בינה עסקית

מבחן סופי - מועד א'

סמסטר א' תשע"ט

מרצה: פרופסור אילן שמשוני

מתרגל : דוד סבן

משך הבחינה שלוש שעות.

ניתן להשתמש בכל חומר עזר כתוב ומחשב כיס.

בבחינה **ארבע** שאלות.

יש לענות על השאלות **במחברת המצורפת**.

שים לב כי בידך 5 דפים כולל דף זה.

הבחינה מיועדת לגברים ונשים כאחד ומנוסחת בלשון זכר מטעמי נוחות בלבד.

שאלה	ניקוד מקסימאלי	ציון
1	30	
2	25	
3	30	
4	25	
סה"כ	110	

בהצלחה!

1) Data Warehouse(30 נקודות)

לאחד מרשתות הפארם המובילות בקשו לבנות מערכת BI שתיתן מענה למגוון השאלות עיסקיות. לרשת קיימים מספר מערכות מידע. מערכת ניהול מלאי וקטלוג מוצרים, מערכת חברי מועדון הלקוחות (פרטי לקוח וכו'), מערכת ביליניג לניהול ושמירת טרנזקציות המכירה (קבלות, חנויות וכו') . הרשת גיבשה מגוון דוחות המתבססות על המערכות המידע הקיימות:

1. סה"כ מכר כספי לפי מחלקה, קטגוריה, תת קטגוריה ומוצרים בודדים
2. סה"כ מכר כספי בציר זמן של שנים, רבעונים וחודשים
3. סה"כ מכר כספי בהשוואה לשנה הקודמת.
4. סה"כ כמות פריטים שנמכרו לפי ציר זמן של שנים וסניפים
5. סה"כ מכירות כספיות וכמות מוצריים שנמכרו בסניפים השונים בשנה האחרונה.
6. גודל סל ממוצע לפי מחלקה, קטגוריה, תת קטגוריה ומוצרים בודדים.
7. ממוצע ביקורים בחנות ותדירות ביקור.
8. כמות לקוחות חברי מועדון רשומים על ציר זמן. (על בסיס תאריך הצטרפות למועדון)
9. מכירות ממוצעות לפי קטגוריה של חברי מועדון בהשוואה ללא חברי מועדון בשנים 2017 אל מול 2018. (סה"כ 4 מדדים חברי מועדון בשנת 2017 לא חברי מועדון בשנת 2018 וכו')
10. פילוח לקוחות – גיל ממוצע של קונים לפי מחלקה, קטגוריה, תת קטגוריה
11. פילוח לקוחות – גודל סל ממוצע לפי גיל.
12. הצגת 10 המוצרים הנמכרים ביותר באתר האינטרנט.
13. השוואת ממוצע סל של אתר האינטרט אל מול החנויות הפיזיות לפי קטגוריות.

<u>הערות:</u>

- נתוני המכירות הם של חברי מועדון הלקוחות וגם של לקוחות אחרים, חברי מועדון הם בעלי סטטוס אחר במערכת מאשר שאר הלקוחות. חברי המועדון הסטטוס שהלקוח מחזיק הוא CM ושאר הלקוחות מחזיקים סטטוס NM.
- ניתן להבדיל בין המכירות של אתר האינטרנט לבין מכירות של החנויות הפיזיות לפי id של החנות, store_id = 123 הוא עבור מכירות אתר האינטרט כל שאר ה id's הם של חנויות הפיזיות השונות.
- סל היא הוצאה של לקוח ביום (חשבונית) .

a) על סמך סיפור הרקע ומגוון הדוחות המבוקשים בנה סכמת כוכב שתיתן מענה לדוחות המבוקשים. (20 נק').

b) ממש את הדוחות הבאים, ציין באיזה אובייקט תשתמש (table, bar chart, pivot, line, gauge וכו') ציין איזה ממדים תכניס ל dimension וכיצד תחשב את המדד המבוקש, המימדים והממדים חייבים להסתמך על המבנה נתונים שבנית בסעיף הקודם.

תבנית של set analysis:

Aggregation_Func({$<Filter_Filed = {} , Filter_Filed = {} > }Required_Field)

a. דוח מס' 4 (5 נק')
b. דוח מס' 5 (5 נק')
c. דוח מס' 9 (10 נק')

2) Cross Validation (25 נקודות)

בקורס גילינו ש Cross Validation עוזר בפתרון הרבה בעיות.

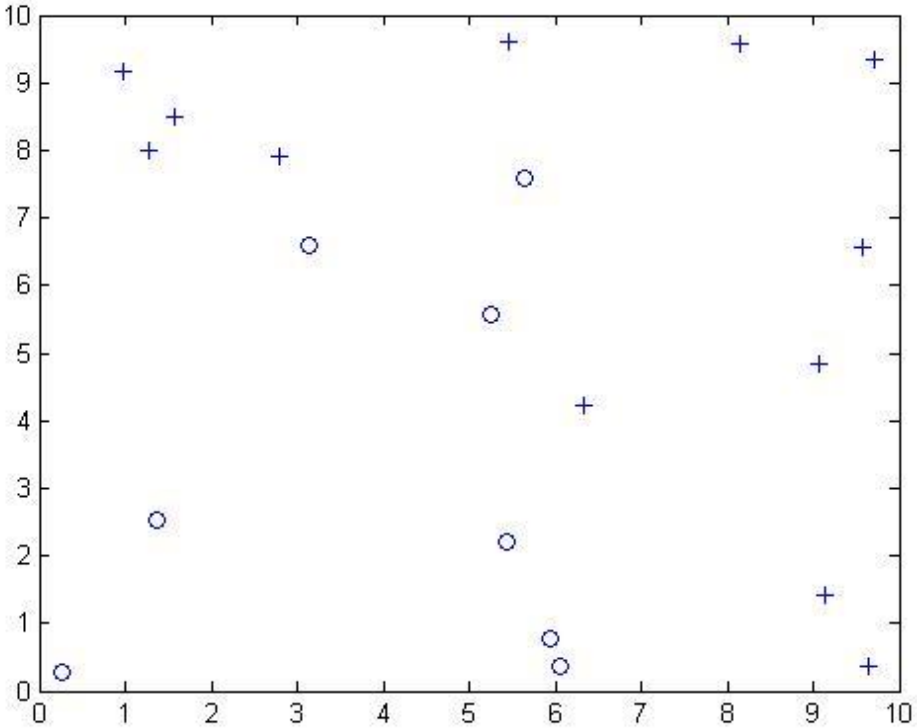
א) תן שלוש דוגמאות לשימוש ב CV לפתרון בעיות.

