מבוא לעיבוד ספרתי של אותות ומידע 300107

שנה״ל: תשפ״ג סמסטר: ב׳

עבודת בית

ד״ר אריק פארן ד״ר בני סלומון

הוראות לנבחן:

1. העבודה תוגש ע"י סטודנט יחיד או בזוג.
2. סטודנט/זוג אינו רשאי לעיין בפתרון, מלא או חלקי, של עבודת בית של סטודנט/זוג אחר (ובכלל זה פתרון השמור במדיה דיגיטלית כלשהי, לרבות רשתות חברתיות והודעות דוא"ל) או להיעזר בפתרון כאמור בכל צורה שהיא.
3. אסור לסטודנטים/זוגות שונים לנסח במשותף מסמך.
4. אין להיעזר בחברים, מכרים, בני-משפחה או גורמים אחרים.
5. יש להגיש מסמך Word/PDF עם פתרונות לשאלות והסברים מפורטים של העבודה שלך לפי ההנחיות שבשאלות. **בנוסף**

- כלל השאלות בחלק I (עיבוד אותות): ההגשה תכלול גם סרטוני MP4 (כתוב קישורים להורדה במסמך), קבצי אודיו במידת הצורך וקבצי Matlab. תעד היטב את הקוד שלך. **ההסברים בקוד ובסרטונים ישפיעו על הניקוד!**
- כלל השאלות בחלק II (למידת מכונה): נא הקפידו לפרט הנחותיכם ולבסס את מסקנותיכם בתוצאות הניסויים שביצעתם, כולל גרפים/טבלאות וכיו"ב במקומות בהם זה יכול לסייע – **למרכיבים אלה יינתן משקל מרכזי בציון**. ההגשה תכלול גם סרטון MP4 (כתוב קישור להורדה במסמך), ואת **כל** הקוד הרלוונטי במחברת Jupyter יחידה כשהיא לאחר הרצה מלאה ומוכנה להרצה מחדש. **ההסברים בקוד ובסרטונים ישפיעו על הניקוד!**

6. למרצים יש אפשרות לזמן את הסטודנטים **להגנה (בחינה בעל פה)** לפני מתן ציון לעבודה.

בהצלחה !!

אנא אשר/י: הנני מתחייב/ת לעבודה עצמאית

חתימה_____

ת"ז לשם איש ור _____

חלק I – עיבוד ספרתי של אותות (50 נקודות)

שאלה 1 (10 נקודות)

חלק 1 (5 נקודות)

נתבונן במסנן בעל פונקציית תמסורת

$$H(z) = \frac{b_0}{1 + r^n z^{-n}}$$

כאשר $b_0 = 1 - r^n$. הנח $n = 10,\ r = 0.98$ וקצב דגימה $f_s = 300\ \mathrm{Hz}$.

צייר את תגובת ההלם של המסנן.

צייר את המגניטודה של תגובת התדר של המסנן כפונקציה של תדר הנמדד ב Hz. התייחס למגניטודה של תגובת התדר והסבר במדויק מה המסנן מבצע.

חלק 2 (5 נקודות)

נתבונן במסנן בעל פונקציית תמסורת

$$H(z) = \frac{b_0(1 - z^{-n})}{1 - r^n z^{-n}}$$

כאשר $b_0 = (1 + r^n)/2$. הנח $n = 10,\ r = 0.96$ וקצב דגימה $f_s = 200\ \mathrm{Hz}$.

צייר את תגובת ההלם של המסנן.

צייר את המגניטודה של תגובת התדר של המסנן כפונקציה של תדר הנמדד ב Hz. התייחס למגניטודה של תגובת התדר והסבר במדויק מה המסנן מבצע.

האיורים, תשובות לשאלות, והסבר של התוצאות צריכים להופיע במסמך.

צרף להגשה קובץ Matlab (script) הניתן להרצה.

