

ביה"ס למדעי-המחשב, אוניברסיטת ת"א  
סמסטר ב' תשע"ז  
מועד: א', 4/8/17  
משך הבחינה: שלוש שעות  
אין להשתמש בחומר עזר

## בחינה בקורס מבוא למדעי המידע

מרצה: פרופ' טובה מילוא  
מרצים אורחים: ד"ר זאב וקס, ד"ר לב פייבישבסקי, ד"ר אמתי ערמון, מר תומר לוי, גב' הגר לאוב  
מתרגל: מר סלבה נובוגרודוב

### הנחיות כלליות

- במבחן ישנן 14 שאלות:  
שאלות 1-10 הינן אמריקאיות (עם תשובה נכונה אחת בלבד),  
שאלות 11-14 הינן שאלות פתוחות.
- את כל הפתרונות יש לכתוב במחברת הבחינה, לרבות שאלות רב-ברירתיות.  
בטופס הבחינה ניתן להשתמש כטיוטה בלבד, והוא לא יבדק.
- אין להשתמש בחומר-עזר.
- מותר להשתמש במחשבון.
- הניקוד של כל שאלה מופיע בסמוך לה. סעיפי כל שאלה הם שווי-משקל.
- מומלץ לעבור על כל המבחן בעיון לפני שמתחילים לענות.
- בדקו בסיום הבחינה כי עניתם על כל השאלות וכל הסעיפים.
- הקפידו לסמן טיוטות שרשמתם במחברת כטיוטה (בראש הדף). לכל שאלה תיבדק אך ורק התשובה הראשונה שתופיע במחברת הבחינה.
- נא לכתוב בכתב קריא וברור.

בהצלחה!

לשימוש הבודקים:

	תעודת זהות
	מספר מחברת
	ציון מבחן

**שאלה 1 (5 נקודות)**

מבין המשפטים הבאים על שיטות k-Means, איזה משפט לא נכון:  
תזכורת: EM - Expectation-maximization

- א. בתהליך האימון של k-Means סכום המרחקים למרכזים לא גדל באף איטרציה
- ב. Soft k-Means הוא מקרה פרטי של EM
- ג. האיתחול ההתחלתי של k-Means אינו משפיע על טיב התוצאה
- ד. k-Means הוא אלגוריתם clustering מסוג Unsupervised Learning

**שאלה 2 (5 נקודות)**

נוסחת בייס קובעת את הקשר הבא:

$$\text{Posterior} = \frac{\text{Prior} \cdot \text{Likelihood}}{\text{Evidence}}$$

לפי הנחת האי-תלות בשיטת Naïve Bayes, איזה ערך אפשר לפרק לגורמים שתלויים רק במשתנה אחד?

- א. Likelihood
- ב. Evidence
- ג. Posterior
- ד. Prior

**שאלה 3 (5 נקודות)**

איזה מבין המשפטים הבאים לא נכון:  
תזכורת: ROC - receiver operating characteristic

- א. עקומת ROC מקשרת בין True positive rate לבין False positive rate
- ב. עקומת ROC של מסווג מושלם היא קו ישר שמחבר את הנקודה (0,0) לנקודה (1,1)
- ג. AUC הוא השטח מתחת לעקומת ROC
- ד. AUC שווה להסתברות שהמסווג ידרג דוגמה חיובית מעל דוגמה שלילית

**שאלה 4 (5 נקודות)**

איזה מבין המסווגים הבאים ניתן לממש עם זמן אימון הכי קצר?

- א. Decision Tree
- ב. SVM
- ג. KNN
- ד. Naïve Bayes

**שאלה 5 (5 נקודות)**

מבין השיטות הבאות של זיהוי שינויים/אנומליות, איזו שיטה מניחה התפלגות מסויימת של הנתונים?

