

המחלקה להנדסת תוכנה

10/07/19  
09:00-12:00

**מבוא לכריית נתונים**  
**מועד ב'**  
**יניב הדר**  
תשע"ט סמסטר ב'

חומר עזר – דף נוסחאות (דף אחד משני הצדדים) ומחשבון מדעי ללא יכולות תכנות.

**חומר עזר : נא סמן במשבצת המתאימה את המתאים**

☒ \* ניתן להשתמש בכל מחשבון ללא יכולת תכנות  
\* ☐ לא ניתן להשתמש במחשבון Casio FX-991EX  
\* ☐ לא ניתן להשתמש במחשבון

\* ☐ לא ניתן להשתמש בחומר עזר  
\* ☒ \* מותר שימוש בדף נוסחאות, כמפורט: דף אחד משני הצדדים  
\* ☐ הבחינה בחומר פתוח – מותר להשתמש בכל חומר עזר מודפס או כתוב

**הערות**  
יש לענות על כל השאלות במקומות המיועדים ע"ג טופס השאלון בלבד  
\* ☒ יש להחזיר את השאלון ביחד עם הכריכה/מחברת.  
\* ☐

אחר:

1. יש לענות תשובה סופית על גבי השאלון !  
2. \_\_\_\_\_  
3. \_\_\_\_\_

השאלון מכיל 8 עמודים (כולל עמוד זה).

**בהצלחה !**

=====



חלק א' (40 נקודות, 4 נקודות לכל שאלה)

יש לבחור תשובה אחת בלבד לכל שאלה ולסמן על גבי השאלון בלבד, יש לתת נימוק קצר לתשובה במקום המסומן – תשובות במחברת לא ייבדקו.

1. במודל K-NN (סמן את התשובה הנכונה) :

1. לכל K התצפיות יש משקל שווה בחישוב הערך החזוי
2. תצפיות קרובות לתצפית החדשה משפיעות יותר מאשר תצפיות רחוקות
3. ערך ה-K קטן בהתאם לחשיבות התצפיות
4. ערך ה-K גדל בהתאם לחשיבות התצפיות

2. בקטע טקסט ארוך מחשבים אנטרופיה לאותיות לעומת האנטרופיה של המילים, מה המשפט הנכון ביותר

?

1. האנטרופיה של המילים גדולה מהאנטרופיה של האותיות.
2. האנטרופיה של המילים קטנה מהאנטרופיה של האותיות.
3. האנטרופיה של המילים שווה לאנטרופיה של האותיות.
4. לא ניתן לחשב אנטרופיה בטקסט

3. חלוקת ערכים לאינטרוולים לפי אנטרופיה

1. מקטינה את סטיית התקן של הנתונים
2. מקטינה את השונות של הנתונים
3. יוצרת אינטרוולים באופן לא מונחה (unsupervised)
4. תשובות א' ו-ב' נכונות

4. עפ"י תורת האינפורמציה, אי וודאות האירוע שווה למקסימום אם

1. כל התוצאות הן ערך דטרמיניסטי (קבוע) יחיד
2. התוצאות מתפלגות התפלגות מעריכית
3. התוצאות מתפלגות התפלגות נורמאלית
4. אף תשובה אינה נכונה



5. באלגוריתם Naïve Bayes, כאשר מכפלת ההסתברויות המותנות שווה לאפס?

1. תוצאה זו אפשרית לאלגוריתם זה ואינה ניתנת לתיקון.
2. באלגוריתם זה לעולם לא יקרה מצב בו מכפלת ההסתברויות המותנות מתאפסת.
3. באלגוריתם זה אנחנו מסכמים את ההסתברויות המותנות.
4. נוסיף אחד במונה ובמכנה של כל הסתברות מותנית בכדי להימנע ממצב זה.

6. אלגוריתם מסוג eager לעומת אלגוריתם מסוג lazy:

1. eager מדויק מ lazy.
2. eager שומר חלק מנתוני האימון ו-lazy שומר מודל.
3. גם בבניית המודל וגם בסיווג eager מהיר ו-lazy איטי.
4. אף תשובה אינה נכונה.

7. ברצוננו לחלק את התצפיות כך שהתצפיות בעלות הסיווג השולט (לפי חוק הרוב) יהיו נפרדות משאר התצפיות. באיזה מדד כדאי להשתמש?

1. Information gain
2. Gain ratio
3. Gini index
4. אף תשובה אינה נכונה

8. כדי לדעת מה הסיכויים שרכישת חיתולים תוביל לרכישת בירה, המדד שצריך לחשב הוא:

1. confidence
2. mutual information
3. conditional probability
4. support



9. בבעיית סיווג בינארי, דיוק האימון של מודל בעל אנטרופיה השווה ל-0.5 הוא:

1. 50%
2. 0%
3. 100%
4. לא ניתן לדעת

10. בבניית עץ סיווג (יש לסמן את המשפט הנכון ביותר)

1. ככל שהעץ בעל יותר קודקודים כך דיוק המודל גדל.
2. ככל שהעץ עמוק יותר כך נסווג מהר יותר.
3. ככל שהעץ בעל יותר קודקודים כך דיוק המודל קטן.
4. ככל שהעץ רחב יותר כך נסווג מהר יותר.

