

ביה"ס למדעי-המחשב, אוניברסיטת ת"א  
סמסטר ב' תשע"ז  
מועד: ב', 25/9/17  
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)**

- איזה מהמשפטים הבאים נכון לגבי NAÏVE BAYES CLASSIFIER? ביחרו תשובה אחת בלבד:
- המסווג מסוגל לסווג נתונים בעלי מימד גבוה, כאשר שערך ההתפלגות המשותפת הרב-מימדית של הנתונים אינו אפשרי.
  - המסווג מחייב המרה של נתונים רציפים לייצוג דיסקרטי על מנת לחשב את הנראות (LIKELIHOOD) של המודל.
  - המסווג מחייב המרה של נתונים רציפים לנתונים דיסקרטיים, אלא אם כן הנתונים מפולגים אחיד במקטע  $[0, 1]$ .
  - המסווג לא מסוגל לפעול על נתונים שכוללים גם ערכים דיסקרטיים וגם ערכים רציפים.

**שאלה 2 (5 נקודות)**

- איזה משפט נכון לגבי מסווג KNN? ביחרו תשובה אחת בלבד:
- ככל ש-K גדול יותר כך יש סכנה גדולה יותר של Overfitting במידול
  - כאשר K קטן נוצרת בהכרח תופעה של Underfitting במידול
  - ככל ש-K גדול יותר כך גבול ההחלטה (DECISION BOUNDARY) יהיה חלק יותר
  - ככל ש-K קטן יותר כך זמן האימון של KNN יהיה גדול יותר

**שאלה 3 (5 נקודות)**

- איזה משפט אינו נכון לגבי מסווג SVM? ביחרו תשובה אחת בלבד:
- Hard margin SVM מנסה להפריד נקודות ממחלקות שונות בצורה מושלמת, ואילו Soft margin SVM מאפשר לנקודות בתהליך האימון ליפול בצד הלא נכון של משטח ההפרדה
  - Kernel SVM ההפרדה במרחב הסופי (אחרי ההטלה מהמרחב המקורי) נעשית על ידי מישור
  - בשיטת Hard margin SVM מתקבל שיוך בינארי של כל נקודה לאחת המחלקות בדיוק, ואילו בשיטת Soft margin SVM מתקבלות הסתברויות לשיוך כל נקודה לכל מחלקה
  - אפשר להשתמש ב-Kernel גם ב-Hard margin SVM וגם ב-Soft margin SVM

**שאלה 4 (5 נקודות)**

- איזה משפט נכון לגבי K-Means? ביחרו תשובה אחת בלבד:
- באימון של K-Means יש תמיד פיתרון גלובאלי, כלומר לאחר אימון מספיק ממושך עם אתחול אקראי תמיד יתקבל אותו שיוך נקודות ל Clusters.
  - באימון K-Means פונקציית המטרה יכולה גם לגדול וגם לקטון מאיטרציה לאיטרציה, כיוון שהאתחול אקראי.
  - ככל ש-K גדול יותר, כך זמן הריצה של K-Means יהיה בהכרח קטן יותר.
  - באימון K-Means יש לקבוע את K לפני האימון, על מנת לחשב את פונקציית המטרה.

**שאלה 5 (5 נקודות)**

- איזה משפט אינו נכון לגבי שיטות Change Detection? ביחרו תשובה אחת בלבד:
- שיטת Kolmogorov-Smirnov עובדת רק עם נתונים רציפים, ולא עובדת עם נתונים דיסקרטיים.
  - שיטת KNN KL עובדת רק עם נתונים רב-מימדיים, ולא עובדת עם נתונים חד-מימדיים.
  - שיטת Kolmogorov-Smirnov עובדת רק עם נתונים חד-מימדיים, ולא רב-מימדיים.
  - שיטת KNN KL מאפשרת חישוב קרבה בין התפלגויות נתונים רב-מימדיים, מבלי לשערך התפלגויות שוליות רב-מימדיות.

**שאלה 6 (5 נקודות)**

איזה מהמשפטים לגבי זיהוי ערכים חריגים (outlier detection) אינו נכון? ביחרו תשובה אחת בלבד:

- ההשפעה של ערכים חריגים על איכות התחזית יכולה להיות שונה בין שיטות מידול שונות.
- אם מזהים חריגים בשיטות של אמידת פונקציית צפיפות, אז אין צורך לבדוק חריגים לפי feature בודד.
- טרנספורמציה לוגריתמית יכולה לצמצם את השפעתם של ערכים חריגים.
- שיטות לזיהוי ערכים חריגים לרוב אינן משמשות כשיטות לצורך הורדת מימד.