- א. Mahalanobis Distance
- ב. One class SVM
- ג. Kolmogorov Smirnov
- ד. KNN-based KL

**שאלה 6 (5 נקודות)**

בפרוייקט לאבחון מחלה גנטית יש מידע על 1000 איש שחלו במחלה ו-1000 איש שלא חלו במחלה. עבור כל אדם נתונים 10,000 מאפיינים (features) גנטיים, אולם לא ידוע איזה מהם קשורים למחלה. יש לפתח מודל לגילוי המחלה על בסיס המאפיינים האלה. איזה מההצעות הבאות איננה הצעה שמומלץ לנסות?

- א. ננסה להוריד את כמות ה-features בעזרת שיטת PCA
- ב. נבחן שימוש במגוון מודלים, כולל Naïve Bayes, SVM ו Random forest
- ג. נבחר 5 מאפיינים שיש להם את הקורלציה הגבוהה ביותר לתשובה, ונבצע את המידול לפיהם
- ד. נפריד 400 איש מההתחלה לצורך בדיקת איכות המודל, ונאמן את המודל רק על שאר ה-1600

**שאלה 7 (5 נקודות)**

לבעיית סיווג קיבלתם נתונים שיש להם 1000 מאפיינים (features), וב-10 מבין המאפיינים יש ערכים חסרים. איזה מהמשפטים הבאים נכון?

- א. תמיד כדאי למלא במקומות חסרים את הממוצע של ערכי ה-feature
- ב. כשרק ב-1% מה-features יש ערכים חסרים, מומלץ להתעלם מה-features האלה
- ג. אם יש מספר קטן של דוגמאות עם ערכים חסרים, והן לא מוטות מבחינת ה-label, אז מומלץ להתעלם מהן
- ד. מומלץ לשים במקום כל הערכים החסרים אפסים

**שאלה 8 (5 נקודות)**

בסיווג בינארי כאשר יש לנו מחלקות לא מאוזנות (unbalanced), איזה משפט אינו נכון:

- א. אם הדיוק (accuracy) של המודל שאימנו הוא גבוה מ-99%, אז השגנו מודל טוב
- ב. שיטה נפוצה לטפל במקרים כאלו היא על ידי random downsampling של המחלקה הגדולה
- ג. שיטה לטיפול בבעיה היא לשלב upsampling של המחלקה הקטנה עם downsampling של המחלקה הגדולה
- ד. שיטה טובה לבעיה כזו היא לבנות אוסף מודלים שכל אחד מהם משתמש בחלקים שונים של המחלקה הגדולה ולאחד בסוף את תוצאותיהם

**שאלה 9 (5 נקודות)**

אנו בונים מודל סיווג על בסיס 25 דוגמאות. ידוע שחלק מהמאפיינים (features) מתפלגים נורמלית, אבל לא בהכרח כולם. להלן 25 הערכים של אחד המאפיינים שנשתמש בהם עבור בניית המודל. [-2, -1, -1, 0, 1, 1, 2, 2.3, 3.9, 4, 4, 4, 4.3, 8, 12, 14, 16, 18, 40, 50, 80, 200, 500, 1000, 40000] איזה משפט נכון?

- א. כיוון שיש ערכים שליליים, כדאי לבצע לפני המידול טרנספורמציה לינארית שתהפוך את כל הערכים לחיוביים
- ב. רצוי לבצע עיגול (rounding) למספר השלם הקרוב ביותר, כי רק מיעוט מהערכים לא שלמים
- ג. אם מבצעים טרנספורמציה על ערכים של feature מסויים, יש לבצע אותה טרנספורמציה על כל feature
- ד. כיוון שההתפלגות נוטה למספרים נמוכים, אך כוללת מספרים גבוהים בכמה סדרי גודל, טרנספורמציה לוגריתמית יכולה לעזור לחלק מהמודלים

**שאלה 10 (5 נקודות)**

איזה מהמשפטים הבאים לגבי mutual information אינו נכון:

- א. זהו מדד של אסוציאציה בין משתנים מקריים
- ב. זהו מדד שיכול לזהות קשרים ליניאריים וגם קשרים לא ליניאריים
- ג. זהו מדד שמכמת את הירידה באי הודאות לגבי הערכים של משתנה מקרי אחד כאשר ידוע הערך של המשתנה המקרי השני
- ד. זהו מדד שניתן לחישוב לפי האנטרופיה השולית של שני משתנים מקריים, בלי צורך בידיעת ההתפלגות המשותפת שלהם

**שאלה 11 (10 נקודות)**

על השולחן מונחים 10 קלפים. 2 שחקנים משחקים את המשחק, שבו כל שחקן בתורו לוקח 2, 3 או 5 קלפים. שחקן שלא יכול לעשות מהלך בתורו (כלומר אין לפחות 2 קלפים על השולחן) - מפסיד.

