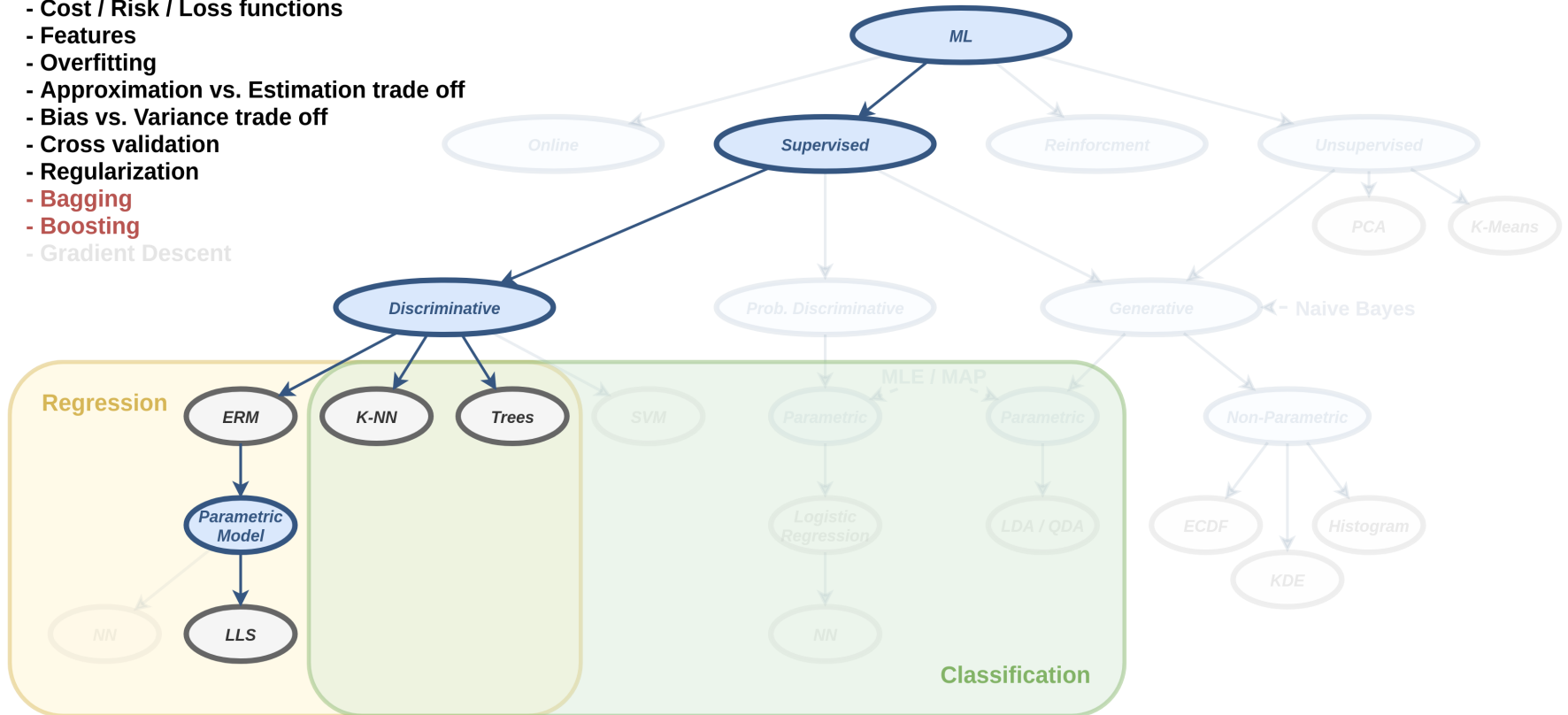


הרצאה 5 - Bagging and Boosting

Subjects Covered in this Course

General concepts:

- Cost / Risk / Loss functions
- Features
- Overfitting
- Approximation vs. Estimation trade off
- Bias vs. Variance trade off
- Cross validation
- Regularization
- **Bagging**
- **Boosting**
- Gradient Descent