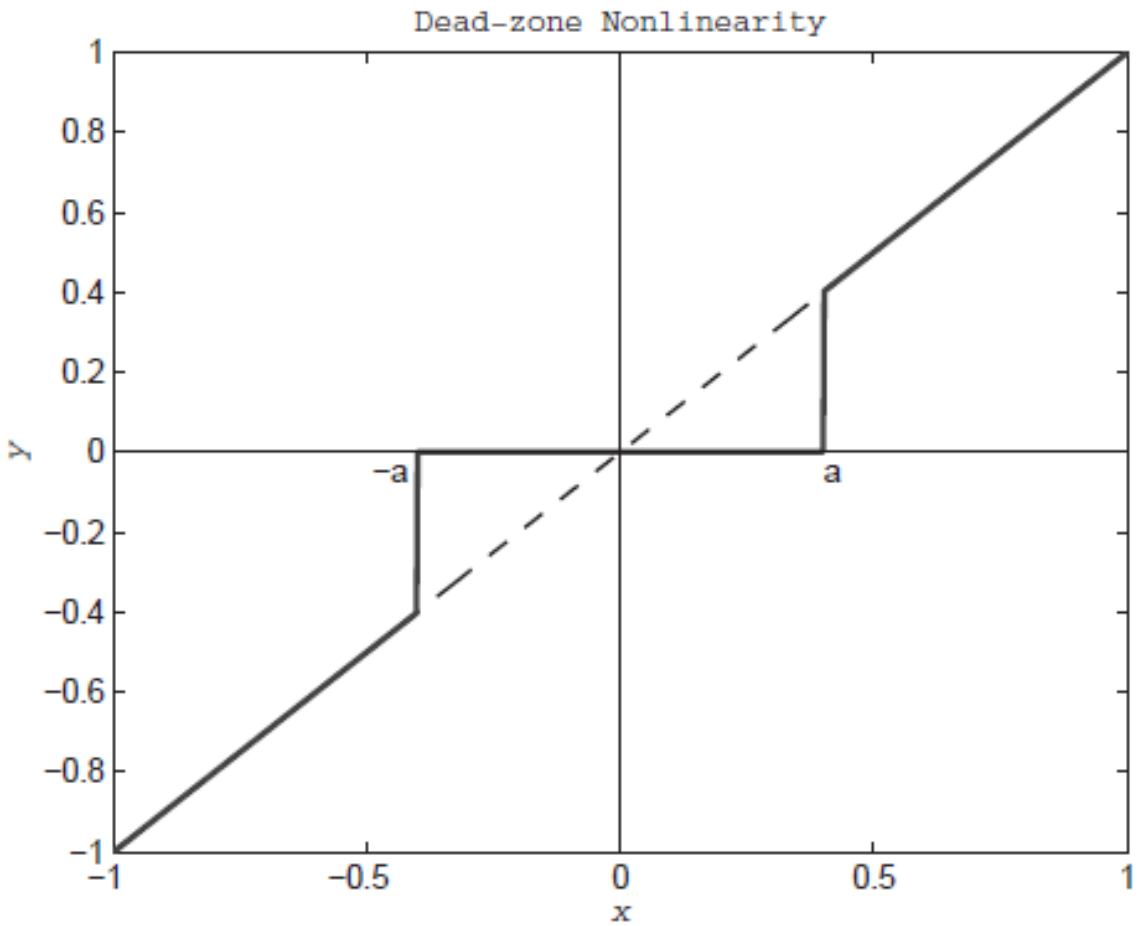
הכן סרטון (קובץ בפורמט MP4) שמסביר היטב את הפתרון שלך לשני החלקים של השאלה.

רשום קישור (link) להורדה של הסרטון. הסרטון צריך להיות זמין להורדה (ע"י הקישור) החל מזמן ההגשה.

שאלה 2 (10 נקודות)

חלק 1 (5 נקודות)

עיוות לא ליניארי נפוץ הוא dead-zone המוצג באיור הבא



כלומר עבור ערך כניסה מסוים x, ערך המוצא הוא

$$y = \begin{cases} 0, & 0 \le |x| \le a \\ x, & a < |x| < \infty \end{cases}$$

נתבונן באות הבא

$$x[n] = \cos(2\pi F_0 n T)\,, \quad 0 \le n \le N-1$$

כאשר $F_0 = 20\ \text{Hz}$, $N = 100$ והאות $x[n]$ התקבל ע"י דגימה בתדר דגימה $f_s = 2000\ \text{Hz}$ ולכן $T = 1/f_s$.

האות $x[n]$ עובר עיוות dead-zone כאשר $a = 0.25$. האי-ליניאריות יוצרת רכיבי תדר נוספים באות המוצא $y[n]$.

ניתן לומר (בקירוב טוב מאוד) שהאות $y[n]$ מתקבל מדגימה של האות המחזורי הבא

$$y(t) = \frac{d_0}{2} + \sum_{i=1}^{N/2-1} d_i \cos(2\pi i F_0 t + \theta_i)$$

ההספק של $y(t)$ הוא

$$P_y = \frac{d_0^2}{4} + \frac{1}{2}\sum_{i=1}^{N/2-1} d_i^2$$

העיוות ההרמוני הכולל (THD) הוא מדד לעוצמה היחסית של רכיבי התדר הנוספים שנוצרו עקב האי-ליניאריות. ה THD מוגדר בשאלה זאת ע"י

$$\mathrm{THD} = \frac{100\left(P_y - d_1^2/2\right)}{P_y} \ \%$$

א. צייר את הערך המוחלט של ה DFT של $y[n]$ כפונקציה של תדר f [Hz] עבור $0 \le f < f_s/2$.
ב. חשב את ה THD.