- תארו את האסטרטגיה המנצחת של השחקן שמתחיל
- תארו דרך לבנות מערכת לומדת. המערכת תקבל כקלט את החוקים של המשחק ותעבור אימון.
- אנחנו מתחילים סדרה של משחקי אימון מול יריב אופטימלי שיודע את האסטרטגיה המנצחת. אחרי כמה משחקי אימון שונים המערכת תלמד מספיק טוב כך תוכל לנצח כל יריב? נמקו את תשובתכם.
- האם תשובתכם לסעיף ג' תשתנה אם עושים את האימון מול שחקן רנדומי? נמקו את תשובתכם.

**שאלה 12 (20 נקודות)**

נתונות 3 טבלאות:

- רשימת הסטודנטים במדעי המחשב (students) שכוללת את ת.ז. הסטודנטית (s\_id), השם (s\_name) והשנה בלימודים (s\_year)
- רשימת הקורסים (courses) שכוללת את מספר הקורס (c\_id), שם הקורס (c\_name) והשנה שבה מומלץ ללמוד את הקורס (c\_year)
- רשימת הסטודנטים שרשומים לקורס מסוים (registered), שכוללת ת.ז. הסטודנטית (r\_s\_id), מספר הקורס (r\_c\_id) ותאריך ההרשמה (r\_date)

טבלאות לדוגמא:

**students**

s_id	s_name	s_year
111111	Alice	2nd
111112	Bob	2nd
111113	Charlie	3rd

**courses**

c_id	c_name	c_year
10001	Calculus 1	1st
20001	Algorithms	2nd
20002	Data Structures	2nd

**registered**

r_s_id	r_c_id	r_date
111111	20001	2017-02-18
111112	20001	2017-02-22

תארו במילים מה עושה כל אחת מהשאליות SQL הבאות וכיתבו תוכניות MAP-REDUCE (פסאודו קוד) עם מספר מינימלי של פונקציות Map ו Reduce שמחשבות אותן:

א) **SELECT \* FROM students, registered WHERE s\_year = '2nd' AND s\_id = r\_s\_id**

ב) **SELECT r\_c\_id, COUNT(r\_s\_id) FROM registered GROUP BY r\_c\_id**

**שאלה 13 (10 נקודות)**

נתונה רשת חברתית שמיוצגת על ידי גרף חברויות לא מכוון. כיתבו תוכנית MAP-REDUCE (פסאודו-קוד) עם מספר מינימלי של פונקציות Map ו Reduce, שמוצאת את כל המשולשים ברשת, ללא חזרות (a,b,c) ו (b,a,c) נחשב כחזרה). משולש זאת קבוצה בת 3 אנשים שבה כולם חברים של כולם. הקלט בנוי בשורות כאשר בכל שורה מצוין קודקוד בגרף, סימן של חץ, ורשימה (מופרדת ברווחים) של הקודקודים שמיצגים חברים שלו.

דוגמא לקלט:

A => B C F  
B => A  
C => A D  
D => C E F  
E => D F  
F => A D E

פלט עבור הקלט לדוגמא:

(D, E, F)

**שאלה 14 (10 נקודות)**

א. ציינו 3 יתרונות של SPARK על HADOOP ופרטו כל אחד מהיתרונות ב-2-3 משפטים.

ב. ציינו 2 מאפיינים של HDFS שמאפשרים לו להתמודד עם כמויות גדולות של נתונים בצורה מהירה ובטוחה (מוגנת מקריסות). פרטו את תשובתכם ב 2-3 משפטים.