



המכללה האקדמית להנדסה סמי שמעון

מדור בחינות ומערכת שעות

המחלקה להנדסת תוכנה

מבוא לכריית נתונים

מועד א'

גב' ניבה חזון

תשע"ח סמסטר ב'

9:00-12:00

20/06/2018

הוראות מיוחדות

- חומר עזר: דף נוסחאות A4 כתוב משני צדדים.
- מחשבון פשוט ללא יכולת תכנות.
- משך המבחן 3 שעות.
- יש לענות על כל השאלות.
- כתבו את תשובתכם בכתב קריא ומסודר.
- יש לענות אך ורק בטופס הבחינה במקום המיועד לכך.

בהצלחה !

השאלון מכיל 6 דפים (כולל דף זה).

=====

חלק א' (40 נקודות, 4 נקודות לכל שאלה)

יש לבחור תשובה אחת בלבד לכל שאלה.

1. עקרון Apriori אומר:
 - א. כל תת קבוצה של קבוצת פריטים שכיחים חייבת להיות שכיחה
 - ב. קבוצת על של קבוצת פריטים שכיחים חייבת להיות שכיחה
 - ג. לפחות תת קבוצה אחת של קבוצת פריטים שכיחה חייבת להיות שכיחה
 - ד. כל התשובות נכונות
2. המרחק בין שני אשכולות בשיטת single link 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 שונות הנתונים
 - ב. יוצרת אינטרוולים בעלי שכיחות שווה
 - ג. יוצרת אינטרוולים בעלי רוחב שווה
 - ד. תשובות א' ו-ב' נכונות
4. האנטרופיה המקסימלית של משתנה בינארי שערכיו האפשריים הם 0 או 1 תתקבל כאשר
 - א. כל הרשומות בעלות ערך 0
 - ב. חצי מהרשומות בעלות ערך 0 וחצי בעלות ערך 1
 - ג. כל הרשומות בעלות ערך 1
 - ד. אף תשובה אינה נכונה
5. בבעיות סיווג, הסיווג של אילו קבוצות תצפיות ידוע מראש?
 - א. קבוצת האימון (training set) וקבוצת המבחן (test set)
 - ב. קבוצת המבחן (test set) וקבוצת הדירוג (score set)
 - ג. קבוצת הדירוג (score set) וקבוצת האימון (training set)
 - ד. אף תשובה אינה נכונה
6. התחזית לשער הביטקוין אתמול הייתה \$8,000 אך מחירו בפועל הגיע ל-\$10,000. מה יהיה השער החזוי של הביטקוין היום (באלפי דולרים) עפ"י הממוצע הנע האקספוננציאלי (exponential moving average)?
 - א. $\alpha + 8$
 - ב. $10 - \alpha$
 - ג. $2\alpha + 8$
 - ד. $10 - 2\alpha$
7. מערכת המאפשרת למכונת אוטומטית לזהות את מצב הרמזור (אדום/ירוק/צהוב) מבצעת משימה של
 - א. רגרסיה (Regression)
 - ב. ניתוח אשכולות (Clustering)
 - ג. סיווג (Classification)

ד. למידה בייסיאנית (Bayesian learning)

8. הבעיה שמדד ה-gain ratio מנסה לפתור היא

- א. ההסתברות של קבוצת תצפיות לקבל סיווג זהה עולה עם הירידה בגודל הקבוצה
- ב. ההסתברות של קבוצת תצפיות לקבל סיווג זהה יורדת עם העלייה בגודל הקבוצה
- ג. שתי התשובות נכונות
- ד. אף תשובה אינה נכונה

9. במודל שסובל מ-overfitting

- א. דיוק קבוצת המבחן (test set) גבוה בצורה משמעותית מדיוק קבוצת האימון (train set)
- ב. דיוק קבוצת האימון ודיוק קבוצת המבחן שווים
- ג. דיוק קבוצת האימון גבוה בצורה משמעותית מדיוק קבוצת המבחן
- ד. אף תשובה אינה נכונה