חלק ב' (60 נקודות)

בכל הסעיפים בחלק זה יש להראות את כל החישובים הרלוונטיים במחברת, יש לרשום תשובה סופית על גבי השאלון בלבד, הניקוד הוא על תשובה סופית מדויקת בלבד, אולם תשובה ללא חישובים רלוונטיים במחברת תיפסל.

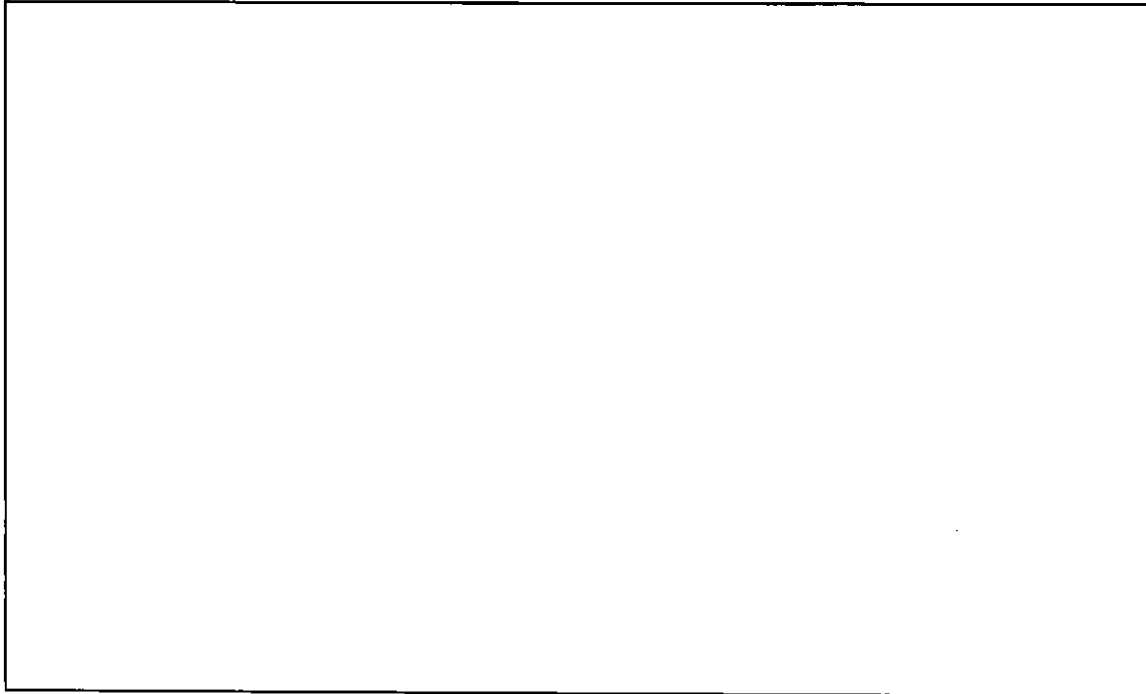
ביקום גיבורי העל של מארוול מנסים לסווג מיהו גיבור על ומי נוול. לשם כך, נאסף מידע על עשר דמויות המכיל 6 משתנים מועמדים ומשתנה מטרה אחד – סיווג. הנתונים בטבלה:

סיווג	נשק	מונולוג	צבע גלימה	גובה קפיצה	מדד כח	מין	רשומה
superhero	TRUE	FALSE	red	17	100	female	1
villain	FALSE	TRUE	black	32	35	female	2
villain	TRUE	FALSE	blue	15	29	male	3
villain	TRUE	TRUE	red	5	45	male	4
superhero	FALSE	FALSE	red	27	87	male	5
superhero	FALSE	TRUE	blue	9	44	male	6
villain	TRUE	FALSE	red	3	50	female	7
villain	FALSE	FALSE	black	2	67	male	8
villain	FALSE	TRUE	black	24	20	male	9
superhero	TRUE	FALSE	red	30	98	female	10