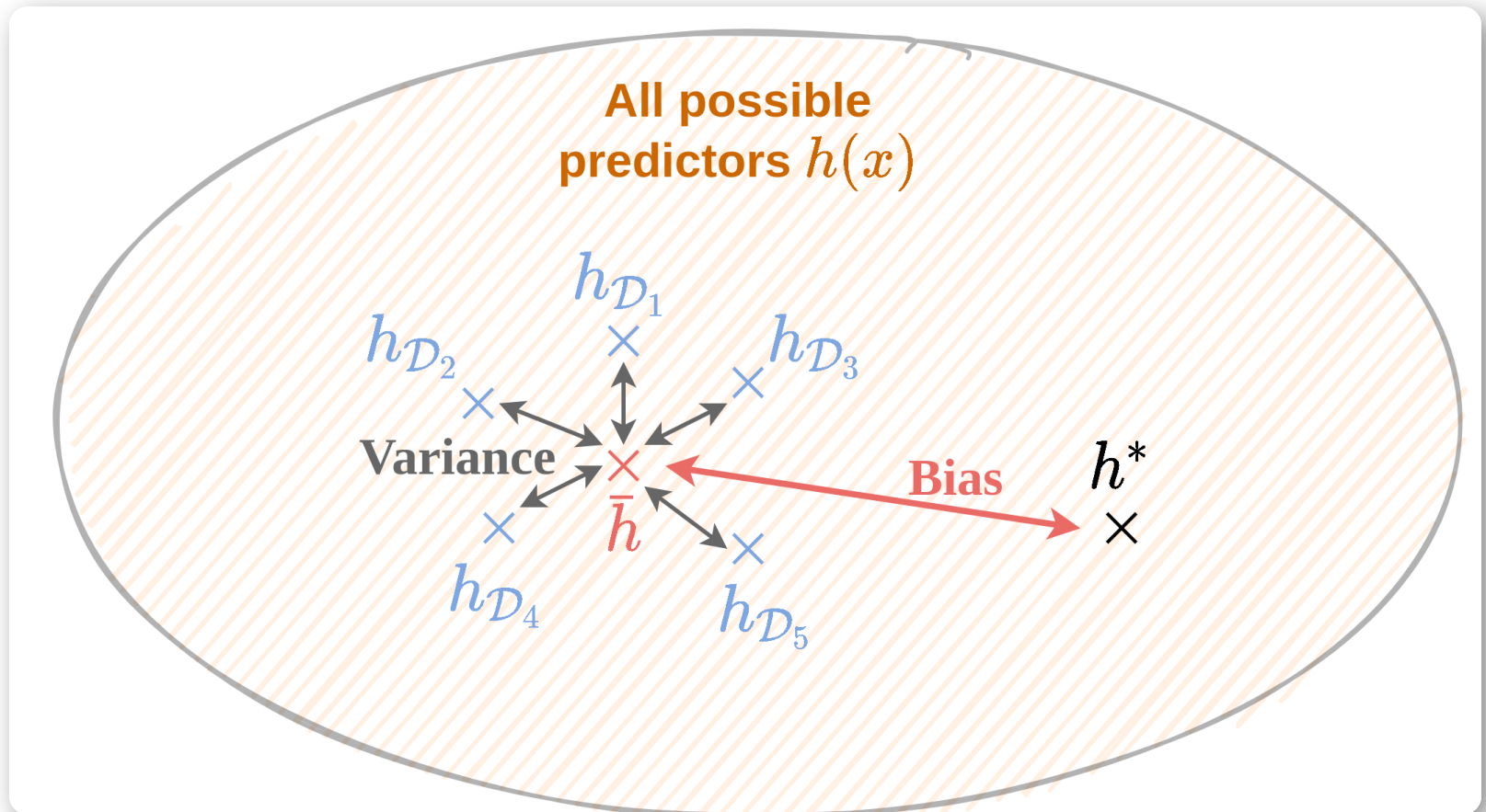


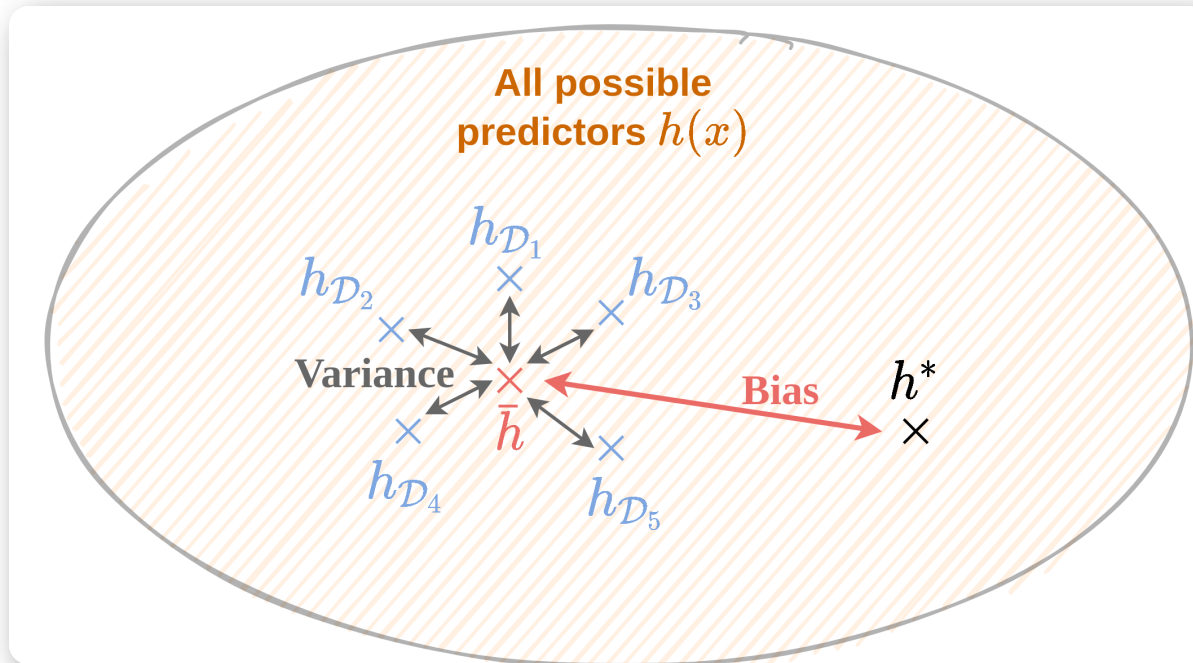
Ensemble Methods

בהרצאה הזו נציג שתי שיטות אשר בעזרתן ניתן לשפר את הביצועים של אלגוריתמים קיימים על ידי שימוש בסט של חזאים. סט זה מכונה לרוב ensemble (מכלול).

תזכורת הטיה ושונויות

- מתייחסים לפילוג של שגיאת החיזוי על פני מדגמים שונים.
- נתייחס למדגם כאל משתנה אקראי.





- החיזוי הממוצע: $\bar{h}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}} [h(\mathbf{x})]$
- החיזוי האופטימלי $h^*(\mathbf{x})$.
- ההטיה היא ההפרש בין החיזוי הממוצע לחיזוי האופטימלי.
- השונות היא $\mathbb{E}_{\mathcal{D}} [(h_{\mathcal{D}}(\mathbf{x}) - \bar{h}(\mathbf{x}))^2]$

תזכורת הטיה ושונות

- מודלים בעלי יכולת ייצוג נמוכה יסבלו לרוב מהתאמת חסר אשר יתבטא בהטיה גבוהה ושונות נמוכה
- מודלים בעלי יכולת ייצוג גבוהה יסבלו לרוב מהתאמת יתר אשר יתבטא בשונות מאד גבוהה והטיה נמוכה

Boosting | Bagging

כעת נוכל להסביר מה **boosting** | **bagging** מנסים לעשות:

- ב **bagging** ננסה לקחת מכלול של חזאים עם שונות גבוהה ולשלב ביניהם כדי ליצור חזאי עם שונות נמוכה יותר.
- ב **boosting** ננסה לקחת מכלול של חזאים עם הטיה גבוהה ולשלב ביניהם כדי ליצור חזאי עם הטיה נמוכה יותר.

שאלה: דוגמאות לכל אחד מהמקרים

נהיה מעוניינים לייצר מכלול (ensemble) של חזאים בעלי הטיה נמוכה אך שונות גבוהה ואז לשלב ביניהם על מנת להקטין את השונות.

