תרגיל בית 4

מבוא לעיבוד ספרתי של אותות ומידע

הוראות הגשה

1. במידה וזה יכול לסייע לכם בתשובותיכם, ניתן (ואף מומלץ) לעשות שימוש חוזר בקוד הנמצא במחברות ה- jupyter שנמסרו לכם בכיתה, או שכתבתם במסגרת המעבדות או במטלות בית קודמות.
2. כפתרון לעבודת בית זאת יש להגיש 5 קבצים נפרדים (לא מאוגדים כקובץ zip):
   א. קובץ עם התשובות לשאלות העיוניות בתרגיל זה (סרוק ברזולוציה גבוהה במידה והתשובות הן בכתב יד).
   ב. מחברות Jupyter נפרדות עם התשובות לכל אחת משאלות 3 ו-5. לאחר הרצתן במלואן הגישו כל מחברת בשתי גרסאות:
      i. כקובץ ipynb כשהוא מוכן להרצה מחדש (כולל קבצי דאטה, במידה ויש כאלה)
      ii. כקובץ pdf (על קובץ זה להיות זהה בכל פרטיו לקבץ בסעיף i, פרט להיותו בפורמט שונה)

בהצלחה!

1. הטלת קובייה

מטילים קובייה הוגנת פעמיים, ומגדירים את המשתנה האקראי X להיות המספר הגבוה מבין השניים שהתקבלו.

א. מהם הערכים האפשריים עבור X?
ב. מה ההסתברות לקבלת כל אחד מהערכים הנ"ל?
ג. מהם הממוצע, השונות וסטיית התקן של X?

2. התפלגות משותפת, שולית ומותנית של משתנים אקראיים בדידים

נתונה P(X,Y) ההתפלגות המשותפת הבאה של שני משתנים אקראיים בדידים

P(X,Y)

Y \ X	0	1	2	3	4
0	0.01	0.02	0.03	0.1	0.1
1	0.05	0.1	0.05	0.07	0.2
2	0.1	0.05	0.03	0.05	0.04

חשבו את

א. $P(X|Y=1)$
ב. $P(Y|X=3)$
ג. $P(Y|X>2)$
ד. $P(X)$
ה. $P(Y)$
ו. האם המשתנים האקראיים X ו Y הם בלתי תלויים סטטיסטית?

3. בעיית מונטי הול

בשאלה זאת תממשו את המשחק מונטי הול שדנו בו בכיתה (תיאור שלו מצורף למטה, ובנוסף ניתן לקרוא עליו כאן), תסמלצו אותו מספר רב של פעמים עבור טקטיקות משחק שונות, ותפרשו את התוצאות הסטטיסטיות של הסימולציה. כפתרון עליכם להגיש מחברת jupyter הכוללת את הקוד שכתבתם ואת הגרפים המתוארים למטה, והסבר קצר שלכם לתוצאות.

תיאור המשחק:
כפי שדנו בשיעור ,במשחק הקרוי 'מונטי הול' מציגים בפני שחקן שלוש דלתות, שמאחורי אחת מהן מצויה מכונית ומאחורי כל אחת משתי הדלתות האחרות ישנה עז. מטרת השחקן היא לבחור את הדלת שמאחוריה נמצאת המכונית - אם יצליח יזכה בה, ואם לא אז לא יקבל דבר. ע"פ כללי המשחק ראשית בוחר השחקן דלת באקראי. לאחר מכן המנחה, היודע מה יש מאחורי כל דלת, פותח אחת משתי הדלתות האחרות ומגלה מאחוריה עז. כעת ניתנת לשחקן האפשרות לדבוק בבחירתו המקורית, או להחליפה בדלת האחרת שעודנה סגורה.
נבחן את תוצאות המשחק עבור שלושה סוגי שחקנים:

- שחקן א תמיד דבק בבחירתו המקורית.
- שחקן ב תמיד מחליף את בחירתו לדלת שנותרה סגורה.
- שחקן ג דבק בבחירתו בסיכוי 0.5, ומחליפה בסיכוי 0.5.

א. כיתבו קוד המדמה את משחק מונטי הול ע"פ השלבים הבאים (מומלץ לממש כל שלב בפונקציה נפרדת):
  - הגרילו מאחורי איזו דלת תימצא המכונית
  - בחרו באופן אקראי את הדלת אותה בוחר השחקן
  - קיבעו מהי הדלת אותה פותח המנחה (בהתאם לכללי המשחק)
  - בהתאם לסוג השחקן (מבין השלושה שתוארו למעלה) קיבעו האם הוא בוחר להישאר עם בחירתו המקורית או להחליפה
  - בידקו האם בחירתו הסופית של השחקן הובילה לזכייה במכונית, או לא

ב. שחקן א משחק 1000 פעמים ברציפות. הריצו את הקוד שכתבתם 1000 פעם ושימרו את התוצאות. ציירו גרף המתאר, לאחר כל משחק, את מספר הפעמים שהשחקן זכה במכונית. בציר האופקי ציינו מספר המשחקים ששוחקו ובציר האנכי את שיעור הזכיות במכונית עד לאותו המשחק (כלומר הערכים בציר האופקי יהיו בין 1 ל 1000, ובציר האנכי בין 0 ל-1).

