FACULTY OF DATA
AND DECISION SCIENCES
MAKING SENSE OF IT

הטכניון
מכון טכנולוגי לישראל

למידה חישובית 1 (096411)

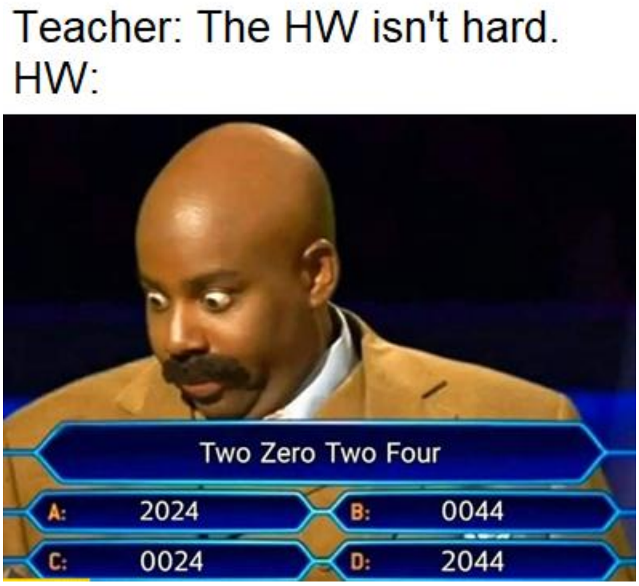
חורף 2025

<u>תרגיל בית 5</u>

<u>תאריך אחרון להגשה: 30.1.2025 בשעה 23:59</u>

<u>הוראות הגשה</u>

- ההגשה בזוגות בלבד, דרך "קבוצה" ייעודית שיצרתם במודל.
- עליכם להגיש קובץ **pdf** בודד:
  - `HW5_ID1_ID2.pdf` – קובץ המכיל תשובות לכל השאלות. עבור שאלה 2, יש להוסיף צילומי מסך של הקוד והפלטים שהוא מפיק. ניתן גם לייצא מחברת בפורמט PDF ולשרשר אותה לתשובות לחלק היבש (במקום המתאים).
- קוד חייב להיות קריא, תמציתי ומתועד היטב. יש להקפיד על שימוש בשמות משמעותיים למשתנים.
- כל גרף חייב להכיל לפחות את האלמנטים הבאים: כותרת, מקרא (legend), כותרות לצירים ויחידות (ticks).
- יש להשתמש בפורום במודל לטובת שאלות על התרגיל. השאלות שלכם עוזרות לסטודנטים אחרים בקורס.



שאלה 1

בהרצאה למדנו על האלגוריתם **AdaBoost** שמשלב קבוצת לומדים חלשים לכדי לומד חזק אחד.

כפי שאתם זוכרים, בהינתן משימת סיווג בינארית עם מדגם אימון $S = \{(x_i, y_i)\}_{i=1}^{m}$,

כאשר $y_i \in \{+1, -1\}$ **AdaBoost**, מבצע את הצעדים הבאים:

1. $D^{(1)} \leftarrow \left(\frac{1}{m}, \dots, \frac{1}{m}\right)$
2. **for t =1,…T:**
   i. $h_t \leftarrow WL(D^{(t)}, S)$
   ii. $\epsilon_t = \sum_{i=1}^{m} \quad D_i^{(t)} 1_{[y_i \neq h_t(x_i)]}$ **and** $w_t = \frac{1}{2} log\ log\ \left(\frac{1}{\epsilon_t} - 1\right)$
   iii. $D_i^{(t+1)} \propto D_i^{(t)} e^{-w_t y_i h_t(x_i)}$
3. **Output** $\hat{h}$ **where** $\hat{h}(x) = sign(\sum_{t=1}^{T} \quad w_t h_t(x))$

1. הסבירו את ההנחה על $WL(\cdot,\cdot)$ וכל צעד באלגוריתם.
2. הוכיחו שהשוויון הבא מתקיים:

$$\sum_{i=1}^{m} \quad D_i^{(t)} e^{-w_t y_i h_t(x_i)} = \varepsilon_t \cdot e_t^{w} + (1 - \varepsilon_t) \cdot e_t^{-w_t} = 2\sqrt{\varepsilon_t(1 - \varepsilon_t)}$$

