תעודת זהות:
מספר מחברת:

מבוא ללמידה חישובית - מבחן מועד א'
סמסטר א' תשע"ו (2015/6)

בית הספר למדעי המחשב, אוניברסיטת תל־אביב

מרצים: פרופ' ליאור וולף, פרופ' ערן הלפרין
מתרגל: רגב שוייגר

26.1.2016

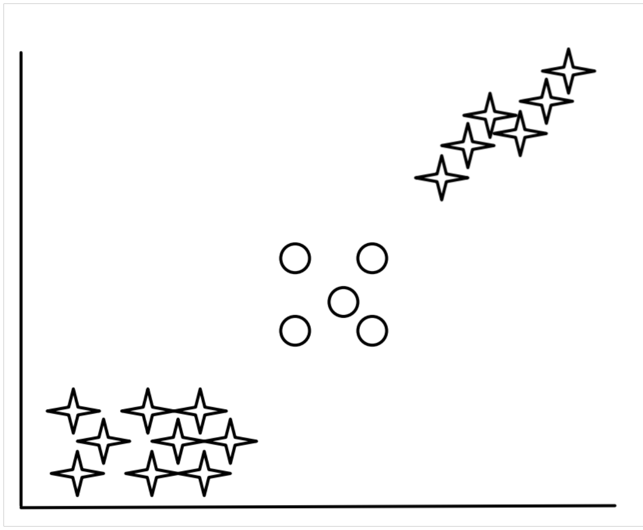
<u>הוראות</u>:

1. מומלץ לקרוא את כל ההנחיות והשאלות בתחילת המבחן לפני תחילת כתיבת התשובות.
2. משך הבחינה – **שלוש שעות**.
3. חומר עזר מותר: דף נוסחאות בגודל A4.
4. **יש לענות על השאלות במקום המיועד לכך בטופס השאלון (טופס זה)**. מחברות הבחינה לא תקראנה, ותשמשנה כטיוטה בלבד.
5. יש למלא בכל דף של השאלון מספר ת.ז. ומספר מחברת.
6. במבחן 5 שאלות:
   - הניקוד לכל שאלה מופיע ליד מספר השאלה.
   - יש לענות תשובות ברורות, ענייניות ותמציתיות.
7. מותר להשתמש בכל טענה שהוכחה בכיתה (בהרצאה, בתרגול, או בתרגיל בית) בתנאי שמצטטים אותה במדויק. טענות אחרות (כאלה שהוכחו בספר, בהרצאות מהסמסטר הקודם, וכו') יש להוכיח.
8. אם לא נאמר אחרת, יש להניח שדגימות במדגם נוצרות באופן בלתי תלוי ומאותה התפלגות (i.i.d).

תעודת זהות:
מספר מחברת:

1 שאלה 1 - 20 נקודות

נתון המדגם הדו־מימדי הבא, בו הנקודות מסווגות לשתי מחלקות:



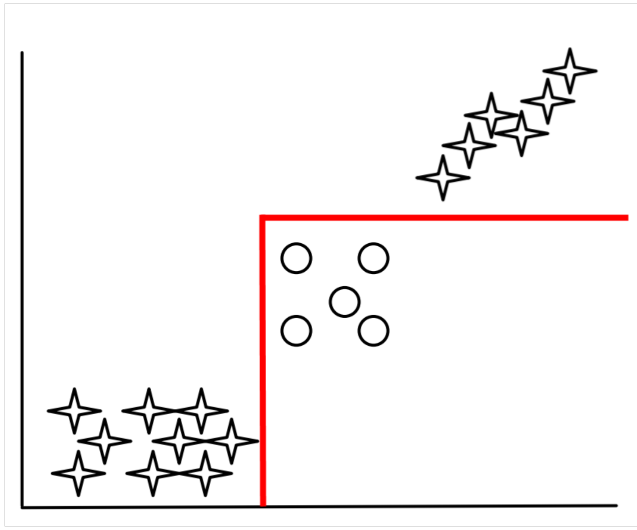
עבור כל אחד מהאלגוריתמים הבאים, קבע/י האם ניתן להריץ אותו עד לקבלת מסווג עם שגיאת למידה אפס, על המדגם הנתון. אם כן, צייר/י קו הפרדה מתאים למסווג המתקבל. אם לא, הסבר/הסבירי מדוע.