<u>חלק 2 (5 נקודות)</u>

חזור על חלק 1 אבל כעת $N = 50$, $f_s = 1000$ Hz והאי-ליניאריות היא

$$y = x^3$$

האיורים, תשובות לשאלות, והסבר של התוצאות צריכים להופיע במסמך.

צרף להגשה קובץ Matlab (script) הניתן להרצה.

הכן סרטון (קובץ בפורמט MP4) שמסביר היטב את הפתרון שלך לשני החלקים של השאלה.

רשום קישור (link) להורדה של הסרטון. הסרטון צריך להיות זמין להורדה (ע"י הקישור) החל מזמן ההגשה.

שאלה 3 (10 נקודות)

נתבונן במערכת RCSR המתוארת ע"י משוואת ההפרשים הבאה

$$y[n] = -r^2 y[n-2] + x[n]$$

א. הנח $r = 0.9$ ו $x[n] = \delta[n]$. מצא את $y[n]$ עבור $n = 0,1 \cdots ,127$ ומצא וצייר את הערך המוחלט של ה DFT של $y[n]$. כלומר צייר את $|Y[k]|\ ,\ k = 0,1,\cdots,127$.

ב. יהי

$$w[n] = 0.92^{-n} y[n]$$

מצא וצייר את הערך המוחלט של ה DFT של $w[n]$. כלומר צייר את $|W[k]|\ ,\ k = 0,1,\cdots,127$.

הסבר את התוצאות תוך התייחסות ל DTFT: היכן אמור להתקבל ערך המקסימום ומידת "החדות" של המקסימום.

ג. חזור על סעיף א עבור $r = 0.5$.

ד. חזור על סעיף ב עבור

$$w[n] = 0.55^{-n} y[n]$$

כאשר $y[n]$ הוא האות שמיוצר בסעיף ג.

הסבר את התוצאות תוך התייחסות ל DTFT: היכן אמור להתקבל ערך המקסימום ומידת "החדות" של המקסימום.

ה. כעת נוסיף רעש לאות $y[n]$ שהתקבל בסעיף ג ע"י

```
y = y + sqrt(0.1)*randn(1,128);
```

מצא וצייר את הערך המוחלט של ה DFT של $y[n]$ לאחר הוספת הרעש וחזור על סעיף ד (ביחס ל $y[n]$ אחרי הוספת הרעש).

הסבר את התוצאות שהתקבלו לעומת הסעיפים הקודמים.

האיורים, תשובות לשאלות, והסבר של התוצאות צריכים להופיע במסמך.

צרף להגשה קובץ Matlab (script) הניתן להרצה.

הכן סרטון (קובץ בפורמט MP4) שמסביר היטב את הפתרון שלך לשאלה.

רשום קישור (link) להורדה של הסרטון. הסרטון צריך להיות זמין להורדה (ע"י הקישור) החל מזמן ההגשה.

שאלה 4 (10 נקודות)

חלק 1 (5 נקודות)

נתבונן במסנן בעל פונקציית התמסורת הבאה

$$H(z) = \frac{0.2[(z+0.5)^2 + 1.5^2]}{z^2 - 0.64}$$

א. צייר את המגניטודה של תגובת התדר של המסנן וצייר מפת קטבים-אפסים של המסנן.
ב. מצא ע"י חישוב אנליטי (כלומר ידני לא באמצעות Matlab) ייצוג של המסנן בצורה הבאה

$$H(z) = H_{MP}(z)H_{AP}(z)$$

כאשר $H_{MP}(z)$ הוא מסנן פאזה מינימלית ו $H_{AP}(z)$ הוא מסנן all-pass. החישוב צריך להופיע במסמך.

ג. צייר את המגניטודה של תגובת התדר של $H_{MP}(z)$ ואת המגניטודה של תגובת התדר של $H_{AP}(z)$.
ד. צייר מפת קטבים-אפסים של $H_{MP}(z)$ ומפת קטבים-אפסים של $H_{AP}(z)$.

האיורים, תשובות לשאלות, והסבר של התוצאות צריכים להופיע במסמך.

צרף להגשה קובץ Matlab (script) הניתן להרצה.

הכן סרטון (קובץ בפורמט MP4) שמסביר היטב את הפתרון שלך לשאלה.

רשום קישור (link) להורדה של הסרטון. הסרטון צריך להיות זמין להורדה (ע"י הקישור) החל מזמן ההגשה.

