varepsilon,\delta>0$ האלגוריתם ילמד בעזרת מספיק מידע היפותיזה ε אופטימלית בהסתברות לפחות $1-\delta$. מטרת לומד חלש היא צנועה בהרבה:

אלגוריתם למידה לומד חלש אם לכל פילוג (בכלל) D_t על הנקודות (בפרט) מתקיים כי השגיאה:

$$\varepsilon_t = \Pr_{i\sim D_t}\left[h_t\left(x_i\right)\neq y_i\right]=\sum_{i:h_t\left(x_i\right)\neq y_i} D_t\left(i\right)$$

מקיימת כי $\varepsilon_t<\frac{1}{2}-\gamma$. נשים לב כי המשמעות היא מסווג הטוב במעט ממסווג אקראי.

שאלה: האם $\gamma>0$ גורר כי למידה חלשה ← למידה חזקה?

התשובה באופן מפתיע היא כן!

הרעיון הבסיסי מאחורי כל אלגורתמי ה- boosting הוא שמובטח לנו כי הלומד החלש יכול ללמוד (במשהו) ביחס לכל פילוג. זו הנחה חזקה למדי. אלגוריתם ה-boosting הראשון של Rob Schapire השתמש בטכניקה של filtering. פרטים נוספים ניתן למצוא במאמר המקורי: http://www.cs.princeton.edu/~schapire/papers/strengthofweak.pdf

טכניקת ה-filtering היא נוחה לעבודה מבחינה קונספטואלית אך איך מעשית. טכניקות נוספות הן דגימה (Sampling) עליה נלמד בסעיף הבא, ומשקול מחדש (re-weighting) עליה נלמד בסעיף שאחריו.

14.2 Bagging

מקור השם: **Bootstrap aggregating (bagging)**. הרעיון הוא לקחת מדגם עם n דגימות ולדגום מתוכו 'n דגימות (עם החלפה: אותה דגימה יכולה להידגם פעמיים) m פעמים. מבצעים סיווג (או רגרסיה) לכל אחד מ- m המדגמים ואז לוקחים את ה- majority vote (סיווג) או הממוצע (ברגרסיה) של m המסווגים השונים.

יתרונות השיטה:

1. יציבות
2. הורדת וריאנס ומניעת התאמת יתר
3. התגברות על outliers

נשים לב שעבור מודלים ליניאריים (רגרסיה) הממוצע ישאר ליניארי ולכן שיטה זו פחות אפקטיבית.

14.3 אלגוריתם ה- Adaboost

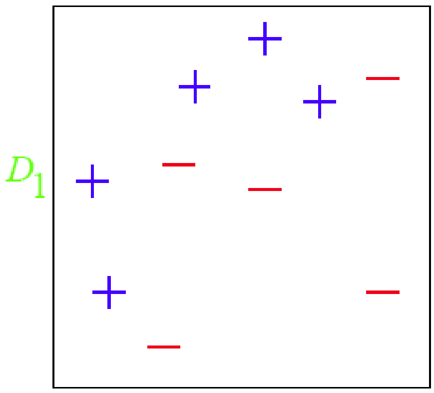
ישנם מספר רב של אלגוריתמי boosting העובדים על אותו רעיון :

1. שמירת משקל לדגימות (פילוג D_t)
2. מציאת מסווג חלש ביחס לפילוג D_t הנוכחי
3. שינוי המשקל תוך הדגשת דוגמאות שסווגו לא נכון
4. חזרה ל- 1.

המסווג הסופי הוא קומבינציה ליניארית של המסווגים החלשים.

כעת נראה אילוסטרציה הלקוחה מתוך www.site.uottawa.ca/~stan/csi5387/**boost**-tut-ppr.pdf.