1. עץ החלטה, כאשר ה־classifiers מתוכם האלגוריתם בוחר הם כל ה־decision stumps המקבילים לצירים (כלומר, עבור כל מימד i, וסף a, ניתן לבחור classifier שמסווג על פי תוצאת ההשוואה $x_i < a$).

   ☐ לא יכול להגיע לשגיאה אפס. הסבר:

   ☐ יכול להגיע לשגיאה אפס. קו ההפרדה של המסווג:

   יכול. ריצה אפשרית של עץ ההחלטה, תחלק תחילה את המישור בציר ה־y, ואז את החצי התחתון בציר ה־x. קו ההפרדה הוא הגבול שמפריד בין נקודות שיסווגו בכוכב או בעיגול.

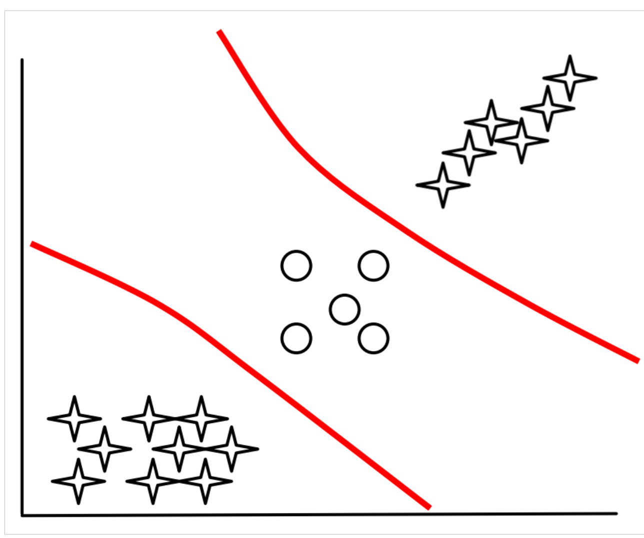


2. 1-Nearest Neighbor, כאשר הנקודה הנבדקת לא נכללת ברשימת הנקודות השכנות.

☐ לא יכול להגיע לשגיאה אפס. הסבר:

☐ יכול להגיע לשגיאה אפס. קו ההפרדה של המסווג:

יכול. במקרה זה, קו ההפרדה יורכב בדיוק מהנקודות שנמצאות במרחק זהה מנקודה כלשהי שמסווגת כעיגול, או מנקודה כלשהי שמסווגת ככוכב.



3. SVM עם Gaussian kernel.

☐ לא יכול להגיע לשגיאה אפס. הסבר:

☐ יכול להגיע לשגיאה אפס. קו ההפרדה של המסווג:

יכול. נזכור שבמקרה של SVM עם Gaussian kernel, המישור המפריד (במרחב שאליו נשלחות הנקודות) מקיים:

$$\mathbf{w} = \sum_n \alpha_n y_n \phi(\mathbf{x}_n)$$

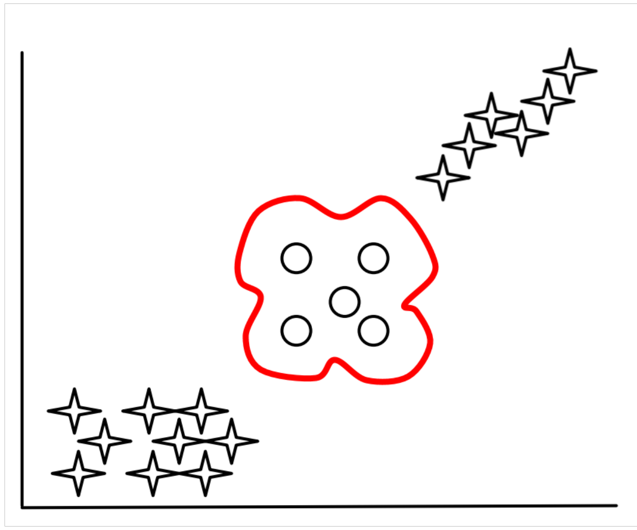
ולכן, קו ההפרדה הוא בדיוק כל הנקודות שהסיווג שלהן נותן 0:

$$\{\mathbf{x} | \langle \mathbf{w}, \phi(\mathbf{x}) \rangle + b = 0\}$$

או, באופן שקול

$$\begin{aligned}
\langle \mathbf{w}, \phi(\mathbf{x}) \rangle + b &= \left\langle \sum_n \alpha_n y_n \phi(\mathbf{x}_n), \phi(\mathbf{x}) \right\rangle + b \\
&= \sum_n \alpha_n y_n \langle \phi(\mathbf{x}_n), \phi(\mathbf{x}) \rangle + b \\
&= \sum_n \alpha_n y_n K(\mathbf{x}_n, \mathbf{x}) + b \\
&= \sum_n \alpha_n y_n e^{\|\mathbf{x}_n - \mathbf{x}\|^2 / 2\sigma^2} + b = 0
\end{aligned}$$