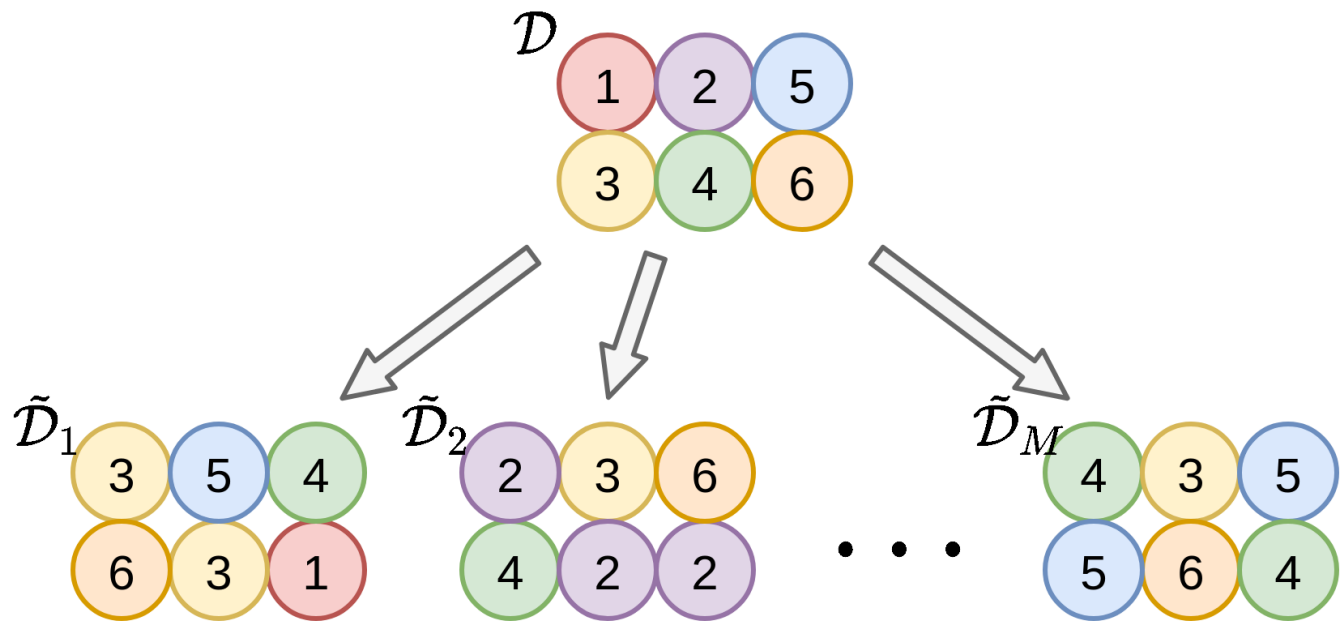
אחת הבחירות הנפוצות לחזאים שכאלה ב bagging היא עצי החלטה עמוקים (ללא pruning).

השם Bagging הוא הלחם של המילים **bootstrapping** ו **aggregation**, שהם שני שלבי השיטה.

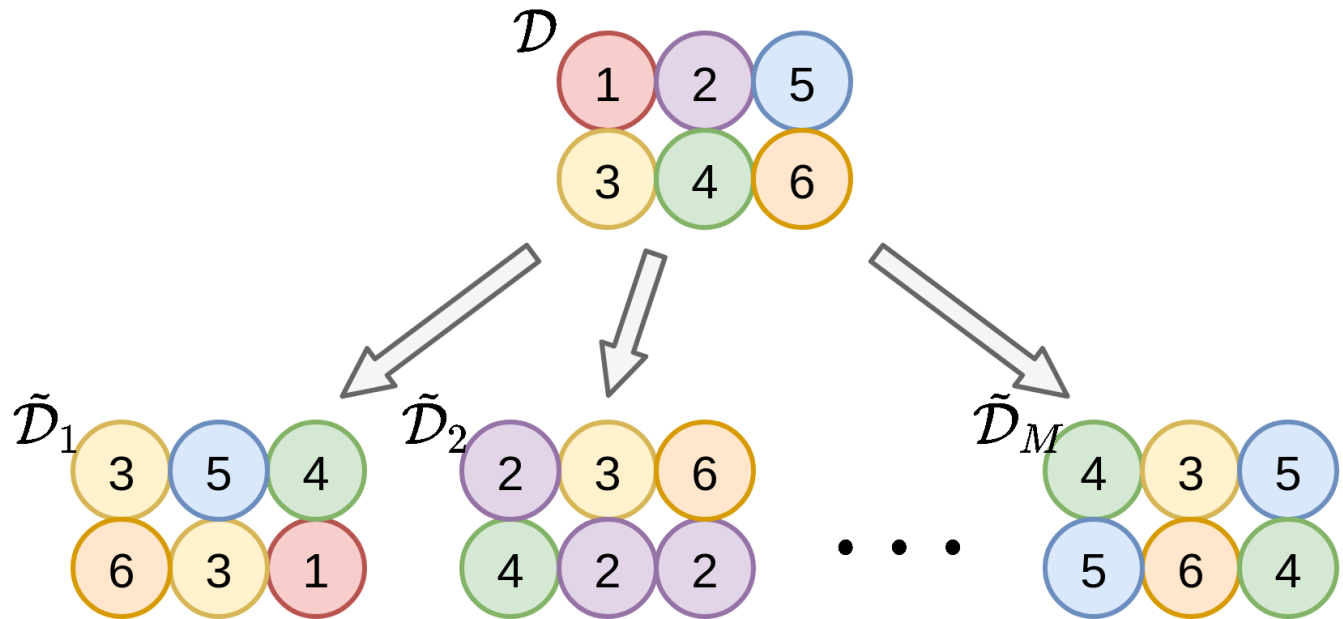
- לוו יכלנו לייצר כמה מדגמים בלתי תלויים, היינו יכולים לבנות חזאי עבור כל מדגם ולמצע על החזאים על מנת להקטין את השונות של השגיאת החיזוי.
- בפועל לרוב יהיה בידינו רק מדגם יחיד שאיתו נצטרך לעבוד.
- ניתן לייצר מספר מדגמים על ידי חלוקת המדגם הקיים, אך לרוב העובדה שהמדגמים הם משמעותית קטנים מהמדגם המקורי **תגדיל** את השונות.