בסעיף הבא תוכיחו כי אם $\forall t \in [T]\,,\ \varepsilon_t \leq \frac{1}{2} - \gamma$ אז:

$$L_s(\hat{h}) \leq e^{-2\gamma^2 T}$$

<u>מספר הגדרות:</u>

נגדיר את $f_0(x) \equiv 0$ ואת $f_t(x) \coloneqq \sum_{p=1}^{t} \quad w_p \cdot h_p(x)$ $\forall t \in N$,

מכאן ש- $\hat{h}(x) \coloneqq sign(f_T(x))$

בנוסף, $\forall t \in N \cup \{0\}$ נגדיר $z_t \coloneqq \frac{1}{m} \sum_{i=1}^{m} \quad e^{-y_i f_t(x_i)}$

3. ענו על הסעיפים הבאים:

   1. הוכיחו כי $\forall i \in [m]: 1_{[y_i \neq \hat{h}(x_i)]} \leq e^{-y_i f_T(x_i)}$ ולכן. $L_s(\hat{h}) \leq z_T$

   2. הוכיחו באינדוקציה כי $\forall t \in N\,, i \in [m]$ מתקיים:

   $$D_i^{(t)} \propto e^{-y_i f_{t-1}(x_i)}$$

   3. הוכיחו כי $\forall t \in N$

   $$\frac{z_t}{z_{t-1}} = 2\sqrt{\varepsilon_t(1 - \varepsilon_t)}$$

   4. הוכיחו כי אם $\forall t \in [T]\,,\ \varepsilon_t \leq \frac{1}{2} - \gamma$ אז:

   $$2\sqrt{\varepsilon_t(1 - \varepsilon_t)} \leq e^{-2\gamma^2}$$

<u>רמז:</u> הפונקציה $g(a) = a \cdot (1-a)$ היא מונוטונית עולה בתחום $\left[0, \frac{1}{2}\right]$. בנוסף, $\forall a: 1 - a \le e^{-a}$

הראו כי $\frac{z_t}{z_{t-1}}$. $z_T = \prod_{t=1}^{T}$ והוכיחו כי אם $\varepsilon_t \le \frac{1}{2} - \gamma$, $\forall t \in [T]$ אז:

$$L_s(\hat{h}) \le e^{-2\gamma^2 T}$$

4. הניחו כי $\varepsilon_t \le \frac{1}{2} - \gamma$, $\forall t \in [T]$. מהו מספר האיטרציות המינימלי T שאלגוריתם ***AdaBoost*** צריך לבצע בכדי להבטיח שהמסווג שיתקבל $\hat{h}$ ישיג בהכרח שגיאת אימון השווה ל- 0?
<u>רמז:</u> איזה ערכים $L_s(\hat{h})$ יכול לקבל?

הוכיחו שהשגיאה של h_t ביחס להתפלגות D^{t+1} היא בדיוק $\frac{1}{2}$. כלומר הוכיחו ש $\forall t \in [T]$ -מתקיים:

$$\sum_{i=1}^{m} D_i^{t+1} \cdot 1_{[y_i \ne h_t(x_i)]} = \frac{1}{2}$$

FACULTY OF DATA
AND DECISION SCIENCES
MAKING SENSE OF IT

הטכניון
מכון טכנולוגי
לישראל

שאלה 2

בשאלה הזאת נעבוד עם סט הנתונים wine-quality שמצורף לתרגיל במודל. סט נתונים זה מכיל1599 תצפיות עם 11 פיצ'רים. העמודה `quality` מתארת את התווית של כל יין (ציון הנע בין 3 ל-8).

בצעו את השלבים הבאים:

- טענו את קובץ הדאטה (`winequality-red.csv`) והמירו אותו ל-DataFrame של pandas.
- המירו את העמודה של `quality` לעמודה בינארית (כלומר הפכו את התווית להיות בינארית). כאשר כל תצפית עם ערך `quality > 5` תקבל תווית 0 ואחרת 1. המטרה תהיה לחזות את התווית הבינארי החדשה.
- חלקו את הנתונים למטריצת פיצ'רים (`X`) ווקטור לייבלים (`y`).)
- חלקו את סט הנתונים למדגם אימון ומדגם מבחן באמצעות הפקודות הבאות:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.40, random_state=42)
```

בסעיפים הבאים נשתמש במודלים הבאים מתוך הספרייה של sklearn:

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
```

1. הגדירו מודל של עץ החלטה `DecisionTreeClassifier`, עם הפרמטרים הבאים:
```
max_depth=12, random_state=0
```
הפעילו את המודל על מדגם האימון. דווחו אחוז דיוק על מדגם האימון ומדגם המבחן.