כלומר, זהו קו גובה של צירוף לינארי של גאוסיאנים, שמרכזם בנקודות המדגם.
בציור מתוארת דוגמה אחת לקו גובה כזה.



שאלה 2 - 20 נקודות

ראינו שהאלגוריתם של בעיית הרגרסיה הלינארית, מוצא פתרון אופטימלי ביחס לפונקציית ה־L2 loss. כעת נגדיר פונקציית loss חדשה, הנותנת לכל נקודה משקל שונה. כלומר, ה־loss על תחזית $\hat{y}_i$ של הנקודה ה־i יהיה:

$$L_i(\hat{y}_i, y_i) = w_i \cdot (\hat{y}_i - y_i)^2$$

1. מהי פונקציית האופטימיזציה המתאימה לבעיה החדשה?

נניח שנתונות לנו נקודות $\mathbf{x}_i \in \mathbb{R}^d$, וערכי $y_i \in \mathbb{R}$ המתאימים להם, נרצה להביא למינימום את סכום פונקציות ה־loss של כל הנקודות:

$$\arg\min_{\mathbf{a}\in\mathbb{R}^d} \sum_{i=1}^{n} w_i \cdot (\mathbf{a}^T\mathbf{x}_i - y_i)^2$$

2. האם המשקולות יכולות לשנות את הפתרון הנמצא, ביחס לבעיה המקורית ללא המשקולות? אם כן, תנ/י דוגמה. אם לא, הוכיח/י.

נוכיח שכן. נניח כי:

$$\mathbf{x}_1 = (1), y_1 = 0, \mathbf{x}_2 = (1), y_2 = 1$$

נפתור את פונקציית האופטימיזציה ביחס למשקולות כלליות:

$$f(a) = w_1 \cdot (a \cdot 1 - 0)^2 + w_2 \cdot (a \cdot 1 - 1)^2 = w_1 a^2 + w_2 (a-1)^2$$
$$f'(a) = 2w_1 a + 2(a-1)w_2 = a(2w_1 + 2w_2) - 2w_2 = 0 \Rightarrow$$
$$a_{opt} = w_2/(w_1 + w_2)$$

ניתן להראות שגם הנגזרת השנייה שלילית. מכאן שזו נקודת האופטימום לבעיה. עבור בעיית הרגרסיה המקורית, מתאימות המשקולות $w_1 = w_2 = 1$, ולכן הפתרון יהיה $a_{opt} = 1/2$. לעומת זאת, אם המשקולות היו $w_1 = 99, w_2 = 1$, אז $a_{opt} = 1/100$.

שאלה 3 - 20 נקודות

האלגוריתם שראינו בכיתה ל-Soft Margin SVM מניח שאנחנו רוצים להביא למינימום את סכום החריגות מהשוליים, כלומר את $\sum_{i=1}^{n} \xi_i$. בשאלה זו, נרצה להביא למינימום את ריבועי החריגות מהשוליים. כלומר, נתונה הבעיה הבאה:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{2}\sum_{n=1}^{N}\xi_n^2$$
$$\text{s.t.}\ \ y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n \qquad \forall n = 1, \dots, N$$

כאשר $\mathbf{x}_n \in \mathbb{R}^d$ ו-$y_n \in \{+1, -1\}$.

1. שימ/י לב שהשמטנו את האילוצים $\xi_n \geq 0$ בבעיה זו. הראה/י שאפשר להוריד את אילוצי אי-השליליות הללו - כלומר, שהערך האופטימלי לא משתנה אם האילוצים מתווספים או לא.