חלק 2 (5 נקודות)

מבוא

הגדרנו את ה DFT ע"י

$$X[k]=\sum_{n=0}^{N-1}x[n]W_N^{kn}\ ,\ \ k=0,1,\cdots,N-1$$

מאחר ש $W_N^{-kN}=1$ מתקיים

$$X[k]=W_N^{-kN}\sum_{n=0}^{N-1}x[n]W_N^{kn}=\sum_{n=0}^{N-1}x[n]W_N^{-k(N-n)}$$

נניח ש $N=4$, מתקיים

$$X[k]=\sum_{n=0}^{3}x[n]W_4^{-k(4-n)}=x[3]W_4^{-k}+x[2]W_4^{-2k}+x[1]W_4^{-3k}+x[0]W_4^{-4k}$$
$$=W_4^{-k}\left\{x[3]+W_4^{-k}\left\{x[2]+W_4^{-k}\{x[1]+W_4^{-k}x[0]\}\right\}\right\}$$

כלומר אם נבצע את ההצבות

$$y_k[0]=x[0]+W_4^{-k}y_k[-1]$$

$$y_k[1]=x[1]+W_4^{-k}y_k[0],\quad y_k[2]=x[2]+W_4^{-k}y_k[1]$$

$$y_k[3]=x[3]+W_4^{-k}y_k[2],\quad y_k[4]=x[4]+W_4^{-k}y_k[3]$$

עם תנאי התחלה $y_k[-1]=0$ ומבוא $x[n]=0$ עבור $n<0$ ו $n\geq 4$, נקבל $X[k]=y_k[4]$.

כלומר ניתן לחשב את $X[k]$ בצורה רקורסיבית

(1) $y_k[n]=W_N^{-k}y_k[n-1]+x[n],\ \ 0\leq n\leq N$

(2) $X[k]=y_k[N]$

עם תנאי התחלה $y_k[-1]=0$ ומבוא $x[n]=0$ עבור $n<0$ ו $n\geq N$.

נשים לב שמשוואה 1 היא למעשה פעולת סינון.

הגישה הזאת משתלמת מבחינה חישובית במקרים שבהם מעוניינים לחשב רק מספר קטן של ערכי DFT ולא את כל ערכי ה DFT. כלומר לחשב ערכי DFT רק עבור מספר קטן של ערכי k.

<u>מטלה לביצוע</u>

א. כתוב פונקציית Matlab שמקבלת כמבואות אות ווקטור ערכי k שבהם מעוניינים לחשב את ה DFT של האות, ומחשבת את ה DFT בערכי k הרצויים ע"י משוואות (2) , (1). הפונקציה שלך חייבת להשתמש בפונקציית filter של Matlab עבור המימוש של משוואה (1).

ב. אנחנו מעוניינים לגלות את התדר f_d של אות סינוסואידלי $\cos(2\pi f_d t)$ כאשר f_d יכול להיות אחד מהתדרים $490, 1280, 2730, 3120$ Hz.

האות נדגם בקצב $f_s = 8$ kHz ומשתמשים ב 100 הדגימות הראשונות

$$x[n] = \cos(2\pi f_d nT) \ , \ n = 0,1,\cdots,99$$

כאשר T הוא מרווח הדגימה. לצורך גילוי התדר,

- יש להשתמש בפונקציה מסעיף א לחישוב ה DFT ב 4 ערכי k מסוימים (בחר את הערכים והסבר את בחירתך במסמך)
- יש להחליט איזה תדר f_d יש לאות בהסתמך על ארבעת ערכי DFT המחושבים.

הדגם שהקוד שלך עובד עבור כל אחד מהערכים האפשריים של f_d והתייחס לכך במסמך.

התשובות לשאלות והסבר של התוצאות צריכים להופיע במסמך.

צרף להגשה קובץ Matlab (script) הניתן להרצה.

הכן סרטון (קובץ בפורמט MP4) שמסביר היטב את הפתרון שלך לשאלה.

רשום קישור (link) להורדה של הסרטון. הסרטון צריך להיות זמין להורדה (ע"י הקישור) החל מזמן ההגשה.

שאלה 5 (10 נקודות)

רקע תיאורטי

במקרים מסוימים יש לעדכן את המקדמים של מסנן FIR בהתאם לאות הכניסה. במקרים אלה, סינון במסנן FIR עם סדר m נראה כך

$$y(k) = \sum_{i=0}^{m} w_i(k)x(k-i)$$

כאשר $w_0(k), w_1(k), \cdots, w_m(k)$ הם מקדמי המסנן בזמן k.

כיצד נעדכן את מקדמי המסנן?

נתון אות כניסה $x(k),\ k = 0,1,\cdots,N-1$ ונתון אות רצוי $d(k)\ , k = 0,1,\cdots,N-1$.