10. מודל אשכולות (clustering) נחשב למודל טוב אם:

- א. המרחק בין אובייקטים בתוך האשכולות (inter-class) גבוה והמרחק בין אובייקטים בין אשכולות שונים (intra-class) נמוך
- ב. המרחק בין אובייקטים בתוך האשכולות (inter-class) נמוך והמרחק בין אובייקטים בין אשכולות שונים (intra-class) גבוה
- ג. המרחק בין אובייקטים בתוך האשכולות (inter-class) נמוך והמרחק בין אובייקטים בין אשכולות שונים (intra-class) נמוך
- ד. המרחק בין אובייקטים בתוך האשכולות (inter-class) גבוה והמרחק בין אובייקטים בין אשכולות שונים (intra-class) גבוה

חלק ב' (60 נקודות)

בכל הסעיפים בחלק זה יש להראות את כל החישובים הרלוונטיים במקומות המיועדים לכך בלבד.

רופאים מ - Cleveland Clinic Foundation מעוניינים לחזות נוכחות של מחלת לב בחולים. לשם כך, נאסף מידע מ-10 חולים המכיל 5 משתנים מתמדים ומשתנה מטרה אחד - prediction. הנתונים בטבלה:

record	gender	bs	cp	fbs	slope	prediction
1	male	24	typical angina	TRUE	downsloping	absence
2	male	32	asymptomatic	FALSE	upsloping	presence
3	male	15	asymptomatic	FALSE	downsloping	presence
4	male	18	non-anginal pain	TRUE	downsloping	absence
5	female	27	typical angina	FALSE	upsloping	absence
6	male	9	typical angina	TRUE	upsloping	absence
7	female	3	asymptomatic	FALSE	downsloping	presence
8	female	18	asymptomatic	FALSE	upsloping	absence
9	male	22	asymptomatic	TRUE	upsloping	presence
10	male	21	non-anginal pain	FALSE	upsloping	absence

א. (6 נק') יש לבצע נרמול בשיטת min-max לערכי המשתנה "bs".

ב. (6 נק') יש לבצע דיסקרטיזציה (לשני טווחים) בשיטת רחב שווה למשתנה "bs" (המנורמל).

המכללה האקדמית להנדסה סמי שמעון

קמפוס באר שבע ביאליק פינת בזל 84100 | קמפוס אשדוד ז'בוטינסקי 77245, 84 | www.sce.ac.il | חייג: *88888888