**שאלה 7 (5 נקודות)**

איזה משפט לגבי סוגי מאפיינים (features) אינו נכון? ביחרו תשובה אחת בלבד:

- משתנה אורדינלי יכול להיות תוצר של דיסקרטיזציה של משתנה רציף.
- כמות הקטעים שמחלקים אליה איננה המאפיין היחיד של שיטת הדיסקרטיזציה של משתנה רציף.
- אסור להפוך משתנה קטגוריאל ללאוסף משתנים בינאריים משום שהקשרים בין הערכים לא ישמרו.
- שיטות נפוצות לבחירת גודל הקטעים בדיסקרטיזציה של משתנה רציף כוללות חלוקה לפי גודל זהה, חלוקה לפי כמות דוגמאות זהה, וחלוקה לפי maximal information gain.

**שאלה 8 (5 נקודות)**

איזה משפט לגבי מדדים להערכת מודלים אינו נכון? ביחרו תשובה אחת בלבד:

- נהוג להשתמש בממד F1 על מנת לשקלל בין precision ו- recall.
- precision גבוה לא בהכרח מעיד על מודל טוב.
- תמיד ניתן לבנות מודל שיתן recall גבוה.
- precision ו- recall הם מדדים נפוצים שמתאימים גם לבעיות סיווג וגם לבעיות רגרסיה.

**שאלה 9 (5 נקודות)**

איזה משפט לגבי אופן חלוקת דוגמאות לצורך הערכה (evaluation) נכון? ביחרו תשובה אחת בלבד:

- k-fold cross validation אין משמעות לאופן החלוקה למקטעים (folds) השונים, כל עוד יש כמות זהה של דוגמאות בכל מקטע.
- בחלוקה לשלושה מקטעים ב- hold out evaluation, נהוג לחלק את הדוגמאות באופן אקראי ל- training set, validation set, ו- test set.
- ב- leave one out cross validation אין חלוקה אקראית של הדוגמאות אבל יש משמעות לסדר החלוקה והחישוב.
- בחלוקה לשלושה מקטעים ב- hold out evaluation, רצוי שה- validation set יהיה לפחות כפול בגודלו מה- training set.

**שאלה 10 (5 נקודות)**

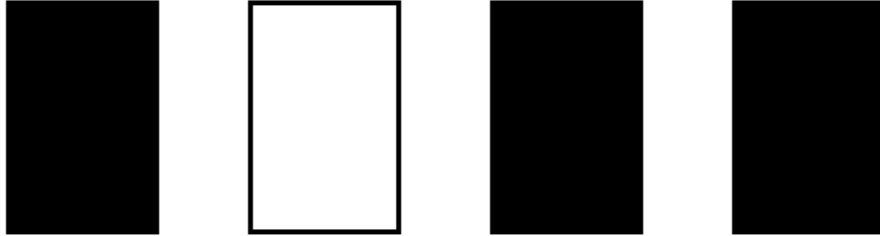
התבקשתם לבנות מסווג בינארי המבדיל בין תקין לפגום. לרשותכם 1000 דוגמאות, מחציתן תקינות ומחציתן פגומות. ידוע כי לא קיימים מאפיינים בודדים המאפשרים סיווג עם תוצאות מספקות, אך ישנם שילובים מסויימים של מאפיינים המאפשרים סיווג עם תוצאות טובות מאוד. איזו שיטת feature selection הכי פחות מתאימה במקרה זה? ביחרו תשובה אחת בלבד.

- נבצע feature selection בעזרת wrapper method.
- נבחר את המאפיינים בעלי הקורלציה הגבוהה ביותר עם התיוג (פגום/תקין).
- נחשב את הקורלציה בין כל זוג מאפיינים. כשהקורלציה בין זוג מסויים גבוהה מאוד, נוותר על אחד המאפיינים בזוג הזה.
- נבצע sequential backward selection בעזרת מודל כלשהו.

**שאלה 11 (10 נקודות)**

על השולחן מונחים 4 קלפים בעלי צד אחד שחור וצד אחד לבן. הקלפים מונחים עם הצד הלבן כלפי מעלה. 2 שחקנים משחקים את המשחק, השחקן הראשון ראשי להפוך קלף אחד בכל מהלך והשחקן השני ראשי להפוך 2 קלפים **צמודים** בכל צעד. השחקן הראשון מנצח אם אחרי לכל היותר 3 מהלכים שלו הוא מצליח להגיע למצב שיש 3 קלפים עם הצד השחור כלפי מעלה. השחקן השני מנצח בכל סיטואציה אחרת.

**דוגמא למצב מנצח:**



- א. תארו את האסטרטגיה המנצחת של השחקן שמתחיל
- ב. תארו דרך לבנות מערכת לומדת. המערכת תקבל כקלט את החוקים של המשחק ותעבור אימון.
- ג. אנחנו מתחילים סדרה של משחקי אימון מול יריב אופטימלי שיודע את האסטרטגיה המנצחת. אחרי כמה הפסדים המערכת תלמד מספיק טוב כך תוכל לנצח כל יריב? נמקו את תשובתכם.