Bootstrapping

אופציה חלופית אשר שומרת על גודל המדגם, אך מתפשרת על דרישת חוסר התלות בין המדגמים ובין הדגימות.



Bootstrapping

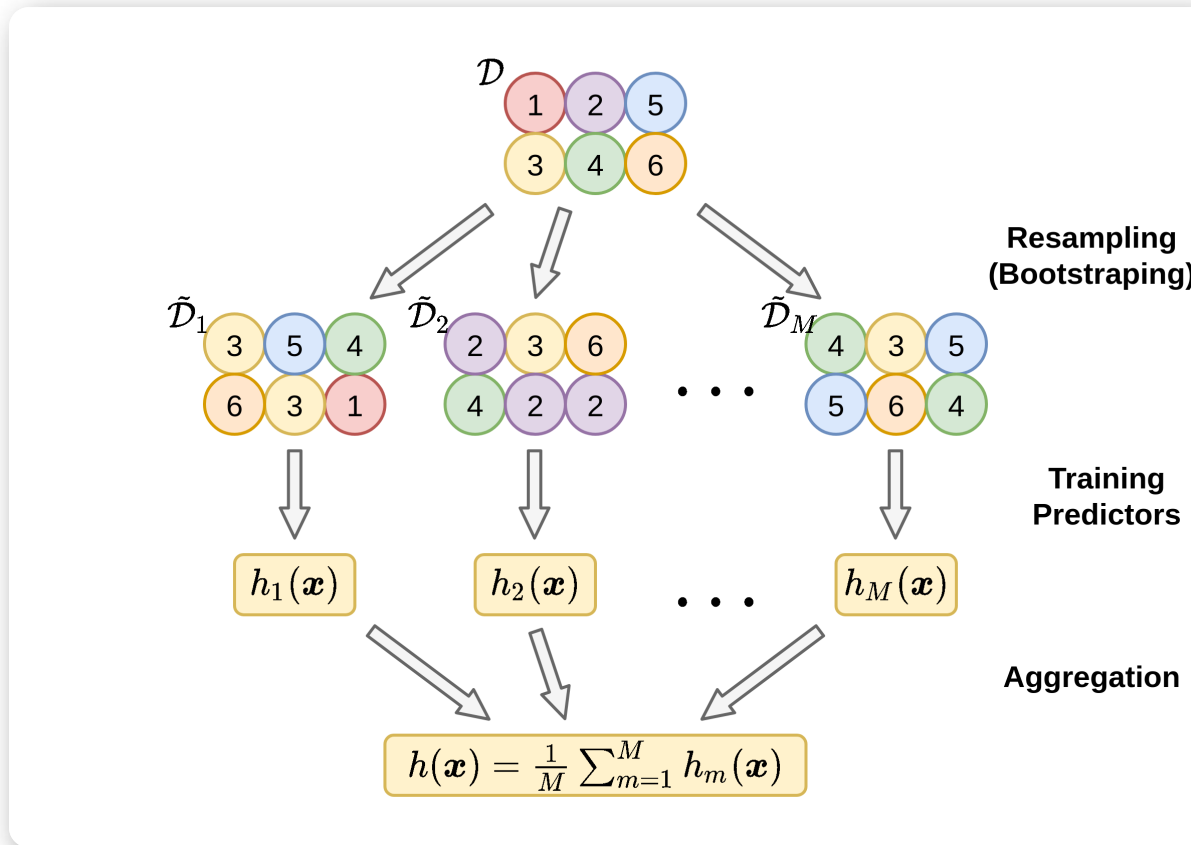


בשיטה זו אנו נייצר מדגמים חדשים על ידי דגימה מחדש של המדגם הנתון. הדגימה הינה **עם חזרות**.
שאלה: מהי דגימה עם חזרות ודגימה ללא חזרות? מה ניתן לומר על התלות בין הדגימות המתקבלות?

Bootstrapping

- הסיכוי של דגימה כלשהיא להופיע ב \tilde{D} הינה $1 - (1 - \frac{1}{N})^{\tilde{N}}$.
- כאשר $\tilde{N} = N$ ו $N \rightarrow \infty$, סיכוי זה הולך ל $1 - e^{-1} \approx 63\%$.

Aggregation: שילוב חזאים לחזאי יחיד

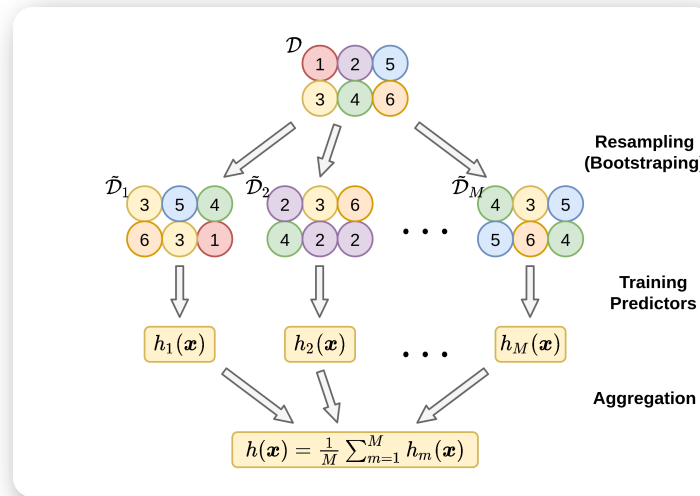


• נייצר M מדגמים חדשים בגודל זהה למדגם המקורי $\tilde{N} = N$.

• עבור כל אחד מהמדגמים $\tilde{\mathcal{D}}_m$ נבנה חזאי \tilde{h}_m .

• נקבץ את כל החזאים על מנת לקבל את החזאי הכולל.