2. הגדירו מודל של `RandomForestClassifier`, עם הפרמטרים הבאים:
```
n_estimators=100, max_depth=12, random_state=0
```
הפעילו את המודל על מדגם האימון. דווחו אחוז דיוק על מדגם האימון ומדגם המבחן.
הציגו גרף המתאר את אחוז הדיוק על מדגם המבחן בלבד כפונקציה של מספר העצים (n_estimators) כאשר טווח הערכים של מספר העצים הוא מ 1 עד 100 עצים.

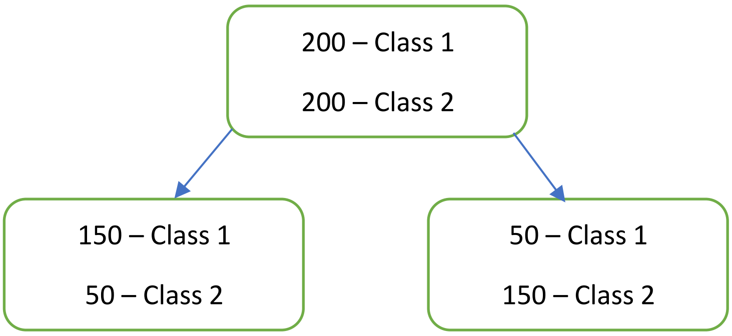
3. הגדירו שוב מודל של `RandomForestClassifier` עם אותם פרמטרים כמו בסעיף (ב) רק שעכשיו נרצה להוריד את מנגנון האקראיות בבחירת הפיצ'רים כך שבכל פיצול לא תהיה דגימה אקראית של תת קבוצת פיצ'רים אלא שבכל פיצול יסתכל על אוסף כל הפיצ'רים. (רמז: השתמשו בפרמטר `max_features`).
הפעילו את המודל על מדגם האימון. דווחו אחוז דיוק על מדגם האימון ומדגם המבחן.

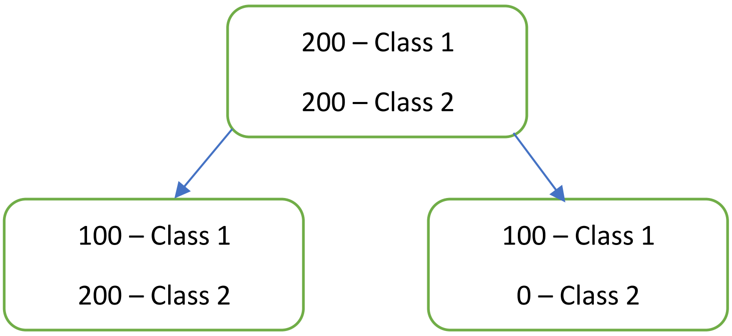
הציגו גרף המתאר את אחוז הדיוק על מדגם המבחן בלבד כפונקציה של מספר העצים (`n_estimators`) כאשר טווח הערכים של מספר העצים הוא מ 1 עד 100 עצים.

4. על סמך התוצאות והגרפים שהצגתם:
   a. האם הביצועים של המודל בסעיף (ב) הוא יותר טוב או יותר גרוע מהביצועים של המודל בסעיף (א). למה הוא יותר טוב/גרוע? הסבירו את תשובתכם.
   b. האם השינוי הנעשה בסעיף (ג) נתן ביצועים טובים יותר או גרועים יותר בהשוואה ל RandomForest הרגיל שהגדרתם בסעיף (ב). למה הוא נתן ביצועים טובים/גרועים יותר? הסבירו את תשובתכם.

סעיף ה הוא סעיף נפרד ובלתי תלוי בסעיפים הקודמים.

5. נניח שיש מדגם של בעיית סיווג בינארית עם 200 תצפיות לכל מחלקה. נסתכל על שני פיצולים אפשריים של התצפיות (בכל צומת כתוב מספר התצפיות שיש בכל מחלקה):





   a. איזה פיצול לדעתכם הוא פיצול טוב יותר?
   b. לכל אחד משני הפיצולים חשבו את מדדי ה- impurity הבאים:
   gini, entropy ו- misclassification rate.
   הסיקו איזה מדד מתוך השלושה לא כדאי להשתמש בשביל בניית עץ החלטה.