נראה שעבור הפתרון האופטימלי $\{\xi_n^{opt}\}_{n=1}^{N}$ בבעיה ללא האילוצים, מתקיים כי $\xi_n^{opt} \geq 0$ לכל n. נניח בשלילה כי קיים n עבורו $\xi_n^{opt} < 0$. נציע פתרון חדש, המוגדר כך:

$$\xi_n^{'} = max(0, \xi_n^{opt})$$

כלומר, כל משתנה שלילי יעוגל ל-0. בפרט, נקבל כי $\xi_n^{'} \geq \xi_n^{opt}$. נראה שפתרון זה עונה לאילוצים. הפתרון המקורי $\{\xi_n^{opt}\}_{n=1}^{N}$ עונה לאילוצים:

$$y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n^{opt}$$

ולכן, הפתרון החדש גם עונה לאילוצים, כיוון ש:

$$y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n^{opt} \geq 1 - \xi_n^{'}$$

מצד שני, הפתרון החדש משיג ערך מטרה נמוך יותר, מכיוון ש:

$$(\xi_n^{'})^2 \leq (\xi_n^{opt})^2$$

כדי לראות זאת, יש לחלק למקרים - אם הערך המקורי אי-שלילי, יש שוויון. אם הוא שלילי, אז זהו אי-שיוויון. מאחר שהנחנו שלפחות נקודה אחת שלילית, הרי שמצאנו פתרון חדש, שעומד באילוצים, ומקבל ערך נמוך יותר בפונקציית המטרה, בסתירה לאופטימליות של הפתרון המקורי.

2. מהו הלגרנז'יאן בבעיה זו?

לפי כללי KKT שלמדנו, הלגרנז'יאן הוא:

$$L(\boldsymbol{\alpha}, \mathbf{w}, b, \boldsymbol{\xi}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{2}\sum_{n=1}^{N}\xi_n^2 - \sum_{n=1}^{N}\alpha_n(y_n(\mathbf{w}^T\mathbf{x}_n + b) - 1 + \xi_n)$$

3. מהי הבעיה הדואלית במקרה זה?

עלינו לפתור את הבעיה

$$\max_{\boldsymbol{\alpha}} \min_{\mathbf{w},b,\boldsymbol{\xi}} L(\boldsymbol{\alpha}, \mathbf{w}, b, \boldsymbol{\xi})$$

בכפוף לאילוצים $\alpha_n \geq 0$. נגזור:

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n = 0 \Rightarrow \mathbf{w}(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n$$

$$\frac{\partial L}{\partial b} = -\sum_{n=1}^{N} \alpha_n y_n = 0 \Rightarrow \sum_{n=1}^{N} \alpha_n y_n = 0$$

$$\frac{\partial L}{\xi_n} = C\xi_n - \alpha_n \Rightarrow \xi_n(\boldsymbol{\alpha}) = \alpha_n / C$$

נציב בביטוי של הלגרנז'יאן ונקבל:

$$\begin{aligned}
L(\boldsymbol{\alpha}) &= \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{2}\sum_{n=1}^{N}\xi_n^2 - \sum_{n=1}^{N}\alpha_n(y_n(\mathbf{w}^T\mathbf{x}_n + b) - 1 + \xi_n) \\
&= \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{2}\sum_{n=1}^{N}\xi_n^2 - \mathbf{w}^T(\sum_{n=1}^{N}\alpha_n y_n \mathbf{x}_n) - b\sum_{n=1}^{N}\alpha_n y_n + \sum_{n=1}^{N}\alpha_n - \sum_{n=1}^{N}\alpha_n\xi_n \\
&= -\frac{1}{2}(\sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i)^T(\sum_{j=1}^{N}\alpha_j y_j \mathbf{x}_j) + \frac{C}{2}\sum_{n=1}^{N}\left(\frac{\alpha_n}{C}\right)^2 + \sum_{n=1}^{N}\alpha_n - \sum_{n=1}^{N}\alpha_n\left(\frac{\alpha_n}{C}\right) \\
&= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j - \frac{1}{2C}\sum_{n=1}^{N}\alpha_n^2 + \sum_{n=1}^{N}\alpha_n
\end{aligned}$$

המשך בדף הבא.

