אוניברסיטת בן-גוריון בנגב
Ben-Gurion University of the Negev

תאריך הבחינה: 29/07/2011
שם המרצה: פרופ' מרק לסט
שם הקורס: כריית נתונים ומחסני נתונים
מספר הקורס: 372-1-3105
שנה: 2011 סמסטר: ב' מועד: א'
משך הבחינה: 3 שעות
חומר עזר: דף נוסחאות (מצורף לבחינה) + מחשבון
גרסה מס' 1

חלק 1 [50 נקודות]

- יש לענות על כל השאלות
- משקל של כל שאלה – **5 נקודות**
- יש לרשום את התשובה בכתב-יד ברור **במקום המיועד לכך על-גבי שאלון הבחינה בלבד**
- תשובה לא מנומקת (במידה ונדרשת הנמקה) **תקבל ציון של אפס**

א. יש להציג שתי דוגמאות של תבניות הסתברותיות שהייתם מצפים למצוא בבסיס נתונים של חברת נסיעות גדולה.

1. _____

2. _____

ב. מה השלב הראשון בתהליך של גילוי ידע בבסיסי נתונים (knowledge discovery in databases)?

_____

ג. נכון/לא נכון: ב-Association Rules, אם נעביר פריט אחד מהצד השמאלי של החוק לצד הימני של החוק, אז ה-confidence של החוק לעולם לא יהיה נמוך יותר. (דוגמא: מעבר מחוק $X,Y \rightarrow Z$ לחוק $X \rightarrow Y,Z$). **יש לנמק בקצרה את תשובתכם.**

_____

_____

ד. נכון/לא נכון: בדיקת השערות סטטיסטית מהווה פעולה של כריית מידע. **יש לנמק בקצרה את תשובתכם.**

_____

_____

בהצלחה!

ה. עליכם לבנות מחסן נתונים עבור נתונים רב-שנתיים של כמות הגשמים היורדים ברחבי הארץ. מה יהיו המימדים (dimensions) והעובדות (facts) בתרשים הכוכב של מחסן זה?

עובדות (לפחות אחת): _____

מימדים (לפחות שניים): _____

_____

ו. מה משמעות התכונה Non-volatility ("אי-נדיפות") של מחסני נתונים?

_____

_____

ז. מהו תחום הערכים של אנתרופיה המחושבת עבור משתנה מקרי נתון?

_____

ח. מה הבדל בין השיטות single link ו-complete link לחישוב המרחק בין שני אשכולות?

_____

_____

_____

ט. יש לציין לפחות שתי שיטות שונות לטיפול בערכים חסרים (missing data)

1. _____

2. _____

י. מה מאפיין כל רמה (שכבה) במודלים של רשתות אינפו-עמומות (Info-Fuzzy Networks)?

_____

_____

_____

חלק 2 [50 נקודות]

- יש להציג את כל התוצאות עם **שלוש ספרות אחרי נקודה עשרונית** אלא אם צוין אחרת!
- יש לרשום את כל התשובות **על-גבי שאלון הבחינה בלבד**
- טיוטות החישוב ייגרסו **ללא בדיקה**

נתון בסיס הנתונים המסוכם הבא:

# id	Age	Gender	Income	Marital Status	Car Type Count		
					Family	Luxury	Sports
1	25-33	Male	10,000	Married	10	5	7
2	34-45	Female	12,000	Single	6	8	9
3	46-55	Female	16,000	Divorced	5	4	19
4	56+	Female	7,000	Single	4	6	10
5	25-33	Male	16,000	Widowed	20	4	7
6	34-45	Male	12,000	Divorced	8	9	10
7	46-55	Female	7,000	Married	4	2	0
8	56+	Male	7,000	Single	3	1	7
9	46-55	Male	12,000	Widowed	9	5	1
10	56+	Female	10,000	Divorced	3	7	31
11	56+	Female	7,000	Married	42	2	1
12	25-33	Male	10,000	Single	20	3	2
13	34-45	Male	16,000	Single	0	10	13
14	46-55	Male	12,000	Divorced	1	5	8
15	25-33	Female	7,000	Married	17	0	2

Attribute	Type	Use
# id	Numeric	ID
Age	Nominal	Input
Gender	Nominal	Input
Income	Numeric	Input
Marital Status	Nominal	Input
Car Type	Nominal	Decision Class

שימו לב:

- משתנה ה-#id הוא לשימוש בסעיף ו'.
- העמודות "Car Type Count" מייצגות עבור כל אחד מהערכים האפשריים של המשתנה Car Type (Family, Luxury, Sports) את מספר המופעים בבסיס הנתונים המקורי. לדוגמא, עבור הרשומה #id=1 בבסיס הנתונים המסוכם:
  - Car Type: Family = 10 ← יש 10 רשומות בבסיס הנתונים המקורי כמו הרשומה שבה #id=1 והסיווג שלהן הינו Family

א. יש לחשב אנתרופיה בלתי מותנית עבור המשתנים Age ו-Car Type . **10 נקודות**
נא להציג את החישובים בטבלאות הבאות:

Age	Count	Probability	Entropy
Total			

Car Type	Count	Probability	Entropy
Total			

ב. יש לחשב את הרווח האינפורמטיבי (Information Gain) של משתנה הסיווג "Car Type" עבור כל נקודות הפיצול האפשריות של המשתנה הרציף "Income". יש לציין בעיגול את נקודת הפיצול הטובה ביותר. **10 נקודות**.