ב) כתוב ב pseudo-code פונקציה שמחשבת CV על מסווג כלשהוא בשם classifier שמחזירה את הדיוק של המסווג

ג) באופן עקרוני איך הדיוק של המסווג שחושב בסעיף הקודם משתנה ככל שמספר החתכים (הפרמטר לפונקציה) גדל.

ד) מה היא שיטת leave one out ואיך היא קשורה ל CV.

3) עצי החלטה (30 נקודות)



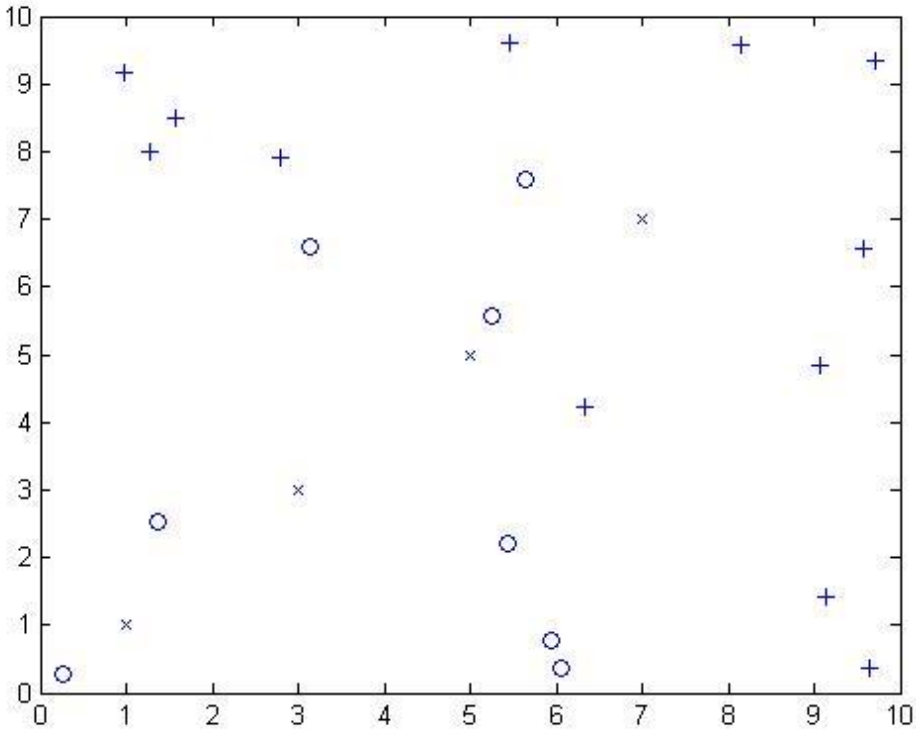
נתונות הנקודות הבאות ששייכות לשתי מחלקות. אנו רוצים לבנות עץ החלטה תוך שימוש בקריטריון mismatch.

א) בחר את החתך האופטימאלי (או קרוב לאופטימאלי) לביצוע החלוקה הראשונה. חשב את מדד ה impurity של החתך הזה. האם יש לבצע את החתך הזה? איך החלטת?

ב) המשך בבניית העץ לעוד רמה אחת. צייר את העץ (צמתים תנאים).

ג) בכל עלה מה יחליט העץ ובאיזה רמת ביטחון (הסתברות).

ד) סווג את ארבע הנקודות המסומנות ב X בתמונה הבאה תוך שימוש בעץ.



ה) איך מסווג K nearest neighbor עם K=3 היה מסווג את הנקודות האלה.

ו) ללא קשר לדוגמא לעיל: מה היתרונות והחסרונות של שימוש ב bagging שמבוסס על עצים לצרכי סיווג לעומת השימוש בעץ בודד.

4) Viola Jones (25 נקודות)

אחד האלגוריתמים שלמדנו הוא האלגוריתם של Viola & Jones לזיהוי מיקום של פרצופים בתמונה. באלגוריתם הזה היו מספר רעיונות מעניינים.

א) מה זה integral image ולמה הוא משמש. הראה דוגמא לשימוש בו באלגוריתם.

ב) מהם ה features (מאפיינים) שבהם משתמשים באלגוריתם ואיך מחשבים אותם.

ג) איך האלגוריתם בחר את המסווגים בהם הוא משתמש.

ד) למה הכוונה ב attentional cascade. למה בחרו הכותבים להשתמש בשיטה זו במקום במסווג יותר פשוט.

ה) מה החשיבות ההיסטורית של המאמר הזה.