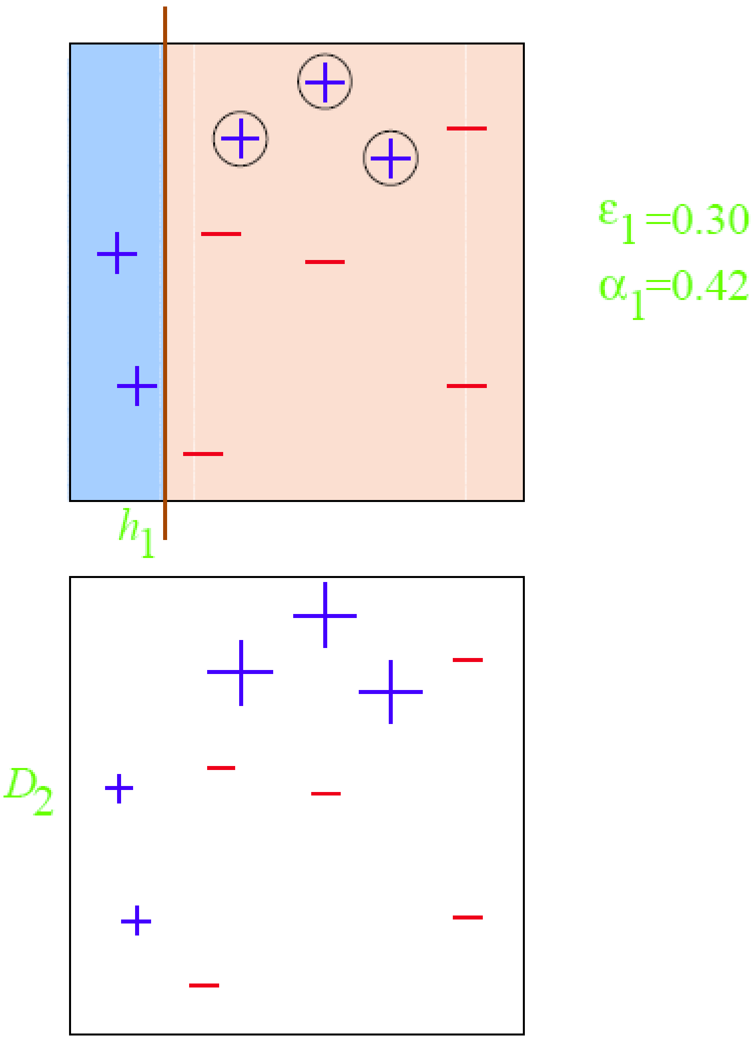
מתחילים ב- data :



נשתמש במסווגים ליניאריים מקבילים לצירים (stumps).

בסיבוב הראשון נבחר את h_1 שיסווג נכונה את שתי הדגימות החיוביות השמאליות ויטעה בשתי בשלושת הדגימות החיוביות הנותרות.