אם אנחנו מגדירים מטריצה M שמקיימת

$$M_{i,j} = y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

אז הבעיה הדואלית היא:

$$\max_{\boldsymbol{\alpha}} -\frac{1}{2}\boldsymbol{\alpha}^T M \boldsymbol{\alpha} - \frac{1}{2C}\boldsymbol{\alpha}^T \boldsymbol{\alpha} + \mathbf{1}^T \boldsymbol{\alpha}$$

בכפוף לאילוצים

$$\alpha_n \geq 0$$
$$\sum_{n=1}^{N} \alpha_n y_n = 0$$

תעודת זהות:
מספר מחברת:

שאלה 4 ־ 20 נקודות

1. נתון לנו מדגם מסווג $(\mathbf{x}_i, y_i)$, כאשר $\mathbf{x}_i \in \mathbb{R}^d$ ו־ $y_i \in \{+1, -1\}$.

מריצים את אלגוריתם Perceptron על מדגם אימון במשך T איטרציות, עד לקבלת התוצאה הסופית $\mathbf{w}^*$. הראה/י ש־$\mathbf{w}^*$ הוא צירוף לינארי של נקודות המדגם $\mathbf{x}_i$.

נוכיח באינדוקציה. בתחילת האלגוריתם, $\mathbf{w}_0 = \mathbf{0}$, ולכן זהו צירוף לינארי טריוויאלי. בצעד האינדוקציה, נניח כי $\mathbf{w}_t$ הוא צירוף לינארי. בכל איטרציה של האלגוריתם, או שהאלגוריתם צודק בסיווג שלו, ואז $\mathbf{w}_{t+1} = \mathbf{w}_t$, ולכן גם $\mathbf{w}_{t+1}$ צירוף לינארי; או שהאלגוריתם טועה, ואז:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + y_t \mathbf{x}_t$$

ולכן $\mathbf{w}_{t+1}$ נשאר צירוף לינארי.

2. השתמש/י בתוצאות הסעיפים הקודמים כדי לשנות את Perceptron כך שישתמש ב־kernels. האלגוריתם לא ישמור במפורש את הערך של $\mathbf{w}^*$, אלא ישמור ויעדכן בכל איטרציה את מקדמי הצירוף הלינארי שיוצר אותו. כתב/י פסאודו־קוד שמתאר את האלגוריתם החדש, והדגימ/י כיצד אפשר להשתמש בתוצאה שלו כדי לסווג נקודת מדגם חדשה $\mathbf{x}$, מבלי לשמור במפורש את $\mathbf{w}^*$.

הראינו בסעיף הקודם ש־$\mathbf{w}_t$ הוא צירוף לינארי של $\mathbf{x}_t$, עם מקדמים שנסמנם ב־α_n, כאשר $\alpha_n \in \{0, 1, -1\}$. נתון לנו קרנל K המקיים $K(\mathbf{x}_i, x_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ עבור פונקציית מיפוי כלשהי. אז הסיווג בכל איטרציה מתבצע על ידי הסימון של:

$$\langle \mathbf{w}_t, \mathbf{x}_t \rangle = \left\langle \sum_{i=1}^{t-1} \alpha_i \phi(\mathbf{x}_i), \phi(\mathbf{x}_t) \right\rangle = \sum_{i=1}^{t-1} \alpha_i \left\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_t) \right\rangle = \sum_{i=1}^{t-1} \alpha_i K(\mathbf{x}_i, \mathbf{x}_t)$$

ולכן האלגוריתם יהיה

input : $\mathbf{x}_1, \ldots, \mathbf{x}_T$
output: $\alpha_1, \ldots, \alpha_T$
for $t = 1, \ldots, T$ **do**
  if $y_t \neq sign\left(\sum_{i=1}^{t-1} \alpha_i K(\mathbf{x}_i, \mathbf{x}_t)\right)$ **then** $\alpha_t \leftarrow y_t$;
  else $\alpha_t \leftarrow 0$;
end

סיווג של נקודת מדגם חדשה יתבצע על ידי

$$h(\mathbf{x}) = sign\left(\sum_{i=1}^{T} \alpha_i K(\mathbf{x}_i, \mathbf{x})\right)$$

שאלה 5 - 20 נקודות

נתון לנו מדגם מסווג $(\mathbf{x}_i, y_i)$, כאשר $\mathbf{x}_i \in \mathbb{R}^d$ ו־ $y_i \in \{+1, -1\}$. נזכיר כי באיטרציה ה־t של אלגוריתם AdaBoost, מעדכנים את ההתפלגות D_t על ידי:

$$\begin{aligned}
h_t &= \arg\min_{h \in H} \Pr_{\mathbf{x} \sim D_t}[h(\mathbf{x}) \neq y] \\
\epsilon_t &= \Pr_{\mathbf{x} \sim D_t}[h_t(\mathbf{x}) \neq y] \\
\alpha_t &= \frac{1}{2} \ln \frac{1 - \epsilon_t}{\epsilon_t} \\
D_{t+1}(\mathbf{x}_i) &= \frac{D_t(\mathbf{x}_i) e^{-\alpha_t y_i h_t(\mathbf{x}_i)}}{Z_t}
\end{aligned}$$

כאשר H היא מחלקה של השערות.