Aggregation: שילוב חזאים לחזאי יחיד

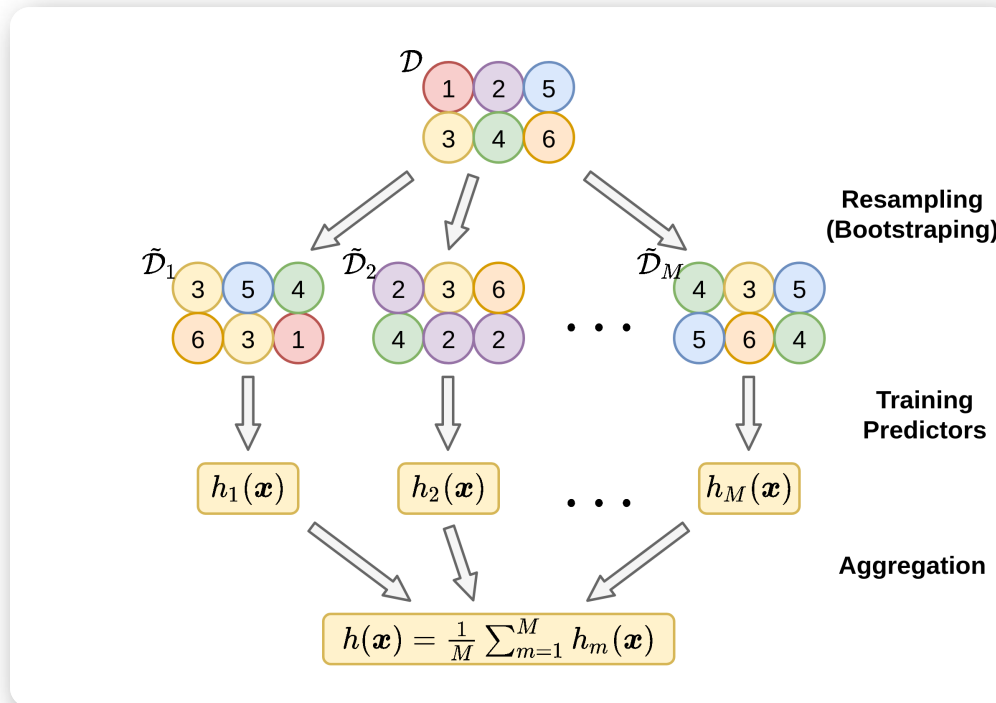


• **עבור בעיות רגרסיה: נמצע את תוצאת החיזוי של כל החזאים:**

$$h(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^M \tilde{h}_m(\mathbf{x})$$

• **עבור בעיות סיווג: נבצע majority voting:**

Aggregation: שילוב חזאים לחזאי יחיד



- מספר המדגמים M נע בין עשרות מדגמים לאלפים.

- לרוב נשתמש באותה השיטה על מנת לבנות את כל החזאים.

- כל מסווג יכול להיות עץ החלטה מלא!

Out Of Bag Error Estimation (לקריאה עצמית - לא למבחן)

- ניתן להעריך את ביצועי המודל ללא צורך ב `test / validation set`.
- הרעיון הינו להשתמש בעובדה שכל אחד מהמדגמים מכיל רק חלק מהדגימות.

Random Forest (לקריאה עצמית - לא למבחן)

- שילוב של עצי החלטה עם bagging + תוספת.
- תוספת - בחירה אקראית של תת-קבוצות של רכיבים בפיצולים בצומת - מקטין את הקורלציה בין העצים השונים.
- שיטה מאד יעילה ונפוצה.

Random Forest (לקריאה עצמית - לא למבחן)

מדוע השיטה עובדת?

- הפחתה בהתאמת יתר
- דיוק גבוה
- חסינות בפני נתונים בעייתיים (כגון נתונים חסרים, **ouliers**)
- קיימות הבטחות תאורטיות (לא נדון)
- מה החסרון לעומת עצי החלטה?

- ננסה להשתמש במכלול של חזאים בעלי **הטיה גבוהה** אך **שונות נמוכה** כדי ליצור חזאי כולל בעל שונות נמוכה.
- נתמקד בבעיות סיווג בינארי.
- התוויות בבעיה הן (מטעמי סימטריה) $y = \pm 1$.

בעיית ה **boosting** המקורית (לקריאה עצמאית - לא למבחן)

- מהו **לומד חזק**?
- מהו **לומד חלש**?
- האם כל לומד הוא בעצם גם לומד חזק?

הרחבה ברשימות

(AdaBoost (adaptive-boosting

בהינתן אוסף של חזאים "חלשים" $\tilde{h}(x)$ בעלי הטיה גבוהה, שיטה זו מנסה לבנות חזאי מהצורה

$$h(x) = \text{sign} \left(\sum_{m=1}^M \alpha_m \tilde{h}_m(x) \right)$$

כך ש $h(x)$ יהיה בעל הטיה נמוכה.

- בחירה פופולרית של מסווגים כאלה הינה עצי החלטה בעומק 1 המכונים **stumps** (עצים עם פיצול יחיד).

החסם על ה misclassification rate

נראה שעבור חזאי מהצורה של

$$h(\mathbf{x}) = \text{sign} \left(\sum_{m=1}^M \alpha_m \tilde{h}_m(\mathbf{x}) \right)$$

שגיאת 0-1 האמפירית חסומה מלמעלה על ידי:

$$\frac{1}{N} \sum_{i=1}^N \exp \left(- \sum_{m=1}^M \alpha_m y^{(i)} \tilde{h}_m(\mathbf{x}^{(i)}) \right)$$

החסם על ה misclassification rate

נתחיל בעובדה שעבור $y = \pm 1$ וערך כל שהוא z מתקיים ש:

$$I\{\text{sign}(z) \neq y\} = I\{\text{sign}(yz) \neq 1\} \leq \exp(-yz)$$

ונשתמש באי השיוון הנ"ל על כל איבר בסכום בנוסחא לשגיאה האמפירית ונקבל כי:

$$\begin{aligned} \frac{1}{N} \sum_i I\{h(\mathbf{x}^{(i)}) \neq y^{(i)}\} &= \frac{1}{N} \sum_i I\left\{\text{sign}\left(\sum_{m=1}^M \alpha_m \tilde{h}_m(\mathbf{x}^{(i)})\right) \neq y^{(i)}\right\} \\ &\leq \frac{1}{N} \sum_{i=1}^N \exp\left(-\sum_{m=1}^M y^{(i)} \alpha_m \tilde{h}_m(\mathbf{x}^{(i)})\right) \end{aligned}$$

בעיית האופטימיזציה של AdaBoost

על פי החסם שהצגנו לעיל נוכל להניח כי מזעור של בעיית האופטימיזציה הבאה:

$$\arg \min_{\{\alpha_m, \tilde{h}_m\}_{m=1}^M} \frac{1}{N} \sum_{i=1}^N \exp \left(- \sum_{m=1}^M \alpha_m y^{(i)} \tilde{h}_m(\mathbf{x}^{(i)}) \right)$$

תגרום כנראה להקטנת השגיאה האמפירית ובכך להקטנת ההטיה.

