המכללה האקדמית להנדסה סמי שמעון

מדור בחינות ומערכת שעות

המחלקה להנדסת תוכנה

25/6/2019
09:00-12:00

מבוא לכריית נתונים

מועד א'

יניב הדר

תשע"ט סמסטר ב'

חומר עזר – דף נוסחאות

השאלון מכיל 8 עמודים.

בהצלחה !

==========

חומר עזר : נא סמן במשבצת המתאימה את המתאים

X * ניתן להשתמש בכל מחשבון (ללא יכולות תכנות)
___ * לא ניתן להשתמש במחשבון Casio FX-991EX
___ * **לא** ניתן להשתמש במחשבון

___ * **לא** ניתן להשתמש בחומר עזר
X * מותר שימוש בדף נוסחאות, כמפורט: דף אחד משני הצדדים (סטודנטים מביאים עימם)
___ * הבחינה בחומר פתוח – מותר להשתמש בכל חומר עזר מודפס או כתוב

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

אחר:
1. _____
2. _____
3. _____

השאלון מכיל 8 עמודים (כולל עמוד זה).

בהצלחה !

==========

חלק א' (40 נקודות, 4 נקודות לכל שאלה)

<u>יש לבחור תשובה אחת בלבד לכל שאלה ולסמן על גבי השאלון בלבד – תשובות במחברת לא ייבדקו.</u>

1. אדם ביצע בדיקה לנוכחות מחלה מסוימת. הבדיקה יצאה חיובית והוא אכן חולה. זוהי דוגמה ל:
   א. False negative
   ב. False positive
   ג. True positive
   ד. True negative

2. המרחק בין שני אשכולות בשיטת Centroid מחושב עפ"י
   א. המרחק בין מרכזי האשכולות
   ב. המרחק הארוך ביותר בין שני אובייקטים בשני האשכולות
   ג. המרחק הממוצע בין שני האשכולות
   ד. המרחק הקצר ביותר בין שני אובייקטים בשני האשכולות

3. חלוקת ערכים לאינטרוולים לפי עומק שווה
   א. מקטינה את סכיית התקן של הנתונים
   ב. יוצרת אינטרוולים בעלי שכיחות שווה
   ג. יוצרת אינטרוולים בעלי טווח שווה
   ד. תשובות א' ו-ב' נכונות

4. עפ"י תורת האינפורמציה, אי וודאות האירוע שווה למינימום אם
   א. כל התוצאות הן ערך דטרמיניסטי (קבוע) יחיד
   ב. התוצאות מתפלגות התפלגות מעריכית
   ג. התוצאות מתפלגות התפלגות נורמאלית
   ד. התוצאות מתפלגות התפלגות אחידה

5. למה אלגוריתם Naïve Bayes נאיבי?
   א. ההנחה שקיימת תלות זהה בין כל משתנה למשתנה היא נאיבית.
   ב. ההנחה שלא קיימת תלות בין אף אחד מהמשתנים היא נאיבית.
   ג. ההנחה שניתן לסכום את המשתנים היא נאיבית.
   ד. ההנחה שקבוצת על של קבוצת פריטים שכיחים חייבת להיות שכיחה היא נאיבית

6. אלגוריתם מסוג eager לעומת אלגוריתם מסוג lazy:
   א. eager שומר את כל נתוני האימון ו-lazy שומר חלק מנתוני האימון
   ב. eager שומר חלק מנתוני האימון ו-lazy שומר את כל נתוני האימון
   ג. eager שומר את נתוני האימון ו-lazy שומר מודל
   ד. eager שומר מודל ו-lazy שומר את נתוני האימון

7. בית אריזה משתמש להפריד בעזרת צילום חמנות, בין תפוחים לבננות וללא סיווג ראשוני. מדובר במשימה של:
   א. ניתוח אשכולות (Clustering)
   ב. סיווג (Classification)
   ג. חיזוי (Prediction)
   ד. למידה מונחית.

8. כדי לדעת מה הסיכויים שקבוצת פריטים אכן תידכש ביחד , המדד שצריך לחשב הוא:
   א. confidence
   ב. mutual information
   ג. conditional probability
   ד. support

9. בבעיית סיווג בינארי, דיוק האימון של מודל בעל אנטרופיה השווה ל-1 הוא:
   א. 50%
   ב. 0%
   ג. 100%
   ד. לא ניתן לדעת

10. נתונה הפונקציה Y=2X+3. האנטרופיה של Y בהינתן X שווה ל:
   א. 1
   ב. 0.5
   ג. 0
   ד. לא ניתן לדעת

חלק ב' (60 נקודות)

<u>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.</u>

זואולוגים מעוניינים לחזות האם נחש הוא ארסי. לשם כך, נאסף מידע מ 10 נחשים המכיל 5 משתנים מועמדים ומשתנה מטרה אחד – סיווג. הנתונים בטבלה:

רשומה	מין	אורך	צבע קשקשים עיקרי	צבע אחיד	ניבים	סיווג
1	male	24	yellow	TRUE	TRUE	yes
2	male	32	black	FALSE	FALSE	no
3	male	15	black	FALSE	TRUE	no
4	male	18	brown	TRUE	TRUE	yes
5	female	27	yellow	FALSE	FALSE	yes
6	male	9	yellow	TRUE	FALSE	yes
7	female	3	black	FALSE	TRUE	no
8	female	18	black	FALSE	FALSE	yes
9	male	22	black	TRUE	FALSE	no
10	male	21	brown	FALSE	FALSE	yes

א. (4 נק') יש לבצע נרמול בשיטת decimal scaling לאורך הנחשים

ב. (4 נק') יש לבצע דיסקרטיזציה לשני טווחים (BINS) בשיטת **עומק שווה** לערכי אורך הוחזלים המנורמל.

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

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

ה. (6 נק') האם מין בשרך male ו – **צבע אחיד** בשרך FALSE הם תבנית שכיחה (Minimum support = 30%)?

ו. (12 נק) יש לחשב את מדד ה-Gini (Gini index) של משתנה המטרה – "סיווג" עבור כל אחת מהתכונות "גיבים" ו – "צבע אחיד" ולהחליט לפי איזו תכונה עדיף לפצל את קדקוד השורש.

ז. (12 נק) יש לחשב את מדד ה – gain ratio של משתנה המטרה – **"סיווג"** עבור כל אחת מהתכונות **"צבע קשקשים"** ו – **"צבע אחיד"** ולהחליט לפי איזו תכונה עדיף לפצל את קודקוד השורש.

ח. (12 נק') יש להשתמש באלגוריתם KNN (K nearest neighbors) על מנת לסווג את התצפית הבאה (K=1):

רשומה	מין	אורך	צבע קשקשים עיקרי	צבע אחיד	ביבים	סיווג
1	female	50	yellow	FALSE	TRUE	???

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

- עבור המשתנים מין, ביבים וצבע אחיד המרחק בין ערכים שונים הוא 1.
- עבור המשתנה אורך יש להשתמש במשתנה לאחר הדיסקרטיזציה (מסעיף ב') – המרחק בין אינטרוולים שונים הוא 1.
- עבור המשתנה "צבע קשקשים עיקרי" יש להשתמש בטבלת המרחקים הבאה:

צבע קשקשים עיקרי	brown	yellow	black
brown	0		
yellow	1	0	
black	2	1	0