שאלה 3

ברשת חברתית ישנם 100 משתמשים אשר מידי יום מעלים תמונות של עצמם. לכל משתמש ישנו אינדקס ייחודי $j \in \{1, \dots, 100\}$. במאגר התמונות של הרשת החברתית נמצאות כעת m תמונות, כאשר ידוע כי בכל תמונה מופיע משתמש אחד בלבד. באופן פורמלי, את מאגר התמונות ניתן לתאר על ידי הקבוצה $\{(x_i, z_i)\}_{i=1}^{m}$ כאשר $x_i \in R^d$ היא התמונה ה-i ו-$z_i \in \{1, \dots, 100\}$ הוא האינדקס של המשתמש שמופיע בתמונה ה-i.

לרוע המזל, בשל באג במערכת, מאגר התמונות הושחת כך שהאינדקס של המשתמש בכל תמונה הוסר. כלומר, המאגר הנגיש היחיד הינו $\{x_1, \dots, x_m\}$. בשל כך, מנהלי הרשת החברתית החליטו להפעיל אלגוריתם clusteringבכדי לחלק את התמונות ל-100 קבוצות, כאשר כל קבוצה אמורה, באופן אידאלי, להכיל תמונות של משתמש אחד בלבד.

1. הסבירו כיצד ניתן לבצע את ה- clusteringבאמצעות אלגוריתם K-Means.
2. נניח שמומחי עיבוד תמונה הצליחו לייצר פונקציית מחרק בין תמונות $d(\cdot,\cdot)$ שמקיימת את התכונה הבאה:
$$\forall (x_1, z_1), (x_2, z_2), (x_3, z_3): if\ z_1 = z_2\ and\ z_2 \neq z_3\ then\ d(x_1, x_2) \leq d(x_1, x_3)$$
הסבירו מה המשמעות של תכונה זו. כיצד ניתן לעדכן את האלגוריתם שכתבתם בסעיף א' בעזרת הפונקציה d?

נניח כעת כי כל תמונה מיוצגת על ידי מספר חד ממדי, כלומר $x_i \in R$. בנוסף, נניח שהתמונות של כל משתמש $j \in \{1,2\}$ מתפלגות לפי $N(j, \sigma_j^2)$ ושהתמונות של כל משתמש $j \in \{3, \dots, 100\}$ מתפלגות לפי $Exp\left(\frac{1}{j}\right)$.

באופן פורמלי, מתקיים כי:

$$x_i | z_i = j \sim \{N(j, \sigma_j^2)\ j \in \{1,2\}\ Exp\left(\frac{1}{j}\right)\ j \in \{3, \dots, 100\}$$

כזכור, פונקציית הצפיפות של ההתפלגויות $N(j, \sigma_j^2)$ ו-$Exp(\lambda)$ נתונות על ידי:

$$f_{normal}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}, \quad f_{exp}(t) = \lambda e^{-\lambda t} \cdot 1_{\{t \geq 0\}}$$

3. הגדירו במפורש את הפרמטרים הלא ידועים, ורשמו את לוג הנראות של הצפיפות.
4. נסחו אלגוריתם EMלהערכת הפרמטרים הלא ידועים. רמת הפירוט צריכה להיות כזאת שמאפשרת מימוש של האלגוריתם. הפרט, הסבירו מהו שלב ה-E ומהו שלב ה-M. יש לכתוב נוסחאות מפורשות לעדכון כל אחד מהאמדים.
רמז: היעזרו בכך שאומד נראות מרבית לשונות של מ"מ נורמלי עם תוחלת ידועה μ הינו $\hat{\sigma}^2 = \frac{\Sigma_{i=1}^{m}(x_i-\mu)^2}{m}$.
5. נניח שידוע כי כל משתמש $j \in \{1, \dots, 100\}$ מעלה j תמונות ביום. כיצד האלגוריתם מהסעיף הקודם משתנה כעת?
רמז: ניתן להשתמש בנוסחה $\Sigma_{k=1}^{n} k = \frac{n(n+1)}{2}$.
6. נניח כעת של תמונה מתפלגת לפי $N(\mu_j, 1)$. הסבירו את הקשר בין אלגוריתם EM לאלגוריתם K-Means. האם מרכזי הקלאסטרים שיתקבלו מ-K-Means יהיו זהים ל-μ_j שיתקבלו מהרצת EM?