**שאלה 12 (20 נקודות)**

נתונות 3 טבלאות:

1. רשימת ספורטאיות אולימפיות (athletes) שכוללת את מספר המזהה של ספורטאית (a\_id), השם (a\_name) והגיל (a\_age)
2. רשימת המדליות שהספורטאיות שזכו (medals) שכוללת את מספר המזהה של הספורטאית (m\_id), סוג המקצה (m\_type) וצבע המדליה (m\_color)
3. רשימת המדינות עליהן שייכות הספורטאיות (countries), שכוללת את המספר המזהה של הספורטאית (c\_id) את המדינה (c\_name) והיבשת (c\_continent)

טבלאות לדוגמא:

**athletes**

a_id	a_name	a_age
1111	Elaine Thompson	25
1112	Vivian Cheruiyot	34
1113	Almaz Ayana	25

**medals**

m_id	m_type	m_color
1111	100 meters	Gold
1111	200 meters	Gold
1112	10,000 meters	Silver

**countries**

c_id	c_name	c_continent
1111	Jamaica	North America
1112	Kenya	Africa

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

א)

```
SELECT * FROM athletes, medals
WHERE a_age < 30 AND m_color = 'Gold' AND m_id = a_id
```

ב)

```
SELECT c_name, COUNT(m_color) FROM countries, medals
WHERE m_color='Silver' AND m_id = c_id GROUP BY c_name
```

**שאלה 13 (10 נקודות)**

נתונה רשימת אתרי אינטרנט וההצבעות בין האתרים (לינקים). כיתבו תוכנית MAP-REDUCE (פסאודו-קוד) עם מספר מינימלי של פונקציות Map ו Reduce, שמוצאת עבור כל אתר את רשימת האתרים שאיתם יש לו הצבעה הדדית (כלומר A מצביע על B ו B מצביע על A).  
 הקלט בנוי בשורות כאשר בכל שורה מצוין אתר, סימן של חץ, ורשימה (מופרדת ברווחים) של אתרים שעליהם הוא מצביע.

דוגמא לקלט:

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

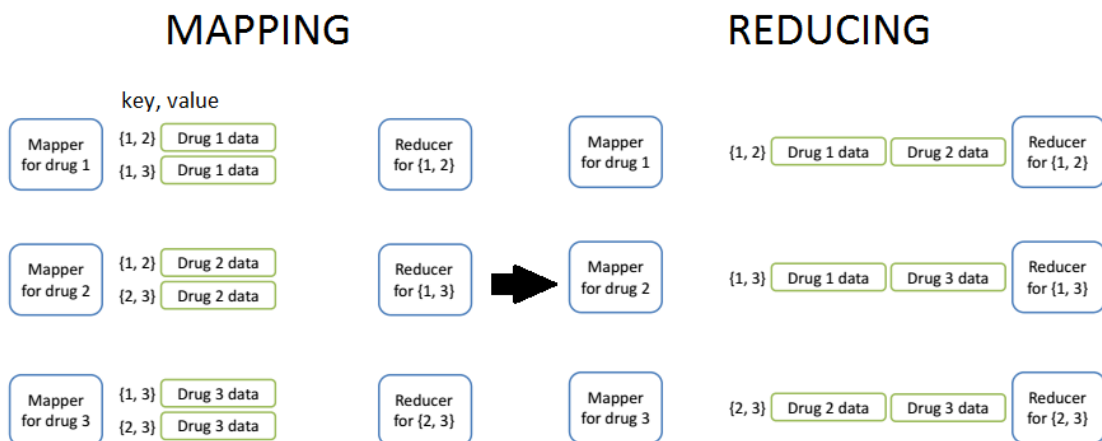
פלט עבור הקלט לדוגמא:

A.com => C.com  
 C.com => A.com D.com  
 D.com => C.com

**שאלה 14 (10 נקודות)**

נתונים 100 סוגי תרופות שונים, כאשר עבור כל סוג של תרופה יש כ מגהבייט אחד (1MB) של מידע רפואי הקשור לתרופה, כולל נתונים של אי-תאימות לתרופות אחרות.  
 הבעיה היא למצוא, עבור כל זוג של תרופות, האם הן מתאימות להילקח ביחד או לא.  
 לצורך השאלה, נניח שכל REDUCER מטפל רק במפתח אחד.  
 נתונות 2 שיטות חלופיות לפתרון מבוססות Map-Reduce:

1. כל MAPPER יקבל נתונים של תרופה אחת וישלח אותה 99 פעמים. כאשר כל מפתח מכיל את מספר התרופה ומספר של MAPPER שונה ממספר התרופה, כששני הערכים האלה מסודרים לקסיקוגרפית. הדגמה של פעולת המערכת עבור 3 תרופות:



2. בשיטה השנייה התרופות מחולקות ל 10 קבוצות (ממוספרות מ 1 עד 10) כל אחת כוללת 10 תרופות, וכל REDUCER מטפל בזוג קבוצות מסויים (במקום בזוג של תרופות).

- א. כמה REDUCERS נצטרך בשיטה הראשונה ובשיטה השנייה?
- ב. ציינו יתרון וחסרון של כל אחת של השיטות לעומת השיטה השנייה.