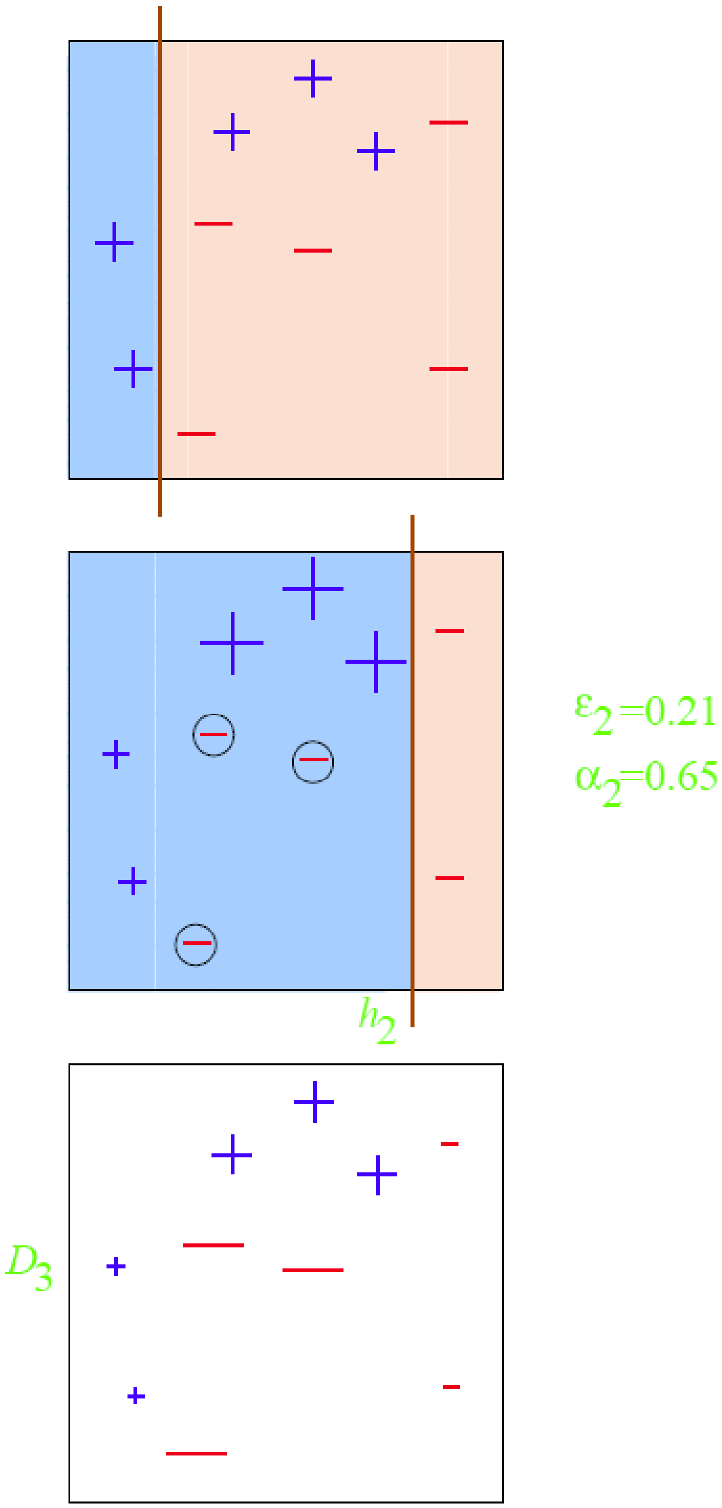
Round 1



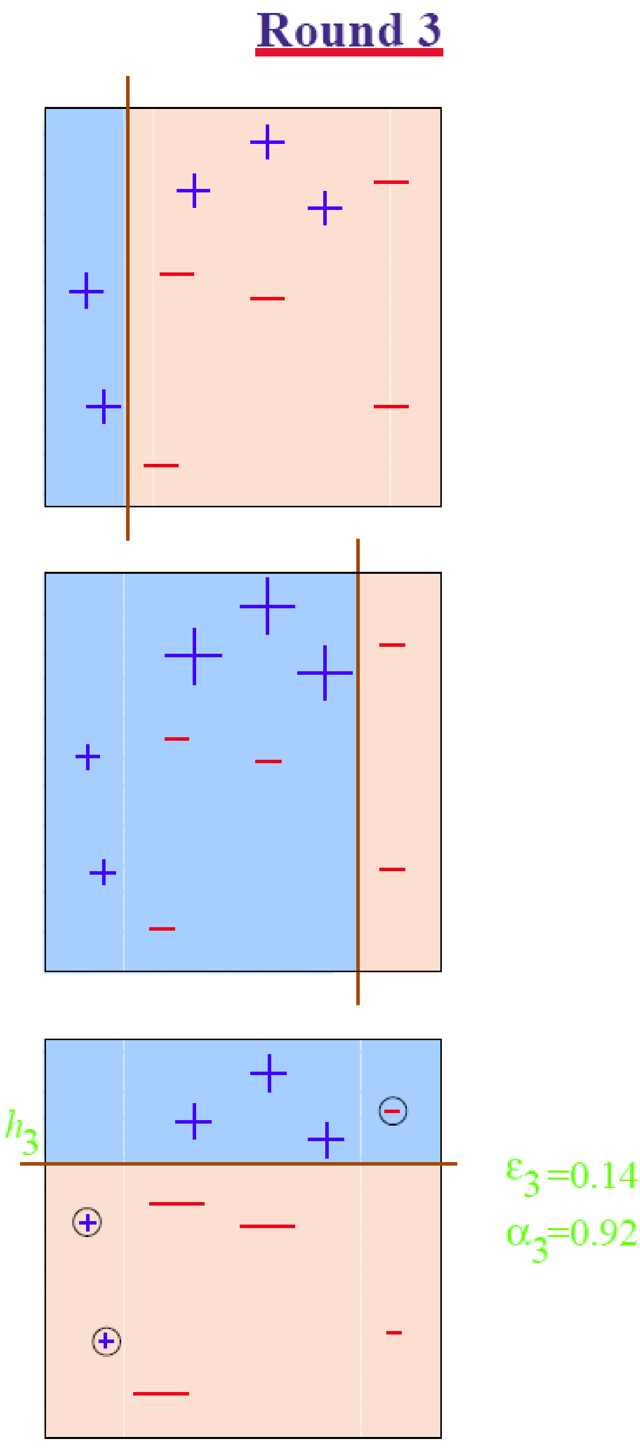
נעיר כי גודל התגיות יחסי למשקל.