• אנו נראה בהמשך כי תחת תנאים מסויימים כאשר $M \rightarrow \infty$ החסם ידעך ל-0.

בעיית האופטימיזציה של AdaBoost

$$\arg \min_{\{\alpha_m, \tilde{h}_m\}_{m=1}^M} \frac{1}{N} \sum_{i=1}^N \exp \left(- \sum_{m=1}^M \alpha_m y^{(i)} \tilde{h}_m(\mathbf{x}^{(i)}) \right)$$

- **AdaBoost מנסה לפתור את בעיית האופטימיזציה בצורה חמדנית.**
- **אנו נגדיל את M בהדרגה כאשר בכל פעם נחפש את ה α_m וה \tilde{h}_m האופטימלים.**

בעיית האופטימיזציה של AdaBoost

$$\arg \min_{\{\alpha_m, \tilde{h}_m\}_{m=1}^M} \frac{1}{N} \sum_{i=1}^N \exp \left(- \sum_{m=1}^M \alpha_m y^{(i)} \tilde{h}_m(\mathbf{x}^{(i)}) \right)$$

נסתכל על המצב בו כבר מצאנו את כל ה α_m וה \tilde{h}_m עד ל $M - 1$, וכעת אנו רוצים למצוא את α_M ו \tilde{h}_M :

$$\alpha_M, \tilde{h}_M = \arg \min_{\alpha, \tilde{h}} \frac{1}{N} \sum_{i=1}^N \exp \left(- \sum_{m=1}^{M-1} \alpha_m y^{(i)} \tilde{h}_m(\mathbf{x}^{(i)}) - \alpha y^{(i)} \tilde{h}(\mathbf{x}^{(i)}) \right)$$

- α_M יכול לקבל כל ערך.
- את \tilde{h}_M עלינו לבחור מתוך מאגר מסווגים נתון.

בעיית האופטימיזציה של AdaBoost

$$\alpha_M, \tilde{h}_M = \arg \min_{\alpha, \tilde{h}} \frac{1}{N} \sum_{i=1}^N \exp \left(- \sum_{m=1}^{M-1} \alpha_m y^{(i)} \tilde{h}_m(\mathbf{x}^{(i)}) - \alpha y^{(i)} \tilde{h}(\mathbf{x}^{(i)}) \right)$$

הדרך לפתור את בעיית האופטימיזציה:

1. רישום מחדש של בעיית האופטימיזציה בצורה יותר פשוטה.
 2. מציאת α_M כפונקציה של \tilde{h}_M על ידי גזירה והשוואה ל-0.
 3. הצבה של α_M בחזרה לבעיית האופטימיזציה על מנת לקבל ביטוי פשוט שאותו יש למזער כתלות ב \tilde{h}_M .
- נציג כעת את הפתרון של בעיה זו, כאשר הפיתוח המלא של הפתרון מופיע בסוף ההרצאה.

בעיית האופטימיזציה של AdaBoost

$$\alpha_M, \tilde{h}_M = \arg \min_{\alpha, \tilde{h}} \frac{1}{N} \sum_{i=1}^N \exp \left(- \sum_{m=1}^{M-1} \alpha_m y^{(i)} \tilde{h}_m(\mathbf{x}^{(i)}) - \alpha y^{(i)} \tilde{h}(\mathbf{x}^{(i)}) \right)$$

נגדיר את הגדלים הבאים:

$$\tilde{w}_i^{(M-1)} = \exp \left(- \sum_{m=1}^{M-1} \alpha_m y^{(i)} \tilde{h}_m(\mathbf{x}^{(i)}) \right)$$

$$w_i^{(M-1)} = \frac{\tilde{w}_i^{(M-1)}}{\sum_{j=1}^N \tilde{w}_j^{(M-1)}}$$

משקל גבוה לדגימות קשות

$$\varepsilon(\tilde{h}, \{w_i\}) = \sum_{i=1}^N w_i I\{y^{(i)} \neq \tilde{h}(\mathbf{x}^{(i)})\}$$

בעיית האופטימיזציה של AdaBoost

α_M ו \tilde{h}_M האופטימאליים בכל שלב יהיו נתונים על ידי:

$$\tilde{h}_M = \arg \min_{\tilde{h}} \varepsilon(\tilde{h}, \{w_i^{(M-1)}\}) = \arg \min_{\tilde{h}} \sum_{i=1}^N w_i^{(M-1)} I\{y^{(i)} \neq \tilde{h}(\mathbf{x}^{(i)})\}$$

$$\alpha_M = \frac{1}{2} \ln \left(\frac{1 - \varepsilon_M}{\varepsilon_M} \right)$$

כאשר סימנו:

$$\varepsilon_M = \varepsilon(\tilde{h}_M, \{w_i^{(M-1)}\})$$

כאשר \tilde{h}_M הינו מסווג טוב לדגימות קשות. הוכחה מלאה ברשימות.

בעיית האופטימיזציה של AdaBoost

אם כן, בכל שלב עלינו לבצע את הפעולות הבאות:

1. חישוב המשקלים $\{w_i^{(M-1)}\}$.

2. מציאת החזאי \tilde{h} אשר ממזער את שגיאת 0-1 האמפירית הממושקלת.

3. חישוב המקדם α_M .

בפועל ניתן לחשב את המשקלים של ה צעד ה M כבר בסוף הצעד ה $M - 1$. בנוסף ניתן להשתמש בעובדה ש:

$$\tilde{w}_i^{(M)} = \tilde{w}_i^{(M-1)} \exp \left(-\alpha_M y^{(i)} \tilde{h}_M(\mathbf{x}^{(i)}) \right)$$

כדי להימנע מלחשב את הסכום על m ולקצר את החישוב.