נגדיר את הווקטורים

$$\underline{\mathbf{u}}(k) = [x(k) \quad x(k-1) \quad \cdots \quad x(k-m)]^T$$

$$\underline{\mathbf{w}}(k) = [w_0(k) \quad w_1(k) \quad \cdots \quad w_m(k)]^T$$

ונבצע את האלגוריתם הבא:

1. תנאי התחלה

$$\underline{\mathbf{w}}(0) = [w_0(0) \quad w_1(0) \quad \cdots \quad w_m(0)]^T = [0 \quad 0 \quad \cdots \quad 0]^T$$

2. בצורה איטרטיבית עבור $k = 0,1,\cdots,N-1$

א. חשב את (כאשר יש להניח שערכי $x(k)$ עבור $k < 0$ הם אפסים)

$$y(k) = \sum_{i=0}^{m} w_i(k)x(k-i) = \underline{\mathbf{w}}^T(k)\underline{\mathbf{u}}(k)$$

$$e(k) = d(k) - y(k)$$

ב. עדכן את תגובת ההלם של מסנן ה FIR

$$\underline{\mathbf{w}}(k+1) = \underline{\mathbf{w}}(k) + 2\mu e(k)\underline{\mathbf{u}}(k)$$

כאשר μ הוא פרמטר לבחירתנו.

שאלת Matlab

נתונה מסנן IIR בעל פונקציית תמסורת

$$H(z)=\frac{2-3z^{-1}-z^{-2}+4z^{-4}+5z^{-5}-8z^{-6}}{1-1.6z^{-1}+1.75z^{-2}-1.436z^{-3}+0.6814z^{-4}-0.1134z^{-5}-0.0648z^{-6}}$$

אנחנו מעוניינים למצוא מסנן FIR כך שהמגניטודה של תגובת התדר שלו תהיה דומה למגניטודה של תגובת התדר של המסנן הנתון. לצורך כך נפעיל את האלגוריתם שתואר ברקע התיאורטי.

הנח שלמסנן ה FIR יש סדר $m=50$, הפרמטר הוא $\mu=0.01$, אות הכניסה נתון בקובץ x.mat והאות הרצוי הוא

```
d = filter(b,a,x);
```

כאשר b ו a הם המקדמים של מסנן ה IIR הנתון.

צייר את $e^2(k),\ k=0,1,\cdots,N-1$ כפונקציה של k.

צייר את המגניטודה של תגובת התדר של מסנן ה IIR הנתון ואת המגניטודה של תגובת התדר של מסנן ה FIR שמתקבל בסיום האלגוריתם באותו איור.

בדוק את ההתנהגות של האלגוריתם בתלות בערכים שונים של הפרמטר μ (צרף איורים וכתוב את מסקנותיך במסמך).

האיורים, תשובות לשאלות, והסבר של התוצאות צריכים להופיע במסמך.

צרף להגשה קובץ Matlab (script) הניתן להרצה.

הכן סרטון (קובץ בפורמט MP4) שמסביר היטב את הפתרון שלך לשאלה.

רשום קישור (link) להורדה של הסרטון. הסרטון צריך להיות זמין להורדה (ע"י הקישור) החל מזמן ההגשה.

חלק II: אימון מסווגים בינאריים

הקדמה

בחלק זה של הפרויקט תאמנו מסווגים בינאריים מסוג רגרסיה לוגיסטית ו-KNN, כולל מדידת ביצועיהם על המידע הנתון והערכת ביצועיהם הצפויים על מידע שאינו ידוע בזמן האימון. כפי שלמדנו, מסווג בינארי הינו פונקציה $f: \mathcal{R}^{\mathrm{D}} \longmapsto \{0,1\}$ המשייך לכל וקטור במרחב המאפיינים $\mathrm{x} \in \mathcal{R}^{\mathrm{D}}$ (כאשר D הוא מימד מרחב המאפיינים) אחד משני ערכים, למשל $y \in \{0,1\}$. המידע $\mathcal{D} = \{(\mathrm{x}_n, y_n)\}_{n=1}^{N}$ שישמש אתכם לאימון המסווגים מורכב מאוסף דוגמאות מתויגות, כשוקטור המאפיינים בכל דוגמא מכיל את מאפייני הנוסעים במסע הבכורה של חללית ליישוב כוכבי לכת במערכות שמש שכנות, והתיוג של כל דוגמא מייצג האם הנוסע כן/לא עבר בטעות טלפורטציה מסוכנת למימד אחר!