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

--

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

--

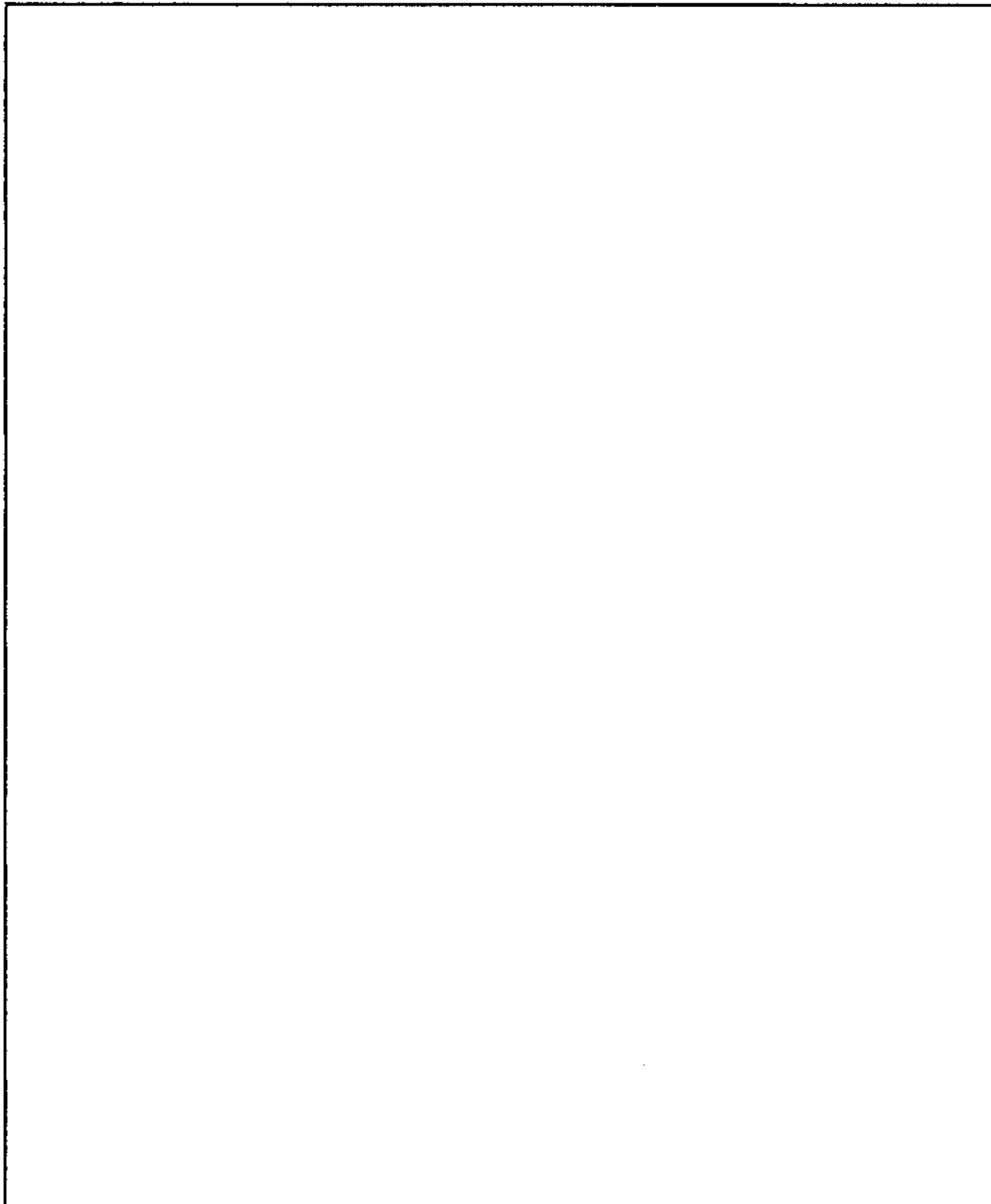
ה. (6 נק') האם $gender = male$ ו- $fbz = FALSE$ הם תבנית שכיחה (Minimum support = 30%)?

--

ו. (14 נק') יש לחשב את הרווח האינפורמטיבי (information gain) ואת מדד ה- $gain\ ratio$ של משתנה המטרה "prediction" עבור כל אחת מהתכונות "cp" ו- "fbz" ולהחליט לפי איזו תכונה עדיף לפצל את קדקוד השורש לפי כל אחד מהמדדים.

--

ז. (12 נק') ידוע כי המשתנים slope ו- fbs תלויים שניהם במשתנה gender , המשתנה bs תלוי במשתנה cp ומשתנה המטרה תלוי במשתנים bs , fbs , slope . יש לבנות רשת בייסיאנית בהתאם לנתונים (כולל טבלת ההסתברויות לכל משתנה. יש להשתמש במשתנה bs לאחר הגרמול והדיסקרטיוזציה).



ח. (6 נק') יש לחשב את ההסתברות לתרחיש הבא לפי הרשת שבנייתם בסעיף ד': גבר עם non-anginal pain , $\text{slope} = \text{upsloping}$, $\text{fbs} = \text{FALSE}$, $\text{bs} = 10$ שנמצא כי אינו סובל ממחלת לב.