1. הראה/י כי

$$\epsilon_t e^{\alpha_t} = \sqrt{\epsilon_t(1 - \epsilon_t)} = (1 - \epsilon_t) e^{-\alpha_t}$$

העלאה בחזקה תיתן לנו

$$e^{\alpha_t} = \sqrt{\frac{1 - \epsilon_t}{\epsilon_t}} \qquad e^{-\alpha_t} = \sqrt{\frac{\epsilon_t}{1 - \epsilon_t}}$$

ואז

$$\epsilon_t e^{\alpha_t} = \sqrt{\epsilon_t(1 - \epsilon_t)} = (1 - \epsilon_t) e^{-\alpha_t}$$

2. הראה/י כי גורם הנורמליזציה Z_t הוא

$$Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}$$

מההגדרה,

$$\sum_{\{i|y_i \neq h_t(\mathbf{x}_i)\}} D_t(\mathbf{x}_i) = \epsilon_t.$$

ולכן גורם הנורמליזציה מקיים

$$\begin{aligned} Z_t &= \sum_{i=1}^{m} D_t(\mathbf{x}_i) e^{-\alpha_t y_i h_t(\mathbf{x}_i)} \\ &= \sum_{\{i|y_i = h_t(\mathbf{x}_i)\}} D_t(\mathbf{x}_i) e^{-\alpha_t} + \sum_{\{i|y_i \neq h_t(\mathbf{x}_i)\}} D_t(\mathbf{x}_i) e^{\alpha_t} \\ &= (1-\epsilon_t) e^{-\alpha_t} + \epsilon_t e^{\alpha_t} \\ &= 2\sqrt{\epsilon_t(1-\epsilon_t)} \end{aligned}$$

כשהשתמשנו בתוצאת סעיף 1.

3. הרא/ה כי השגיאה של ההשערה הקודמת ביחס להתפלגות החדשה היא בדיוק חצי, כלומר:

$$\Pr_{\mathbf{x}\sim D_{t+1}}[h_t(\mathbf{x}) \neq y] = \frac{1}{2}$$

$$\begin{aligned}
\Pr_{\mathbf{x}\sim D_{t+1}}[h_t(\mathbf{x}) \neq y] &= \sum_{\{i|y_i\neq h_t(\mathbf{x}_i)\}} D_{t+1}(\mathbf{x}_i) \\
&= \sum_{\{i|y_i\neq h_t(\mathbf{x}_i)\}} \frac{D_t(\mathbf{x}_i)e^{-\alpha_t y_i h_t(\mathbf{x}_i)}}{Z_t} \\
&= \sum_{\{i|y_i\neq h_t(\mathbf{x}_i)\}} D_t(\mathbf{x}_i)\frac{e^{\alpha_t}}{Z_t} \\
&= \frac{e^{\alpha_t}}{Z_t}\epsilon_t = \frac{\sqrt{\frac{1-\epsilon_t}{\epsilon_t}}\epsilon_t}{2\sqrt{\epsilon_t(1-\epsilon_t)}} = \frac{1}{2}
\end{aligned}$$

4. האם בהכרח $h_{t+1} \neq h_t$?

כן. חלק מהנחות האלגוריתם הוא שמחלקת ההשערות היא של weak learners ולכן מתוך ההגדרה, אנחנו יודעים שהשגיאה של ההשערה h_{t+1} גדולה מחצי, שזו השגיאה של ההשערה הקודמת, לפי סעיף 3.

הערה: הנחה שלא נכתבה במפורש היא ש־$\epsilon_t > 0$, שכן אחרת אין טעם להריץ עוד איטרציה, וגם אז המשקולות וגורם הנורמליזציה לא מוגדרים. מכיוון שהיתה אי־ודאות, כל הסטודנטים קיבלו את מלוא הנקודות על סעיף זה.