ראינו כי מסווג מסוג KNN הינו בעל אופי הסתברותי ומבוסס על המודל

$$P(y = c|\mathrm{x}, \mathrm{K}) = \frac{\sum_{n \in B_K(x)} \mathbb{I}(y_n == c)}{\mathrm{K}} \qquad \forall \qquad c = 0,1$$

כש- $B_K(x)$ בנוסחא למעלה מייצג את אוסף כל האינדקסים של K הדוגמאות הקרובות ביותר לדוגמא x (כפי שהגדרנו בהרצאה), ופונקציית האינדיקטור $\mathbb{I}(A)$ עבור טענה כלשהי A הינה

$$\mathbb{I}(A) = \begin{cases} 1 & \text{if } A \text{ is TRUE} \\ 0 & \text{otherwise} \end{cases}$$

עבור מסווג מסוג רגרסיה לוגיסטית ראינו כי הוא מבוסס על המודל

$$P(y = 1|\mathrm{x}, \mathrm{w}) = \sigma\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}\right) = \frac{1}{1 + e^{-\mathrm{w}^{\mathrm{T}}\mathrm{x}}}$$

וכי אימון המסווג נעשה על-ידי חישוב הפרמטרים הממקסמים את פונקציית הסבירות של $\mathcal{D}$

$$\begin{aligned} \mathrm{w}_{\mathrm{ML}} &= \underset{\mathrm{w} \in \mathcal{R}^{\mathrm{D}}}{\operatorname{argmax}}\, P\left(\{y_n\}_{n=1}^{N} \middle| \{x_n\}_{n=1}^{N}, w\right) \\ &= \underset{\mathrm{w} \in \mathcal{R}^{\mathrm{D}}}{\operatorname{argmax}} \prod_{n=1}^{N} \sigma\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_{\mathrm{n}}\right)^{y_n} \cdot \left(1 - \sigma\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_{\mathrm{n}}\right)\right)^{1-y_n} \end{aligned}$$

שניתנת לניסוח שקול כבעיית האופטימיזציה

$$\mathrm{w}_{\mathrm{ML}} = \underset{\mathrm{w} \in \mathcal{R}^{\mathrm{D}}}{\operatorname{argmin}}\, \mathcal{L}_{\mathrm{CE}}(\mathcal{D}, \mathrm{w})$$

כאשר $\mathcal{L}_{\mathrm{CE}}(\mathcal{D}, \mathrm{w})$, פונקציית ה cross-entropy, נתונה על-ידי

$$\mathcal{L}_{\mathrm{CE}}(\mathcal{D}, \mathrm{w}) = -\sum_{n=1}^{N} \left(y_n \log \sigma\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_n\right) + (1 - y_n) \log\left(1 - \sigma\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_n\right)\right)\right)$$

לבסוף, כלל ההחלטה של המסווג תלוי בערך סף $\rho \in (0,1)$ הנתון לבחירתנו והינו

$$f(\mathrm{x}|\rho,\theta) = \begin{cases} 1 & \text{if} \quad P(y|\mathrm{x},\theta) > \rho \\ 0 & \text{otherwise} \end{cases}$$

כאשר $\theta = \mathrm{K}$ עבור מסווג מסוג KNN ו $\theta = \mathrm{w}$ עבור מסווג מסוג רגרסיה לוגיסטית.

דיוק (accuracy) המסווג $f(\mathrm{x}|\rho,\theta)$, עבור הדאטה $\mathcal{D}$, נתון על ידי

$$P_c(f,\theta,\mathcal{D}) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}\big(y_n == f(\mathrm{x}_n|\rho,\theta)\big)$$

<u>הערה:</u> בפתרונכם ניתן להשתמש בקוד שכתבתם במהלך הסימסטר בתרגילי הקידוד במעבדות ובמטלות הבית.

פרק א: הכנת המידע לאימון מסווג בינארי

כפי שדנו בהרצאה, המידע שישמש אותנו במסגרת הפרויקט נלקח מקישור <u>זה</u>, כולל תיאור כל אחת מהעמודות בו המצורף למטה (התיאור אף הוא נלקח מהקישור הנ"ל). מטרתכם היא לממש אלגוריתם למידה מסוג gradient descent למסווג מסוג רגרסיה לוגיסטית במטרה לזהות את הנוסעים שעברו טלפורטציה למימד אחר.

<u>שימו לב:</u> לנוחותכם צורפו הדאטה ותיאורו, ואין חובה לעשות שימוש בקישור הנ"ל.

- `PassengerId` - Each Id takes the form `gggg_pp` where `gggg` indicates a group the passenger is travelling with and `pp` is their number within the group. People in a group are often family members, but not always.
- `HomePlanet` - The planet the passenger departed from, typically their planet of permanent residence.
- `CryoSleep` - Indicates whether the passenger elected to be put into suspended animation for the duration of the voyage. Passengers in cryosleep are confined to their cabins.
- `Cabin` - The cabin number where the passenger is staying. Takes the form `deck/num/side`, where `side` can be either `P` for *Port* or `S` for *Starboard*.
- `Destination` - The planet the passenger will be debarking to.
- `Age` - The age of the passenger.
- `VIP` - Whether the passenger has paid for special VIP service during the voyage.
- `RoomService`, `FoodCourt`, `ShoppingMall`, `Spa`, `VRDeck` - Amount the passenger has billed at each of the *Spaceship Titanic*'s many luxury amenities.
- `Name` - The first and last names of the passenger.
- `Transported` - Whether the passenger was transported to another dimension. This is the target, the column you are trying to predict.

