מס׳ מחברת : מס׳ ת.ז. :

ביה״ס למדעי-המחשב, אוניברסיטת ת״א
סמסטר ב׳ תשע״ו
מועד : א׳, 23/6/16
משך הבחינה : שלוש שעות
אין להשתמש בחומר עזר

בחינה בקורס מבוא למדעי המידע

מרצה הקורס : אסף עראקי
מרצים נוספים : ערן בולס, ג׳רמי דרייפוס, נעמה פיקסלר, ערן אבידן, עמית מור, ד״ר שחר כהן, ד״ר לב פייבישבסקי, ד״ר אמתי ערמון

הנחיות כלליות

- 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)

על איזה אתגר מיועדת לענות הנחת האי-תלות במודל Naïve Bayes Classifier? ביחרו את התשובה הנכונה.

א. בחישוב ההסתברויות המשותפות יש חשש שההסתברויות לא ייצגו את הבעיה בעולם האמיתי.

ב. יש קושי לחשב הסתברות משותפת של משתנים מסוגים שונים.

ג. יש צורך לחשב ולשמור את ההסתברויות לכל הקומבינציות האפשריות של ערכי המשתנים בדאטה.

ד. לא ניתן לחשב הסתברויות משותפות כאשר יש ערכים חסרים בדאטה.

שאלה 2 (4 נקודות)

איזה מהמשפטים הבאים אינו נכון לגבי מודל K-Nearest Neighbors? (ביחרו תשובה אחת בלבד)

א. במודל KNN, ככל שנבחר K קטן יותר, כך השכונה תהיה יותר הומוגנית.

ב. במודל KNN, ככל שנבחר K גדול יותר, כך השונות תהיה קטנה יותר.

ב. במודל KNN ניתן להשתמש בדטה עם משתנים קטגוריאליים.

ג. במודל KNN נרצה להשתמש בכמה שיותר משתנים כדי לקבל מודל סיווג טוב יותר.

שאלה 3 (4 נקודות):

במודל רגרסיה לוגיסטית: (ביחרו תשובה אחת בלבד)

א. המודל הלינארי משמש לשיערוך **ערכי המשתנה הבינארי** המיוצגים על ידי 0 ו-1.

ב. המודל הלינארי משמש לשיערוך **ההסתברות** לאחד מערכי המשתנה הבינארי.

ג. המודל הלינארי משמש לשיערוך **היחס בין ההסתברויות** לכל אחד מערכי המשתנה הבינארי.

ד. המודל הלינארי משמש לשיערוך **לוג היחס בין ההסתברויות** לכל אחד מערכי המשתנה הבינארי.

שאלה 4 (4 נקודות)

עבור כל סעיף, ביחרו נכון/לא נכון:

א. במודל Decision Tree, בכל אחד מהפיצולים ישתתפו בהכרח פחות תצפיות מאשר בצומת שמעליו. **נכון/לא נכון**

ב. במודל Decision Tree, המשתנה שייבחר בכל פיצול הוא זה שממקסם את מדד ה information gain. **נכון/לא נכון**

ג. במודל Decision Tree, מדד ה-information gain המתקבל עבור המשתנה הנבחר בכל צומת החלטה הוא בהכרח קטן יותר מאשר בצומת שמעליו. **נכון/לא נכון**

ד. במודל Decision Tree, המשתנה שייבחר לפיצול הראשון עשוי להשפיע גם על בחירת המשתנים במורד העץ. **נכון/לא נכון**

שאלה 5 (4 נקודות)

נתונים training set המכיל N דוגמאות ו- test set המכיל M דוגמאות. אלגוריתם SVM לינארי עם שוליים רכים (soft) אומן על ה- training set והשיג accuracy של A_{train}. לאחר מכן המודל הופעל על ה- test-set והשיג accuracy של A_{test}. בשלב הבא ערך הפרמטר C הוגדל, והייתה $A_{train} > A_{test}$ חזרה על התהליך, שנתנה ערכי דיוק B_{train}, B_{test}. איזה מהתשובות הבאות היא הסבירה ביותר? (ביחרו תשובה אחת בלבד)

א. $B_{train} \geq A_{train}$

ב. $B_{train} < A_{train}$

ג. $B_{test} < A_{test}$

ד. $B_{test} \geq A_{test}$

מס׳ מחברת : מס׳ ת.ז. :

שאלה 6 (5 נקודות)

מקלסטרים N אובייקטים $O_1, \ldots, O_N$ בשיטת k-Means. מריצים את האלגוריתם שתי הרצות נפרדות, עם מס' קלסטרים k_1 , k_2 , והשמה התחלתית l_1 , l_2 בהתאמה. מספר האיטרציות הוא זהה בשתי ההרצות. הציון הסופי בהרצה הראשונה גדול יותר מאשר בהרצה השניה. איזה מהתשובות הבאות **איננה יכולה להיות נכונה**? (ביחרו תשובה אחת בלבד)

א. $k_1=k_2$, $l_1 \neq l_2$

ב. $k_1 = k_2$, $l_1 = l_2$

א. $k_1 > k_2$, $l_1 \neq l_2$

ב. $k_1 < k_2$, $l_1 \neq l_2$

שאלה 7 (6 נקודות)

א. מהי בעיית Overfitting? הסבירו בקצרה.

