בינה עסקית

מבחן סופי - מועד ב'

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	40	
2	27	
3	26	
4	27	
סה"כ	120	

בהצלחה!

1) Data Warehouse(40 נקודות)

אחת מרשתי ביתי הקולנוע המובילות בארץ החליטה להכניס מערכת bi על מנת שתוכל לעקוב באופן יעיל ולקבל מידע עסקי בצורה נוח וקלה מכל מערכות המידע שלהם.

מקורות המידע הקיימים הם: נתוני מכירות סרטים (סרט רגיל, vip וכו'), קטלוג סרטים (ז'אנרים), נתוני עובדים ועמלות ונתוני מכירות בקפיטריה. בנוסף לפני כשנתיים החלה הרשת בתי הקולנוע בהקמת מועדון חברים, המועדון בתמורה לתשלום שנתי מזכה הנחה בקפיטריה ובתשלום לכרטיס לסרט.

<u>ההנהלה גיבשה מס' דוחות שחושב לה שהיו במערכת החדשה:</u>

1. סה"כ מכירות כספיות בשנת 2018 של סרטים מסוג vip בחלוקה לז'אנרים
2. סה"כ מכירות כספיות של פופקורן בציר זמן של שעות (עבור על שעה ביום יוצג סהכ מכר כספי של פופקורן) .
3. חמשת הערים עם כמות הצופים הגדולה ביותר בשנת 2018
4. הוצאה כספית של לקוחות המועדון לעומת שאר הלקוחות בהשוואה לשנים 2017 ו-2018.
5. עשרת הלקוחות שצפו בכמות הסרטים הגדולה ביותר בשנת 2018.
6. עשרת הסרטים הנצפים ביותר בשנת 2018.
7. עשרת העובדים שקיבלו את העמלות הגבוהות ביותר בשנת 2018.
8. כמות המבקרים הממוצעת היומית (בכל יום בשבוע) בשנת 2018.
9. כמות ההזמנות הממוצעות ביום בשבוע (ראשון עד שבת) בשנת 2018 מהאינטרנט בהשוואה להזמנות מהסניף. (עבור כל יום בשבוע יוצג כמות ההזמנות הממוצעות ביום מהזמנות באינטרנט אל מול ההזמנות בסניף)
10. בחלוקה לסניפים הצג את ההוצאה הממוצעת לאדם בהשוואה ללקוח מועדון לשאר הלקוחות.

<u>הנחות:</u>

1. ניתן להניח שהזמנה מהאינטרנט היא הזמנה שהסניף הרשום הוא 123
2. לוקח מועדון הוא לקוח שהסטטוס שלו הוא CM שאר הלקוחות הסטטוס שלהם הוא NM

<u>שאלה 1:</u>

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

<u>שאלה 2:</u>

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 (במידה הצורך שימוש ב set analysis), המימדים והממדים חייבים להסתמך על המבנה נתונים שבנית בסעיף הקודם.

תבנית של set analysis:

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

1. דוח מס' 3 (5 נק')
2. דוח מס' 7 (5 נק')
3. דוח מס' 9 (10 נק')

2) Clustering (27 נקודות)

א) תאר איך האלגוריתם k-means עובד.

ב) מהי פונקציית המטרה שהאלגוריתם מנסה למזער?

ג) הוכח שבכל איטרציה של האלגוריתם ערך הפונקציה יורדת.

ד) האם האלגוריתם מוצא את המינימום הגלובלי של הפונקציה? אם לא אז מה הוא מוצא?

ה) האם אלגוריתם ה SOM מנסה אף הוא למצוא את המינימום של אותה פונקציה?

ו) מה ההבדלים בתהליך הלימוד בין SOM ו k-means?

ז) מי משני האלגוריתמים ימצא לרוב ערך נמוך יותר? הסבר.

3) SVM (26 נקודות)

א) מה זה SVM לינארי? איך מסווגים איתו נקודת test.

ב) מהו Kernel SVM? מה צריך להיות מוגדר כדי לבנות כזה.

ג) תן דוגמא של kernel מסוים.

INDUCTIVE SVM – THE DUAL PROBLEM

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{k=1}^{n} \alpha_k (y_k(\langle w^T, x_k\rangle + b) - 1)$$

$$\frac{d}{dw} L(w,b,\alpha) = 0 \,, \frac{d}{db} L(w,b,\alpha) = 0$$

$$w = \sum_{k=1}^{n} \alpha_k y_k x_k \qquad \sum_{k=1}^{n} \alpha_k y_k = 0$$

$$\max_{\alpha} \sum_{k=1}^{n} \alpha_k - \frac{1}{2} \sum_{k,l=1}^{n} \alpha_k \alpha_l y_k y_l \langle x_k, x_l\rangle$$

$$s.t: \alpha_k \geq 0 \,, k = 1,2,\dots,n \qquad \sum_{k=1}^{n} \alpha_k y_k = 0$$

ד) נתון השקף שמפתח את תהליך האופטימיזציה הריבועית לפתרון בעיית ה SVM. בהינתן ה kernel שבחרת, מה היית עושה כדי להכין את בעיית האופטימיזציה לפתרון (אחרי שהבעיה הוכנה לפתרון ישנה פונקציה סטנדרטית שפותרת אותה). אחרי שהתקבל הפתרון איך היית מסווג איתו נקודה לא מסווגת (test).

ה) מה ההבדל מבחינת המשתמש בשימוש ב kernels שונים לצורך פתרון בעיית סיווג. למה הוא יעדיף kernel אחד על השני.

4) Adaboost (27 נקודות)

נתון dataset מסווג שכל נקודה שייכת לאחת משתי מחלקות. היית רוצה ללמוד מסווג adaboost כשכל מסווג בסיסי הוא עץ החלטה.

א) איך תבחר את המסווג הבסיסי הראשון.

ב) איך תבחר את המסווג הבסיסי השני. האם הוא זהה לראשון?

ג) האם ניתן להפעיל את ה adaboost כשיש יותר משתי מחלקות אפשריות? אם כן אז איך תשנה את האלגוריתם כדי שזה יהיה אפשרי?

ד) איך היית מממש אלגוריתם bagging במצב הזה כשיש רק שתי מחלקות.

ה) ענה על סעיפים א ב ג ו-ד במקרה שהמסווג הבסיסי הוא svm לינארי. הערה: בsvm כל נקודה שנמצאת בתוך הmargin או בצד הלא נכון שלו מוגדרת כטעות סיווג.