בסיבוב השני נבחר את h_2 שיסווג נכון את שתי הדוגמאות השליליות מצד ימין ויטעה בשלושת הדוגמאות השליליות השמאליות.

Round 2

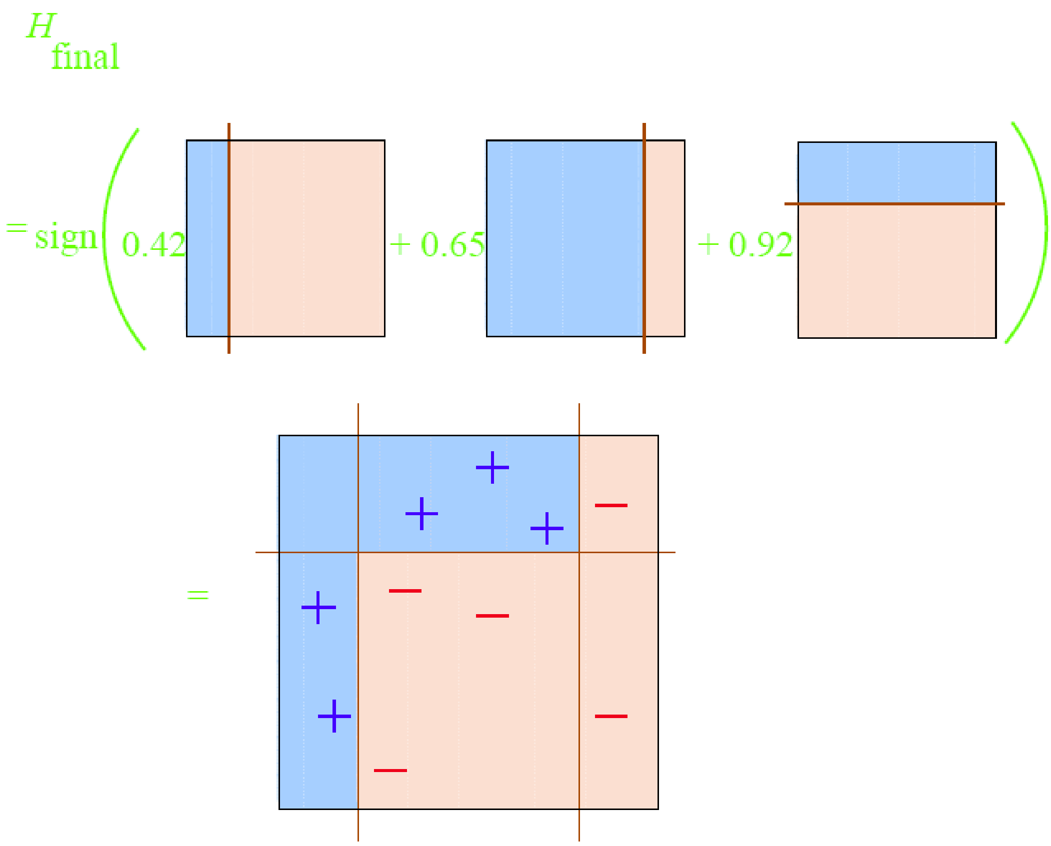


בסיבוב השלישי נבחר את h_3 שיסווג נכון את הדגימות המרכז (שמשקלן גבוה כי טעינו בהן קודם)



ההיפוטיזה הסופית היא הקומבינציה הליניארית של ההיפותיזות:

Final Hypothesis



אלגוריתם ה- Adaboost מאופיין על ידי בחירה מסויימת המאפשרת לו להיות אדאפטיבי (Adaboost = Adaptive Boosting):

1. איתחול: פילוג אחיד $D_1 = 1/m$
2. בהנתן D_t מצא מסווג חלש $h_t : X \to \{1,-1\}$ בעם שגיאה ממוצעת נמוכה ביחס אליו
3. נסמן את השגיאה: $\varepsilon_t = \Pr_{i\sim D_t}\left[h_t(x_i) \neq y_i\right]$
4. קבע: $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$
5. עדכן:

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

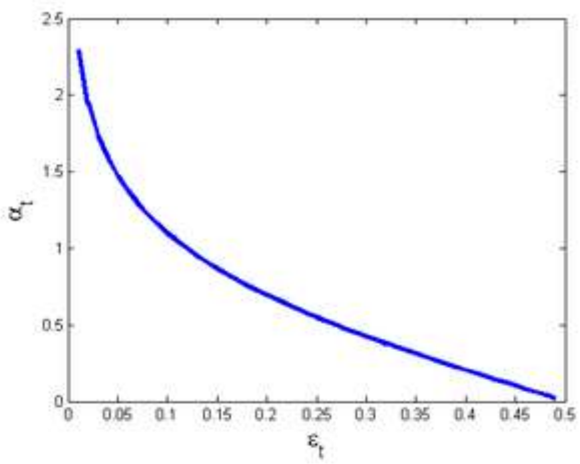
6. חזור לשלב 2 עד שאיזשהו תנאי עצירה מסופק.
7. ההיפותיזה הסופית היא:

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

האלגוריתם אדאפטיבי כי אין צורך לדעת מראש את T (מספר האיטרציות) או את השגיאות או חסם עליהן.

נשים לב כי קיים יחס לינארי בין הקבוע $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$ ויחס הצלחת האימון לשגיאתו.

כמובן, אם השגיאה גדולה מחצי וקרובה לאחד, נוכל להפוך את החיזוי של האלגוריתם ולקבל שגיאה קטנה מחצי וקרובה ל-0. במקרה כזה ערך הקבוע יהיה שלילי.



ניתוח שגיאת האימון: נרשום $\varepsilon_t = \frac{1}{2} - \gamma_t$. אפשר להוכיח ש:

$$\text{Training Error (H)} \le \frac{1}{m}\sum_i e^{-y_i H(x_i)}$$

$$\text{Training Error (H)} \le \prod_{t=1}^{T} 2\sqrt{\varepsilon_t(1-\varepsilon_t)} = \prod_{t=1}^{T}\sqrt{\left(1-4\gamma_t^2\right)} \le \exp\left(-2\sum_{t=1}^{T}\gamma_t^2\right)$$

ז"א אם $\gamma_t > \gamma > 0$ שגיאת האימון תלך ל-0: $\text{Training Error (H)} \le \exp\left(-2T\gamma^2\right)$

רגולריזציה מתבטאת במספר דרכים:

1. בחירת מסווג חלש "חלש" (מסווג חזק מדי יגרום ל- overfitting).

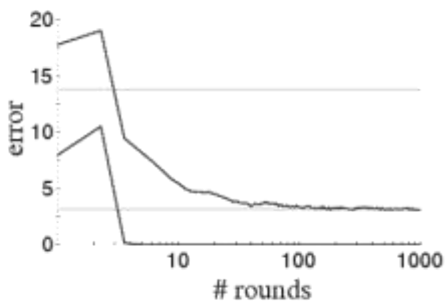
2. בחירת T.

משפט 1 : שגיאת ההכללה היא בהסתברות גבוהה חסומה ע"י :

$$\frac{1}{m}\sum_{i=1}^{m} 1\left\{H\left(x_i\right) \neq y_i\right\} + C\sqrt{\frac{Td}{m}}$$

כאשר התעלמנו מהגורמים ההסתברותיים, d הוא המימד VC של הלומד החלש, 1 היא פונקציית האינדיקטור ו- T הוא מספר האיטרציות.

במקרים רבים ביצועי אלגוריתם ה- Adaboost ישתפרו אם נמשיך לאמן גם אחרי ששגיאת האימון היא 0. ראו הגרף הבא ממסמך של Rob Schapire :



ההסבר לתופעה זו נובע ממשפט ההכללה הבא. עבור דוגמא (x, y) נגדיר את ה- margin באיטרציה t להיות :

$$\text{margin}(x, y; t) = \frac{y\sum_{\tau=1}^{t} \alpha_\tau h_\tau\left(x\right)}{\sum_{\tau=1}^{t} \alpha_\tau}$$

ונשים לב שזהו מספר באינטרול [-1,1] שהינו חיובי אם H מסווג את x נכון. ככול שה- margin גדול יותר, הבטחון בסיווג רב יותר. ניתן להוכיח את המשפט הבא :

משפט 2 : שגיאת ההכללה באיטרציה t היא בהסתברות גבוהה חסומה ע"י :

$$\frac{1}{m}\sum_{i=1}^{m} 1\left\{\text{margin}(x_i, y_i; t) \leq \theta\right\} + C\sqrt{\frac{d}{m\theta^2}}$$

כאשר 1 היא פונקציית האינדיקטור והתעלמנו מגורמים הסתברותיים.

הערה : חסם ההכללה אינו תלוי במספר האיטרציות אלא רק ב- margin.