ב. תארו בקצרה שיטה אחת שמאפשרת לזהות את קיום הבעיה הזו.

ג. כיצד ניתן להתגבר על הבעיה הזו? הזכירו בקצרה שתי שיטות.

שאלה 8 (6 נקודות)

א. תארו יתרון אחד וחיסרון אחד של חישוב Mutual Information כשיטת Feature selection.

ב. בשיטת Forward selection מוסיפים בכל פעם משתנה נוסף, ומחשבים את המודל מחדש. כיצד יודעים מתי לעצור? (להפסיק להוסיף משתנים).

שאלה 9 (6 נקודות)

שאלה זו מתייחסת למצב של Imbalanced data בבעיית קלסיפיקציה בינארית.

א. הסבירו בקצרה מהי הבעייתיות הנגרמת כאשר הנתונים הם לא מאוזנים.

ב. תארו בקצרה שתי שיטות לטיפול בנתונים לא מאוזנים.

שאלה 10 (6 נקודות)

נתונה להלן טבלה ובה סוגים שונים של זוגות משתנים. התאימו לכל זוג משתנים את סוג מבחן הקורלציה המתאים ביותר לבדיקת הקשר בין זוג המשתנים הזה. ביחרו מבין סוגי מבחני הקורלציה הבאים:

א. Pearson $r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}*\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

ב. Spearman $r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}*\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

(The x_i & y_i are the ranked features)

ג. Kendall's Tau : $\tau_{X,Y} = \frac{(\#\ of\ concordant\ pairs) - (\#\ of\ discordant\ pairs)}{\frac{1}{2}n(n-1)}$

סעיף	משתנה A	משתנה B	מבחן התאמה
א	רציף עם חשיבות לערך	רציף עם חשיבות לערך	
ב	אורדינלי	אורדינלי	
ג	רציף	אורדינלי	
ד	רציף עם חשיבות לסדר	רציף עם חשיבות לסדר	

מס׳ מחברת : מס׳ ת.ז. :

שאלה 11 (4 נקודות)

במה תומכת יכולת ה Lineage ב- SPARK ? (**בחרו את התשובה המתאימה ביותר**)

א. Partial Failure Support
ב. Data Recoverability
ג. Component Recovery
ד. Consistency

שאלה 12 (8 נקודות)

תכננו את האלגוריתם K-Means ב - Map-Reduce. לרשותכם שני קבצים. קובץ ראשון גדול ומכיל מליוני נקודות. קובץ שני קטן ומכיל איתחול של K נקודות מרכז.

א. מספר הפונקציות הנדרשות הוא _____ (בין 1 ל – 4)
ב. פונקציה ראשונה
   i. הפונקציה מסוג Map / Reduce (מחק את המיותר).
   ii. הפונקציה מחשבת _____
   iii. ה- Output Key של הפונקציה הוא _____
   iv. ה- Output Value של הפונקציה הוא _____
ג. פונקציה שנייה
   i. הפונקציה מסוג Map / Reduce (מחק את המיותר).
   ii. הפונקציה מחשבת _____
   iii. ה- Output Key של הפונקציה הוא _____
   iv. ה- Output Value של הפונקציה הוא _____
ד. פונקציה שלישית
   i. הפונקציה מסוג Map / Reduce (מחק את המיותר).
   ii. הפונקציה מחשבת _____
   iii. ה- Output Key של הפונקציה הוא _____
   iv. ה- Output Value של הפונקציה הוא _____
ה. פונקציה רביעית
   i. הפונקציה מסוג Map / Reduce (מחק את המיותר).
   ii. הפונקציה מחשבת _____
   iii. ה- Output Key של הפונקציה הוא _____
   iv. ה- Output Value של הפונקציה הוא _____

שאלה 13 (4 נקודות)

ההבדל בין Transformation ל- Action הינו (**בחרו את התשובה המתאימה ביותר**)

א. Transformation הינו אופרטור בעוד Action אינו אופקטור
ב. Transformation אינו אופרטור בעוד Action הינו אופקטור
ג. Transformation מבצע טרנספומציה של המידע בעוד Action אינו מבצע טרנספורמציה למידע
ד. Transformation מיצר DataSet בעוד Action מיצר Value

