מס' מחברת : מס' ת.ז. :

ביה"ס למדעי-המחשב, אוניברסיטת ת"א
סמסטר ב' תשע"ח
מועד: ב', 7/08/2018
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 אחת נכונה 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	

מס' מחברת : מס' ת.ז. :

שאלה 1 (5 נקודות):
בתהליך חיפוש אחר מודל מיטבי לבעיה מסויימת, חילקנו את סט הנתונים לסט אימון (train set) וסט בחינה (test set). מי מבין התרחישים הבאים עשוי להעיד על כך שהמודל סובל מתת התאמה (underfitting)?

א. השגיאה ב test גדלה אם משתמשים רק בחצי מה-features עבור האימון.
ב. אם מאמנים את המודל רק עם חצי מהדוגמאות ב- training, השגיאה בtest-לא גדלה.
ג. השגיאה ב training גדלה אם משתמשים רק בחצי מה- features עבור האימון.
ד. אף תשובה לא נכונה.

שאלה 2 (5 נקודות):
כאשר אנו מנסים לבנות מודל, בדרך כלל חשוב לבדוק את התפלגות המאפיינים בסט הנתונים. באיזה מהסיטואציות הבאות **אין צורך** לעשות בדיקה כזו?

א. אנו רוצים לדעת אם יש בנתונים ערכים חסרים (missing values).
ב. אנו בונים מודל של רגרסיה ליניארית.
ג. אנו רוצים לפתור בעיית קליסטור (clustering).
ד. אנו בונים מודל סיווג מסוג רגרסיה לוגיסטית.

שאלה 3 (5 נקודות):
בנינו מודל סיווג בינארי, בדקנו את המודל על 100 דוגמאות וקיבלנו את ה-confusion matrix הבאה. מי מבין המשפטים הבאים **אינו נכון**?

		ניבוי המודל	
		שלילי	חיובי
התיוג האמיתי	שלילי	20	30
	חיובי	10	40

א. ערך ה- precision גדול מ- 0.5
ב. ערך ה- recall שווה 0.8
ג. ערך ה- accuracy שווה 0.6
ד. ערך ה- f1 measure שווה 1.37

שאלה 4 (5 נקודות):
במקרים רבים משתמשים במידול בתהליך בחירת מאפיינים(feature selection). באיזה מקרה מבין הבאים בחירת מאפיינים **אינה** אפשרות מקובלת?

א. אנחנו חוששים ממצב של overfitting.
ב. ידוע כי יש מאפיינים עם קורלציה ביניהם.
ג. אנחנו בונים מודל מסוג CNN בשביל לסווג תמונות.
ד. אנחנו מעוניינים להוריד את מימד הדאטה שלנו.

שאלה 5 (5 נקודות):
5. איזה מבין שיטות הסיווג הבאות יכולה לייצר גבול הפרדה לינארי למקוטעין?
א. Linear SVM
ב. Polynomial Kernel SVM
ג. Gaussian Kernel SVM
ד. Decision Tree

מס' מחברת : מס' ת.ז. :

שאלה 6 (5 נקודות):
איזה מבין השיטות הבאות יכולה לתת הסתברות לגילוי שינוי/אנומליה ללא אימון מוקדם על דאטה נפרד?

א. ONE CLASS SVM
ב. NAÏVE BAYES
ג. Change Detection by Kolmogorov Smirnov Test
ד. Change Detection by Kullback Leibler Distance

שאלה 7 (5 נקודות):
איזה מבין המשפטים הבאים מבטא **חיסרון** של סיווג בעזרת אלגוריתם KNN ?

א. KNN נותן תמיד גבול החלטה לינארי
ב. שימוש ב-KNN דורש שמירת מידע רב בזיכרון
ג. המידול באמצעות KNN הוא תמיד underfitted, לא משנה מה ערכו של K
ד. שימוש ב- KNN דורש זמן ריצה ארוך מאוד עבור אימון המודל

שאלה 8 (5 נקודות):
איזה מהמשפטים הבאים **נכון** לגבי סיווג בעזרת אלגוריתם SVM?

א. ניתן להשתמש ב- SVM רק כאשר יש הפרדה מלאה בין ה- classes
ב. עבור נתונים רציפים, SVM תמיד נותן תוצאות טובות יותר מאשר Decision Tree
ג. בשימוש ב- SVM אין צורך לחלק את הנתונים ל- train ו- test
ד. תשובות א'-ג' אינן נכונות

שאלה 9 (5 נקודות):
איזה מהמשפטים הבאים **אינו נכון** לגבי רשתות נוירונים:

א. פונקציית האקטיבציה של הנוירונים חייבת להיות לינארית
ב. אין הגבלה על מספר השכבות שיכולות להיות ברשת
ג. כל נוירון חייב להיות מחובר לנוירון אחר כלשהו ברשת
ד. אין הגבלה על מספר הנוירונים שיכולים להיות בשכבה ברשת

שאלה 10 (5 נקודות):
איזה מהמשפטים הבאים **נכון** לגבי אלגוריתם k-Means ואלגוריתם GMM?

א. GMM לא מניח הנחה על אופן התפלגות הנתונים, בעוד ש- k-Means כן מניח הנחה כזו
ב. k-Means משייך כל דגימה ל-cluster ספציפי, בעוד ש- GMM נותן ציון התאמה לכל cluster
ג. ב- k-Means צריך לבחור את מספר ה- clusters שיבנו, בעוד שב-GMM לא בוחרים את מספרם
ד. ב-k-Means לא מובטח שהאלגוריתם יתכנס, בעוד שב- GMM ההתכנסות מובטחת

שאלה 11 (10 נקודות)

על השולחן מונחות 6 מטבעות זהות, ב3 שורות: שורה של מטבע אחד, שורה של שתי מטבעות ושורה של 3 מטבעות. 2 שחקנים משחקים משחק, שבו כל שחקן בתורו לוקח כמות זהה של מטבעות **משתי שורות שונות** (לדוגמא: 2 מטבעות מהשורה השניה ו2 מטבעות מהשורה השלישית). שחקן שלא יכול לעשות מהלך בתורו - **מנצח במשחק**.