ג. חיזרו על הנ"ל בנפרד עבור שחקן ב ושחקן ג. מי השחקן שהטקטיקה שלו הובילה לסיכוי הגבוה ביותר לזכות במכונית? הסבירו בקצרה את התוצאות.

ד. ענו בקצרה על השאלה הבאה: בהינתן התוצאות של שחקנים א ו-ב, האם ניתן היה להעריך את סיכויי ההצלחה של שחקן ג ללא הרצת הסימולציה עבורו?

4. KNN – חישוב ידני

נתון מידע המכיל 8 דוגמאות מתויגות $\mathcal{D} = \{(\boldsymbol{x_n}, \boldsymbol{y_n})\}_{n=1}^{8}$ כמפורט בטבלה הבאה, כאשר וקטור המאפיינים הוא בעל שני מימדים $\boldsymbol{x} = [x_1, x_2] \in \mathbb{R}^2$, התיוג הוא בינארי, וקטור המאפיינים של הדוגמא ה-n-ית מיוצג על ידי $\boldsymbol{x_n} = [x_{n,1}, x_{n,2}]$, והתיוג הבינארי של הדוגמא ה-n-ית מיוצג כ- $y_n = 1$ או $y_n = 0$

n	$x_{n,1}$	$x_{n,2}$	y_n
1	8.1	-1.1	0
2	9.9	2.8	1
3	7.8	2.1	1
4	9.2	0.7	0
5	10.6	0.7	1
6	6.4	2.1	0
7	6.6	1.0	0
8	9.5	-1.1	0

הדוגמאות הנתונות ב-$\mathcal{D}$

א. ציירו את המידע על גרף דו-מימדי עם הצירים x_1 ו- x_2

ב. הוסיפו לגרף הנ"ל את הנקודה הלא מסווגת $\boldsymbol{x_9} = [8.0, 1.5]$

ג. חשבו את המרחק האויקלידי של $\boldsymbol{x_9}$ מכל אחת מהנקודות המסווגות, והסבירו את אופן החישוב (די בהסבר עבור אחת מהדוגמאות המתויגות)

ד. סדרו את הנקודות המסווגות על-פי מרחקן האויקלידי מ $\boldsymbol{x_9}$

ה. חשבו את $\hat{y}_9$, סיווג KNN של $\boldsymbol{x_9}$, עבור כל אחד מהערכים $\mathrm{K} = [1, 2, 3, 4, 5, 6, 7, 8]$. הסבירו את אופן החישוב והתייחסו לסוגיה הבאה: האם יש מקום לבחון את כל ערכי K הנ"ל, או שישנם ערכים שאינם מתיישבים עם ההיגיון הפנימי של מסווג KNN?

ו. חיזרו על סעיפים ב-ה עבור $\boldsymbol{x_{10}} = [7.0, 0]$. התייחסו בקצרה לסוגית מובהקות הסיווג: האם לדעתכם ניתן להסתמך על הסיווגים $\hat{y}_9$ ו-$\hat{y}_{10}$ באופן דומה?

5. KNN – מימוש בפייתון

בתרגיל זה תממשו מסווג KNN בפייתון ע"פ השלבים שתארנו בכיתה (ובאופן התואם את שלבי החישוב הידני שביצעתם בשאלה הקודמת), תבחנו את ביצועיו על המידע של מאפייני זני האירוס אותו ראינו בהרצאה, ותבחרו ערך מתאים של K עבור מידע זה.

א. ממשו את אלגוריתם KNN, כולל כל השלבים הבאים:
   i. חלוקת המידע לסדרת אימון ומבחן
   ii. מדידת הדיוק עם סדרת האימון
   iii. מדידת הדיוק עם סדרת המבחן

הריצו את הקוד והדפיסו גרף המסכם את דיוקי המסווג על סדרת המבחן וסדרת האימון עבור כל ערכי K מ K=1 עד K=20 (בציר אופקי הציגו את ערכי K ובציר האנכי את דיוקי המסווג כשני גרפים, אחד עבור סדרת האימון והאחר עבור סדרת המבחן).
הסבירו את התוצאות המוצגות בגרף.

ב. בסעיף זה הנכם מתבקשים לחזור על התהליך שביצעתם בסעיף א 100 פעם, כשבכל פעם תוגרל חלוקה שונה של המידע לסדרות אימון ומבחן, ובסופו של דבר עליכם ליצור גרף יחיד הדומה במהותו לזה שבסעיף א, אך עבור כל K לדווח בו את ממוצע הדיוק (accuracy) שהתקבל על פני כל 100 ההרצות.
הסבירו את הגרף שקיבלתם, והשתמשו בו על מנת לקבוע ערך K מתאים למידע הנתון, שבו תמליצו להשתמש לסיווג של פרחים שאינם ידועים לכם בשלב האימון. נמקו בחירתכם.
שימו לב: בסעיף א וקטור הדיוקים שחישבתם הינו באורך Kmax=20. בסעיף ב הינכם נדרשים ראשית לחשב מטריצת דיוקים במימד 20 x 100, ואז לחשב את הממוצע עבור כל ערך של K ולקבל וקטור של ממוצעי דיוקים (שוב באורך 20).