6. (1 נקודות) טענו את הדאטה שקיבלתם וחלקו את המידע לשתי סדרות ללא דוגמאות משותפות, סדרת אימון $\mathcal{D}_{\text{train}} = \{\mathrm{x}_{\text{train},n}, y_{\text{train},n}\}_{n=1}^{N_{\text{train}}}$ וסדרת מבחן $\mathcal{D}_{\text{test}} = \{\mathrm{x}_{\text{test},n}, y_{\text{test},n}\}_{n=1}^{N_{\text{test}}}$ כאשר $N_{\text{train}} + N_{test} =$ 8693. הסבירו בפירוט את שיקוליכם בחלוקת הדאטה לשתי הסדרות כולל התייחסות לבחירת הערכים עבור $N_{\text{train}}, N_{test}$.

7. (6 נקודות) תארו את המאפיינים שבחרתם להשתמש בהם לאימון המסווג, תוך התייחסות מפורטת לנקודות הבאות

א. אם ויתרתם על מאפיין (עמודה במידע) שניתן לכם, נמקו את בחירתכם.

ב. אם הגדרתם מאפיינים חדשים (feature engineering) על-סמך המידע שקיבלתם, הסבירו בפירוט כיצד חישבתם אותם ונמקו בקצרה מדוע לדעתכם הם עשויים לשפר את ביצועי המסווג.

פרק ב: אימון מסווג בינארי מסוג KNN

8. ממשו את אלגוריתם KNN בהתאם למה שלמדנו בקורס. שימו לב שהאלגוריתם מקבל כקלט

- אוסף דוגמאות מסווגות
- מספר השכנים K
- אוסף דוגמאות לא מסווגות

ומחזיר כפלט את הסיווגים של הדוגמאות הלא מסווגות.

א. (1 נקודות) השתמשו בקוד שכתבתם ובדאטה שהכנתם בחלק א בכדי לאמן מסווג מסוג KNN עם K=3, וחשבו דיוקי המסווג שאימנתם עבור $\mathcal{D}_{\text{train}}$ ו- $\mathcal{D}_{\text{test}}$.

ב. (5 נקודות) כעת השתמשו בקוד שכתבתם בכדי לבחון את ביצועי המסווג עבור ערכי K שונים, החל מ K=1, ובחרו את ערכו של K איתו צפוי המסווג להיות בעל הדיוק הגבוה ביותר בסיווג דוגמאות עתידיות. תארו בפירוט את הניסויים שביצעתם ואת התוצאות שהתקבלו, ונמקו את בחירתכם ב K המבטיח ביותר.

פרק ג: אימון מסווג בינארי מסוג רגרסיה לוגיסטית

9. (15 נקודות) כפי שלמדנו, אלגוריתם gradient descent עבור רגרסיה לוגיסטית לומד את מקדמי המסווג על ידי עדכון איטרטיבי של ערכי וקטור המקדמים ע"פ הנוסחה

$$\mathrm{w}^{(t+1)} = \mathrm{w}^{(t)} - \alpha \sum_{n=1}^{N} \mathrm{x}_n \left(\sigma\left(\left(\mathrm{w}^{(t)}\right)^{\mathrm{T}} \mathrm{x}_n \right) - y_n \right)$$

כאשר t הינו אינדקס האיטרציה של נוסחת העדכון והפרמטר α מגדיר את קצב עדכון הפרמטרים. ממשו בפייתון את אלגוריתם gradient descent עבור מסווג מסוג רגרסיה לוגיסטית ע"פ נוסחת העדכון הנ"ל.

שימו לב לממדים של כל איבר בנוסחה – בפרט, מה הממד במימושכם של x_n, $\mathrm{w}^{(t)}$, $\sigma\left(\left(\mathrm{w}^{(t)}\right)^{\mathrm{T}} \mathrm{x}_n\right)$ ו- y_n?

מומלץ לבחון במהלך המימוש את תקינותו על מידע דו-מימדי פשוט (למשל כמו המידע הדו-מימדי שקיבלתם במעבדות בקורס), ולוודא שאימון המסווג מתקדם כפי שאתם מצפים.