שאלה 14 (4 נקודות)

במה תומכת יכולת ה Lazy ב- SPARK ? (**בחרו את התשובה המתאימה ביותר**)

א. Partial Failure Support
ב. Efficiency
ג. Storage
ד. Consistency

מס׳ מחברת : _____ מס׳ ת.ז. : _____

שאלה 15 (5 נקודות)

איזה מאפיין של HDFS אחראי על Data Recovery?(**בחרו את התשובה המתאימה ביותר**)

א. חלוקת הקובץ לבלוקים בגודל קבוע של 128MB
ב. שיכפול הבלוקים לשלושה העתקים
ג. איזון מספר הבלוקים בין מספר ההמחשבים ב- Cluster
ד. תשובות ב ו-ג נכונות
ה. תשובות א ו-ג נכונות

שאלה 16 (4 נקודות)

פעולת ה JOIN הינה מורכבת מהפעולות MapReduce הבאות ?(**בחרו את התשובה המתאימה ביותר**)

א. פעולת Map בלבד
ב. פעולת Map ו- Reduce
ג. פעולת Reduce בלבד
ד. תשובות א' ו-ב' נכונות

שאלה 17 (8 נקודות)

יש לנו שני קבצים. הקובץ הראשון מכיל רשימת מצביעים בבחירות
(voter-id, name, age, zip, income, Kids_number)
הקובץ השני מכיל מידע על מחלות (zip, age, average_income)
נא חשבו את ממוצע הילדים של האנשים שהכנסתם גדולה מהממוצע לגילם ואיזור המגורים.

א. מספר הפונקציות הנדרשות הוא _____ (בין 1 ל – 4)
ב. פונקציה ראשונה
   v. הפונקציה מסוג Map / Reduce (מחק את המיותר).
   vi. הפונקציה מחשבת _____
   vii. ה- Output Key של הפונקציה הוא _____
   viii. ה- Output Value של הפונקציה הוא _____
ג. פונקציה שנייה
   ix. הפונקציה מסוג Map / Reduce (מחק את המיותר).
   x. הפונקציה מחשבת _____
   xi. ה- Output Key של הפונקציה הוא _____
   xii. ה- Output Value של הפונקציה הוא _____
ד. פונקציה שלישית
   xiii. הפונקציה מסוג Map / Reduce (מחק את המיותר).
   xiv. הפונקציה מחשבת _____
   xv. ה- Output Key של הפונקציה הוא _____
   xvi. ה- Output Value של הפונקציה הוא _____
ה. פונקציה רביעית
   xvii. הפונקציה מסוג Map / Reduce (מחק את המיותר).
   xviii. הפונקציה מחשבת _____
   xix. ה- Output Key של הפונקציה הוא _____
   xx. ה- Output Value של הפונקציה הוא _____

שאלה 18 (5 נקודות)

עבור כל משפט סמנו 'נכון' או 'לא נכון' עבור ההבדלים בין Spark Batch ו- Spark Streaming (**ציין נכון או לא נכון**)

א. מימוש Spark Batch ו- Spark Streamlining, כאשר מבוצע כהלכה חולק זהות לוגית וקוד רב בין המימושים. **נכון / לא נכון**

ב. מימוש Spark Batch ו- Spark Streamlining, כאשר מבוצע כהלכה שונה בעיקר בקוד ה - Driver. **נכון / לא נכון**

ג. Spark Streaming מעבד מידע מיד עם הגיעו ממקור המידע. **נכון / לא נכון**

ד. Spark Batch יכול לקרוא קבצים מה – HDFS בעוד Spark Streaming אינו יכול. **נכון / לא נכון**

ה. Spark Batch משתמש ב RDDs בזמן ש Spark Streaming לא משתמש ב – RDDs. **נכון / לא נכון**

שאלה 19 (4 נקודות)

כיצד Spark Streaming עמיד לכשלונות של איבוד מידע (**בחרו את התשובה המתאימה ביותר**)

א. המידע נשמר במערכת הקבצים

ב. המידע שנקלט משוכפל למכונות אחרות

ג. נשמר ה Lineage של ה - RDD

ד. לא ניתן לגבות כל רשומה ב – Spark Streaming , אפשר להשתמש בטכניקות שלעיל כדי להוריד למינימום את איבוד המידע

ה. אפשר לגבות הכל בעזרת א', ב', ג'

שאלה 20 (4 נקודות)

המתודה DStream#transform(**func**) (**ציינו את כל התשובות הנכונות**)

א. Transformation

ב. Action

ג. יוצר DStream חדש

ד. מוטציה של DStream

ה. מאפשרת לנו להשתמש במתודות של RDD אשר אינן מצויות ב - DStream