א. תארו את האסטרטגיה המנצחת של השחקן המתחיל

ב. תארו דרך לבנות מערכת לומדת שלומדת את האסטרטגיה המנצחת של השחקן המתחיל. המערכת תקבל כקלט את החוקים של המשחק ותעבור אימון.

ג. אנחנו מתחילים סדרה של משחקי אימון (שבהם אנחנו השחקן המתחיל) מול יריב אופטימלי שיודע את האסטרטגיה המנצחת. אחרי כמה משחקי אימון שנפסיד בהם המערכת תלמד מספיק טוב כך תוכל לנצח כל יריב? נמקו את תשובתכם.

ד. האם תשובתכם לסעיף ג' תשתנה אם עושים את האימון מול שחקן רנדומי? נמקו את תשובתכם.

שאלה 12 (20 נקודות)

נתונות 3 טבלאות שמייצגות את הנתונים באתר של הזמנת חופשות:

1. רשימת בתי מלון (hotels) שכוללת את שם המלון (h_name), והעיר של המלון (h_city)
2. רשימת הטיסות (flights) שכוללת את מספר הסידורי של הטיסה (f_id), שם עיר מקור (f_from), ושם עיר היעד (f_to).
3. רשימת הזמנות של הטיסות (orders), שכוללת מזהה ההשמנה (o_id), מזהה של הטיסה (f_id) והמחיר ששולם (price)

טבלאות לדוגמא:

hotels

h_name	**h_city**
Novotel	London
David Intercontinental	Tel Aviv

flights

f_id	**f_from**	**f_to**
U2 4784	Berlin	Tel Aviv
LH 2053	Berlin	London

orders

o_id	**f_id**	**price**
771823	U2 4782	275
981723	LH 2053	208
196630	LH 2053	255

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

a) SELECT h_name **FROM** hotels, flights **WHERE** f_from = 'Berlin' **AND** f_to = h_city

b) SELECT f_id, **SUM** (price) **FROM** orders **GROUP BY** f_id

שאלה 13 (10 נקודות)

נתונה רשימה של נתונים על גברים, מופרדת בפסיק. המידע מחולק באופן הבא:
שם משתמש, גיל, מדינת מגורים, כמות ילדים, מעשן או לא
חברת מחקר סוציולוגית רוצה לבדוק האם יש קשר בין גיל וכמות ילדים של גבר לכך שהוא מעשן או לא מעשן.
הסטטיסטיקה סופית היא גם לפי המדינות. הוחלט לחלק את האוכלוסיה ל2 קטגוריות גיל: 30+ ופחות מ30, ו3 קטגוריות לפי כמות ילדים: 0, 1-2 ו3 או יותר.
במוסף, הוחלט שאם בקטגוריה מסויימת יש פחות מ3 אנשים, להחזיר שאין מספיק מידע.

כיתבו תוכנית Map-Reduce (פסאודו-קוד) עם מספר מינימלי של פונקציות Map ו Reduce, שעוזרת לחשב את הסטטיסטיקה הדרושה. הסבירו בקצרה (3-4 שורות לכל היותר) מה כל פונקציה עושה.

דוגמא לקלט:

andrew, 33, italy, 0, yes
jack, 39, usa, 3, yes
slava, 28, israel, 3, no
pablo diego juan antonio, 26, spain, 2, yes
rodrigo, 55, spain, 2, no
...

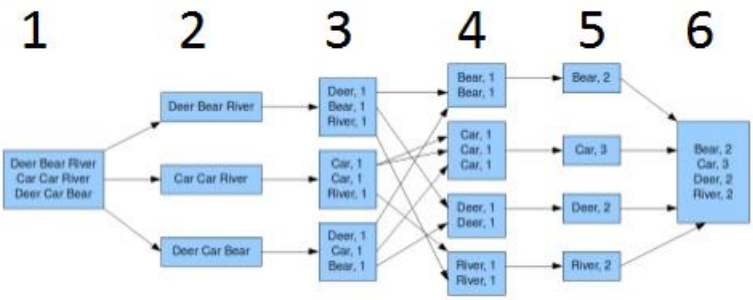
פלט עבור הקלט לדוגמא:

spain, age <30, 1-2 kids: 55% yes, 45% no
spain, age >30, 0 kids: 11% yes, 89% no
spain, age >30, 0 kids: not enough data

...

שאלה 14 (10 נקודות)

א. (4 נקודות) נתון תרשים של הפעלת WordCount בשיטת MapReduce.



ציינו עבור כל מספר שלב את השם שלו מתוך רשימה הבאה:
Output, Input, Shuffle, Reduce, Map, Split

ב. **(6 נקודות) סמנו נכון/לא נכון על המשפטים הבאים:**

1. הגדלת כמות הMAPPERS והקטנת כמות הREDUCERS תמיד תגרום לפתרון יותר איטי של המשימה (נכון / לא נכון)

2. למפתח שיוצא אחרי פעולת הMAP אין השפעה על המשך החישוב (נכון / לא נכון)

3. הREDUCERS יכולים לדבר אחד עם השני לצורך חלוקה יותר טובה של עבודה(נכון / לא נכון)