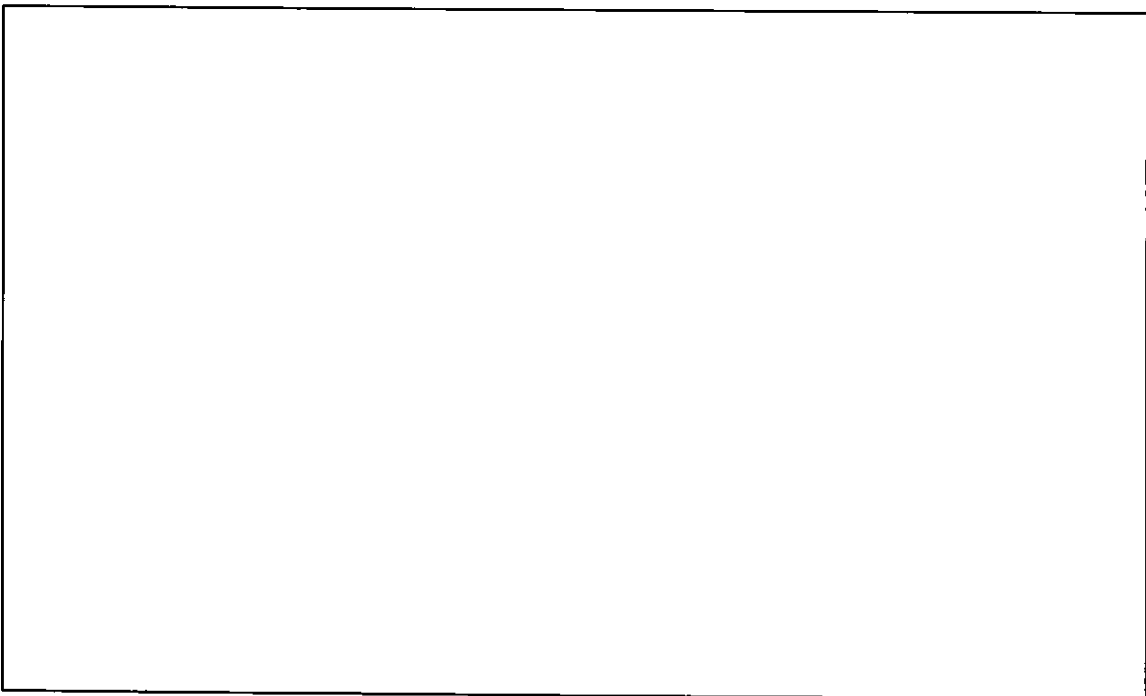




1. (4 נק') יש לבצע נרמול בשיטת min-max לגובה הקפיצה



2. (4 נק') יש לבצע דיסקרטיזציה לשני טווחים (BINS) בשיטת רוחב שווה לערכי גובה הקפיצה המנורמלים.





3. (4 נק') מהו דיוק חוק הרוב בנתונים שלהלן?

4. (6 נק') מהי האנטרופיה של משתנה המטרה סיווג?

5. (6 נק') האם אישה בעלת נשק וגלימה אדומה היא תבנית שכיחה?  
(Minimum support = 30%)

6. (12 נק') יש להשתמש במדד ה-Gini של משתנה המטרה ביחס למשתנה "מדד כח" על מנת לבדוק לפי איזה ערך (Threshold) כדאי לבצע דיסקרטיזציה למשתנה "מדד כח" 40 או 65

N(L,R)	Yes		No		Total	
	L	R	L	R	L	R
40						
65						

P(L,R)	Yes		No		Gini		Gini Split
	L	R	L	R	L	R	
40							
65							

הערך האופטימלי לדיסקרטיזציה של המשתנה "מדד כח" הוא ? נמק בקצרה :



7. (12 נק) יש להשתמש באלגוריתם Naïve Bayes על מנת לסווג את התצפית הבאה

רשומה	מין	מדד כח	גובה קפיצה	צבע גלימה	מונולוג	נשק	סיווג
1	female	95	20	blue	FALSE	TRUE	???

עבור המשתנים גובה קפיצה ומדד כח יש להשתמש במשתנים לאחר הדיסקרטיזציה (מסעיף 2 ו 6)

שימו לב ייתכן ותצטרכו להשתמש ב Laplacian estimate



## המחלקה להנדסת תכנה

8. (12 נק') יש להשתמש באלגוריתם (KNN (K nearest neighbors על מנת לסווג את התצפית הבאה ( $K=3$ ):

סיווג	נשק	מונולוג	צבע גלימה	גובה קפיצה	מדד כח	מין	רשומה
???	TRUE	FALSE	blue	20	95	female	1

על מנת לחשב מרחק בין תצפיות יש להשתמש במרחק מנהטן כאשר המרחקים מוגדרים באופן הבא:

- עבור המשתנים מין, נשק ומונולוג המרחק בין ערכים שונים הוא 1.
- עבור המשתנים גובה קפיצה ומדד כח יש להשתמש במשתנים לאחר הדיסקרטיזציה (מסעיף 2 ו 6)
- המרחק בין אינטרוולים שונים הוא 1.
- עבור המשתנה "צבע גלימה" יש להשתמש בטבלת המרחקים הבאה:

צבע גלימה	red	blue	black
red	0		
blue	1	0	
black	2	1	0