	Left Interval		Right Interval		Total	Information
Threshold	Percentage	Entropy	Percentage	Entropy	Entropy	Gain

ג. יש לנרמל את כל הערכים של המשתנה Income בשיטת min-max normalization לתחום שבין אפס לאחד. **5 נקודות**.

# id	Income	Normalized Income
1	10,000	
2	12,000	
3	16,000	
4	7,000	
5	16,000	
6	12,000	
7	7,000	
8	7,000	

# id	Income	Normalized Income
9	12,000	
10	10,000	
11	7,000	
12	10,000	
13	16,000	
14	12,000	
15	7,000	

Min =
Max =

בהצלחה!

ד. סווג את הרשומה הבאה על-ידי שימוש באלגוריתם K-NN, כאשר K=1. **10 נקודות**.

# id	Age	Gender	Income	Marital Status	Car Type
16	46-55	Female	8,200	Married	

עליך להשתמש **<u>בהנחות הבאות</u>**:

- יש להתייחס לכל שורה בטבלת הנתונים המסוכמת כאילו שהיא מייצגת <u>תצפית אחת</u> שהסיווג שלה נקבע עפ"י <u>חוק הרוב</u>.

- חישוב המרחקים:

  - חישוב המרחק יתבסס על המשתנים: Income, Marital Status בלבד.
  - עבור משתנה נומרי – חשב את המרחק בין <u>הערכים המנורמלים</u> שנקבעו עפ"י השיטה המוגדרת בסעיף הקודם.
  - עבור משתנה נומינלי – חשב את המרחק על סמך Simple matching.

# id	Income	Marital Status	Car Type	income-distance	marital status-distance	total distance
1	10,000	Married				
2	12,000	Single				
3	16,000	Divorced				
4	7,000	Single				
5	16,000	Widowed				
6	12,000	Divorced				
7	7,000	Married				
8	7,000	Single				
9	12,000	Widowed				
10	10,000	Divorced				
11	7,000	Married				
12	10,000	Single				
13	16,000	Single				
14	12,000	Divorced				
15	7,000	Married				
16	**Minimal distance**					

ה. בהנחה שכל רשומה בטבלת הנתונים המסוכמת מייצגת תצפית אחת בלבד (בדומה לסעיף הקודם), יש לבצע את <u>האיטרציה הראשונה בלבד</u> של האלגוריתם k-means לניתוח אשכולות על המשתנים הנומינליים הבאים: Age, Gender, Marital Status. **15 נקודות**

עליך להשתמש **<u>בהנחות הבאות</u>**:

- K=4 (מספר האשכולות שווה לארבעה).

בהצלחה!

- החלוקה הראשונית של התצפיות לאשכולות נתונה בטבלה של האיטרציה מס' 1 (בעמודה "old cluster").
- לחישוב המרחקים יש להשתמש ב-Simple Matching, המרחק בין תצפית לאשכול יחושב כמרחק בין תצפית לווקטור המייצג את האשכול (centroid).
- הערכים של הווקטור המייצג את האשכול (centroid) ייקבעו לפי ה-majority rule. במידה ויש שוויון בין שכיחות הערכים של משתנה מסוים, יש לקבוע את הערך המייצג לפי הערך של התצפית עם ה-#id הנמוך ביותר. לדוגמא:

# id	Age	Gender
1	25-33	Male
2	34-45	Female
3	46-55	Female
Centroid	**25-33**	**Female**

הווקטורים המייצגים (centroids):

Cluster	Age	Gender	Marital Status
1			
2			
3			
4			

איטרציה מס' 1:

# id	Age	Gender	Marital Status	old cluster	Distance to cluster1	Distance to cluster2	Distance to cluster3	Distance to cluster4	new cluster
1	25-33	Male	Married	1					
5	25-33	Male	Widowed	1					
9	46-55	Male	Widowed	1					
13	34-45	Male	Single	1					
2	34-45	Female	Single	2					
6	34-45	Male	Divorced	2					
10	56+	Female	Divorced	2					
14	46-55	Male	Divorced	2					
3	46-55	Female	Divorced	3					
7	46-55	Female	Married	3					
11	56+	Female	Married	3					
15	25-33	Female	Married	3					
4	56+	Female	Single	4					
8	56+	Male	Single	4					
12	25-33	Male	Single	4					