האלגוריתם של AdaBoost

נאתחל את המשקולות $w_i^{(0)} = \frac{1}{N}$.

1. נבחר את המסווג אשר ממזער את:

$$\tilde{h}_M = \arg \min_{\tilde{h}} \sum_{i=1}^N w_i^{(M-1)} I\{y^{(i)} \neq \tilde{h}(\mathbf{x}^{(i)})\}$$

2. נחשב את המקדם α_M של המסווג:

$$\varepsilon_M = \sum_{i=1}^N w_i^{(M-1)} I\{y^{(i)} \neq \tilde{h}_M(\mathbf{x}^{(i)})\}$$
$$\alpha_M = \frac{1}{2} \ln \left(\frac{1 - \varepsilon_M}{\varepsilon_M} \right)$$

האלגוריתם של AdaBoost

3. נעדכן את וקטור המשקלים:

$$\tilde{w}_i^{(M)} = w_i^{(M-1)} \exp\left(-\alpha_M y^{(i)} \tilde{h}_M(\mathbf{x}^{(i)})\right)$$

$$w_i^{(M)} = \frac{\tilde{w}_i^{(M)}}{\sum_{j=1}^N \tilde{w}_j^{(M)}}$$

המשמעות של המשקלים

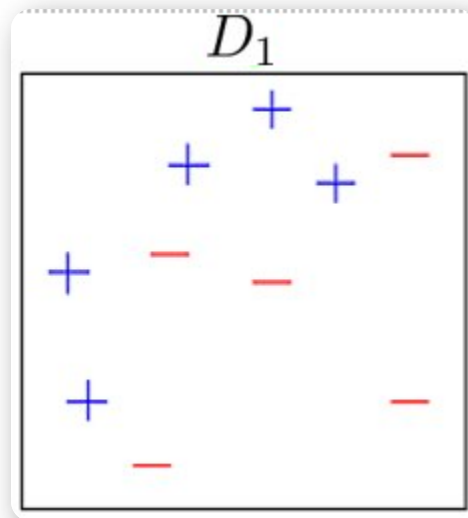
המשקל ללא הנרמול של הדגימה ה i שווה ל:

$$\tilde{w}_i^{(M)} = \exp \left(-y^{(i)} \sum_{m=1}^M \alpha_m \tilde{h}_m(\mathbf{x}^{(i)}) \right)$$

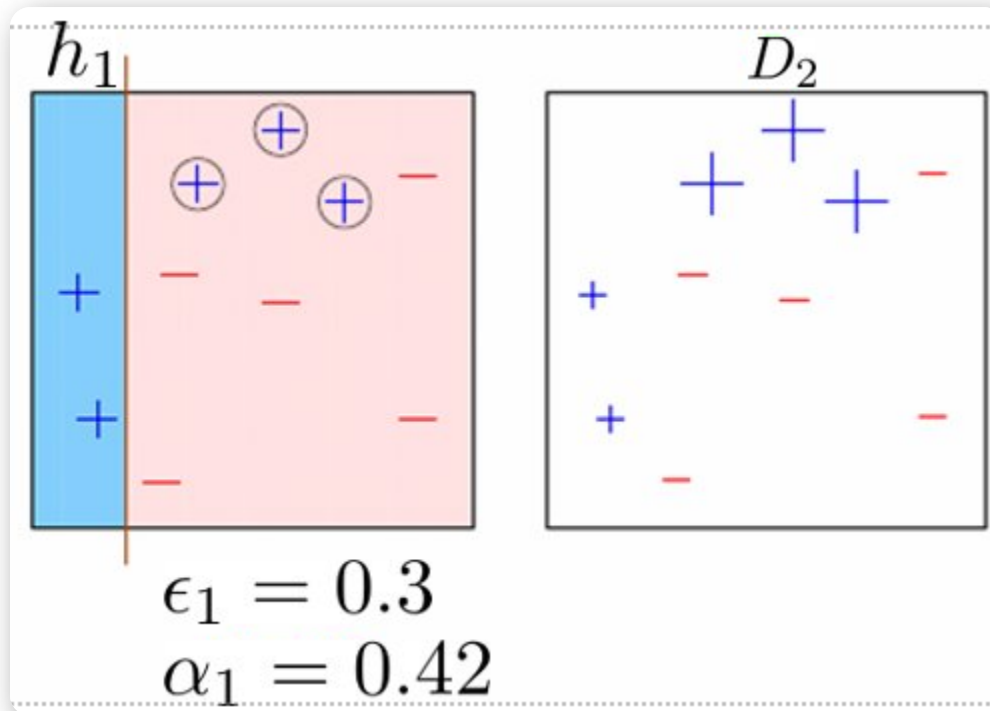
- משקל זה מציין עד כמה טוב האלגוריתם מסווג את הדגימה ה i .
- דגימות שלא מסווגות נכון יהיו בעלות משקל גדול.
- התפקיד של המשקלים הוא לדאוג שהאלגוריתם יבחר בכל צעד את החזאי אשר ישפר את הסיווג בעיקר על הדגימות שעליהן החזאי הנוכחי טועה.

- בחלק גדול מהמקרים AdaBoost ילך ויקטין את שגיאת החיזוי על המדגם עד שהוא יגיע לסיווג מושלם.
- AdaBoost ממשיך לשפר את יכולת ההכללה שלהו גם אחרי שהוא הגיע לסיווג מושלם.
- כמות התאמת היתר גדלה בקצב מאד איטי (אם בכלל).
- נרצה להריץ את האלגוריתם מספר רב של צעדים ולבדוק את הביצועים על validation set.