השתמשו בדאטה שהכנתם בחלק א ובקוד שכתבתם בכדי לאמן מסווג מסוג רגרסיה לוגיסטית. תארו את השלבים שביצעתם במהלך האימון, כולל ערכם של הגדלים שמדדתם לאורכו, תוך שימוש בגרפים ובתוצאות מספריות רלוונטיות. **<u>לכל הפחות</u>**, התייחסו בתשובתכם לגדלים והגרפים הבאים:

א. תארו גרפית את השתנות ערך פונקציית הקנס $\mathcal{L}_{\text{CE}}\left(\mathcal{D}_{\text{train}}, \mathrm{w}^{(t)}\right)$ ודיוק הסיווג $P_c\left(f, \mathrm{w}^{(t)}, \mathcal{D}_{\text{train}}\right)$ עבור ערכי הפרמטרים המתקבלים לאורך האיטרציות של עדכוני ה gradient descent (כלומר כפונקציה של t)

ב. חשבו את דיוק המבחן של המסווג $P_c\left(f, \mathrm{w}^{(\mathrm{t})}, \mathcal{D}_{\mathrm{test}}\right)$, עם המשקלים שתהליך האימון התכנס אליהם $\mathrm{w}^{(\mathrm{T})}$, כאשר T מייצג את מספר האיטרציות הכולל שבוצעו עד הפסקת אלגוריתם ה- gradient descent

ג. חשבו והציגו את מטריצת הערבול ואת גרף ה-ROC של המסווג עבור $\mathcal{D}_{\mathrm{train}}$ לפני תחילת האימון (כלומר עם ערכי $\mathrm{w}^{(0)}$ שהגרלתם), לאחר $\mathrm{T}/2$ איטרציות, ובסיום האימון.

ד. חשבו והציגו את מטריצת הערבול ואת גרף ה-ROC של המסווג עבור $\mathcal{D}_{\mathrm{test}}$ לאחר סיום האימון.

10. (10 נקודות) במסגרת האימון, מאופי הגדרתו של אלגוריתם ה gradient descent, נדרשתם לבחור ערכים עבור נקודת ההתחלה $\mathrm{w}^{(0)}$, גודל הצעד α, ומספר העדכונים T. בחנו את ההשפעה של ערכים שונים עבור פרמטרים אלו על תוצאות האלגוריתם וביצועי המסווג שהתקבל, ו**נמקו בפירוט תוך שימוש בתוצאות מספריות ובגרפים רלוונטיים** את בחירתכם בערכים בהם השתמשתם. **לכל הפחות**, התייחסו בתשובתכם לסוגיות הבאות:

א. כיצד קבעתם את T כך שלא יהיה קטן מידי (מה שעשוי למנוע מהאלגוריתם למצוא את מסווג רגרסיה לוגיסטית המיטבי) או גדול מידי (מה שעשוי לצרוך משאבי מחשוב וזמן ריצה יקרים ללא שילוו בהכרח בשיפור ביצועי המסווג)?

ב. האם היתה ל- $\mathrm{w}^{(0)}$ ו-α השפעה על מספר הצעדים הנדרש T?

ג. האם היתה ל- $\mathrm{w}^{(0)}$ ו-α השפעה על $\mathrm{w}^{(\mathrm{T})}$, נקודת הסיום של האלגוריתם?

ד. האם הסתפקם בחלוקה בודדת לסדרת אימון $\mathcal{D}_{\mathrm{train}}$ וסדרת מבחן $\mathcal{D}_{\mathrm{test}}$? במידה ובחנתם חלוקות שונות, איך השתמשתם בתוצאות ההרצות של אלגוריתם הלמידה?

ה. האם לבחירותיכם בחלוקת הדאטה לסדרות $\mathcal{D}_{\mathrm{train}}$ ו $\mathcal{D}_{\mathrm{test}}$ היתה השפעה על תוצאות ניסוייכם?

ו. האם שגיאת האימון היוותה מדד אמין לשגיאת המבחן?

ז. **האם יש היבטים נוספים שבחנתם המהלך האימון? אם כן, נא פרטו. מומלץ להעזר בנימוקים כמותיים ובהמחשות ויזואליות (כגון גרפים, gif–ים, סרטונים קצרים וכו') ככל שיש כאלה הרלוונטיים לתשובתכם.**

פרק ד: השפעת נרמול המאפיינים על ביצועי האלגוריתמים

11. (12 נקודות) שיטה מקובלת לעיבוד מוקדם של מידע לפני אימון מסווג כוללת, עבור כל מאפיין בנפרד,

- החסרת הממוצע, כך שהממוצע החדש של כל מאפיין הינו 0
- נרמול השונות, כך שהשונות החדשה של כל מאפיין הינה 1

השתמשו בשיטה זאת לנירמול המאפיינים בדאטה, וחיזרו על אימון המסווגים שביצעתם בפרקים א+ב למעלה. השוו **בפירוט**, תוך **התייחסות מפורשת** לתוצאות המספריות שקיבלתם, בין המסווגים שאימנתם עם ובלי נרמול המאפיינים

א. האם היה הבדל בתהליך האימון של המסווגים? אם נדרשו התאמות, ציינו בפירוט מה הן היו.

ב. האם היה הבדל בדיוקים שהתקבלו? במידה וכן, דונו בהרחבה בסיבות שהביאו לדעתכם להבדל.

בהצלחה!