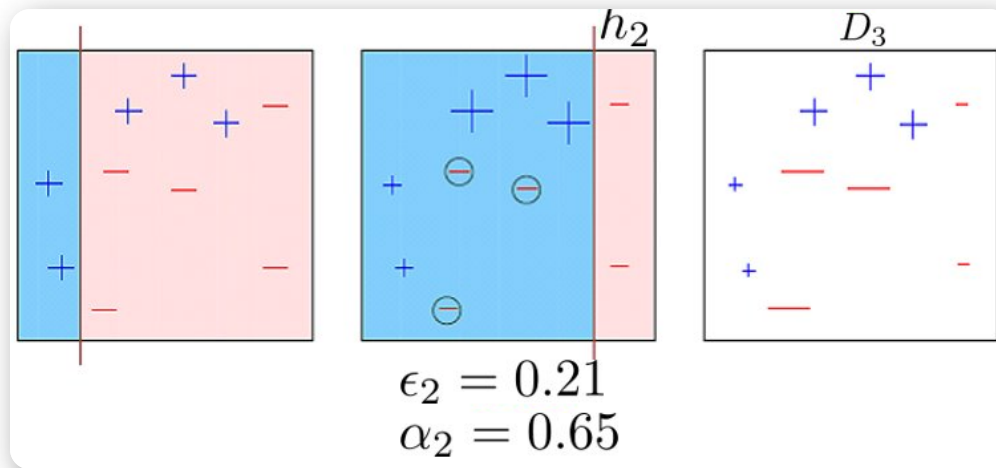
Before training



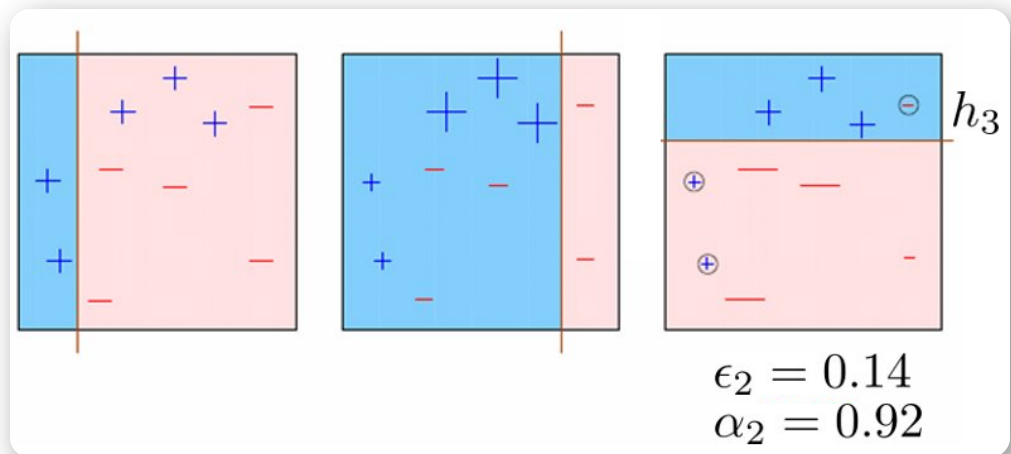
Round 1



Round 2



Round 3



$$H = \text{sign} \left(\begin{array}{c} \begin{array}{|c|c|} \hline \text{blue} & \text{red} \\ \hline \end{array} \\ +0.42 \\ \begin{array}{|c|c|} \hline \text{blue} & \text{red} \\ \hline \end{array} \\ +0.65 \\ \begin{array}{|c|c|} \hline \text{blue} & \text{red} \\ \hline \end{array} \\ +0.92 \end{array} \right)$$

$$= \begin{array}{|c|c|} \hline \text{blue} & \text{red} \\ \hline \end{array}$$

The diagram illustrates the sign function applied to a sum of three rectangular regions. The first region is blue on the left and red on the right, with a weight of +0.42. The second region is blue on the left and red on the right, with a weight of +0.65. The third region is blue on the left and red on the right, with a weight of +0.92. The resulting region is a 2x2 grid where the top-left and bottom-left quadrants are blue and marked with '+', and the top-right and bottom-right quadrants are red and marked with '-'.

קצב ההתכנסות של החסם

ראינו קודם כי שגיאת 0-1 האמפירית (הלא ממושקלת) של החזאי על המדגם חסומה על ידי הביטוי:

$$\frac{1}{N} \sum_i I\{h(\mathbf{x}^{(i)}) \neq y^{(i)}\} \leq \frac{1}{N} \sum_{i=1}^N \exp \left(- \sum_{m=1}^M y^{(i)} \alpha_m \tilde{h}_m(\mathbf{x}^{(i)}) \right)$$

נראה כעת כי תחת תנאים מסויימים מובטח כי חסם זה ידעך ל-0 ככל שנגדיל את M .

טענה (הוכחה מלאה ברשימות)

נסמן את שגיאת 0-1 האמפירית הממושקלת בצעד ה m ב $\varepsilon_m = \frac{1}{2} - \gamma_m$. נטען כי מתקיים הקשר הבא:

$$\frac{1}{N} \sum_{i=1}^N \exp \left(- \sum_{m=1}^M y^{(i)} \alpha_m \tilde{h}_m(\mathbf{x}^{(i)}) \right) \leq \exp \left(-2 \sum_{m=1}^M \gamma_m^2 \right)$$

מכאן שבמידה וקיים γ אשר מקיים $\gamma_m \geq \gamma > 0$ אזי מתקיים ש:

$$\begin{aligned} & \frac{1}{N} \sum_i I\{h(\mathbf{x}^{(i)}) \neq y^{(i)}\} \\ & \leq \frac{1}{N} \sum_{i=1}^N \exp \left(- \sum_{m=1}^M y^{(i)} \alpha_m \tilde{h}_m(\mathbf{x}^{(i)}) \right) \leq \exp(-2M\gamma^2) \end{aligned}$$

זאת אומרת, שקיים חסם לשגיאת 0-1 האמפירית אשר דועך באופן מעריכי עם M .

- הקטנת ההטיה
- מימוש פשוט ויעיל
- ניתן לשימוש עם מגוון מסווגים בסיסיים

- התאמת יתר ורגישות לרעש (התמקדות בדוגמאות קשות)
- קשה למימוש מקבילי
- רגישות לבחירה של מסווגי בסיס
- פרשנות מורכבת (למשל, ביחס לעצי החלטה)

הרחבות

- רבות ומגוונות. בחלקן משלבות רגולריזציה, אפשרות לעיבוד מקבילי, שילוב עם bagging ועוד