

**מבחן במסדי נתונים**

**עמוס עזריה ומירב שקרון**

**7029210-1,3,4,5**

סמסטר ב' מועד א' ג' באב התשפ"א, 12.7.2021

**הנחיות כלליות:**

* משך הבחינה: 150 דקות.
* **יש לענות בגוף השאלון!** המחברת תשמש כטיוטא בלבד, מענה במחברת עלול לגרור ציון 0.
* אין להכניס שום חומר עזר.
* השימוש במחשבון **אסור**.
* בשאלות האמריקאיות רק תשובה אחת נכונה. שימו לב שיש לנמק את הבחירה.
* בסיום הבחינה - נא למסור את השאלון ואת המחברת.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Total | 6 | 5 | 4 | 3 | 2 | 1 |  |
| 105 | 18 | 15 | 15 | 10 | 19 | 28 | Max points |
|  |  |  |  |  |  |  | Grade |

**ב ה צ ל ח ה!**

שאלה 1-SQL (28 נק')

נתונות הטבלאות הבאות המתארות נתונים מקופת חולים מסוימת הקשורים למחלת הקורונה:

patients(id, first\_name, last\_name, birth\_date, gender, first\_dose\_date, second\_dose\_date)

[טבלת מטופלים: ת"ז, שם פרטי, שם משפחה, תאריך לידה, מגדר, תאריך מנת חיסון ראשונה, תאריך מנת חיסון שניה]

isolated(patient\_id, start\_date, end\_date, isolation\_type\_id)

[טבלת מבודדים: ת"ז, תאריך כניסה לבידוד, תאריך סיום הבידוד, מזהה סוג בידוד]

isolation\_types(id, desc)

[טבלת סוג הבידוד: מזהה סוג בידוד, תיאור סוג בידוד]

הערכים בטבלה זו הם:

|  |  |
| --- | --- |
| **desc** | **id** |
| חדר ושירותים נפרדים. בבית גרים אנשים נוספים | 1 |
| חדר נפרד בבית. שירותים עם שאר יושבי הבית | 2 |
| לבד בבית | 3 |
| מלונית | 4 |

confirmed\_positive\_cases(patient\_id, start\_date, is\_symptomatic, end\_date)

[טבלת חולים מאומתים: ת"ז, תאריך תחילת מחלה, האם יש סימפטומים, תאריך סיום המחלה]

1. כתבו שאילתה שמחזירה את פרטי המטופלים העונים על כל התנאים הבאים:
   * חלו לאחר שהתחסנו במנה ראשונה (ולא השנייה)
   * כבר החלימו מהמחלה
   * הם בקבוצת הגיל 20-33

עבור כל מטופל/ת יש להציג את הנתונים הבאים:

* + שם פרטי
  + שם משפחה
  + ת"ז

(14 נק')

Select p.id, p.first\_name, p.last\_name from

patients as p join confirmed\_positive\_cases as cpc

on (p.id = cpc.patient\_id)

Where (cpc.start\_date id between p.first\_dose\_date and p.second\_dose\_date)

and cpc.end\_date is not NULL

and p.birth\_date.year between 1986 and 2001

1. כיתבו שאילתא שתציג את מס' האנשים שחלו בתקופת הבידוד שלהם לפי סוג הבידוד. כלומר, לכל סוג בידוד ייספרו רק מטופלים שאומתו בזמן שהותם בבידוד.

בכל רשומה יש להציג:

* מזהה סוג הבידוד
* תיאור סוג הבידוד
* מס' המטופלים שחלו בזמן שהותם בבידוד זה

(14 נק')

Select it.id, it.desc, count(i.patient\_id) as confirmed\_in\_isolation from

isolation\_types as it join isolated as i

On(it.id = i.isolation\_type\_id) join confirmed\_positive\_cases as cpc

On(cpc.patien\_id = i.patient\_id)

Where cpc.start\_date between i.start\_date and cpc.end\_date

Group by it.id

2. Normalization (19 נק')

נתונה הרלציה הבאה:

R(A, B, C)

נתונות השורות (הרשומות) הבאות ברלציה:

a1, b2, c1

a1, b1, c2

a2, b2, c1

כמו כן, נתון שמהחוקיות של הרלציה, ומהשורות הנתונות נובעות גם השורות הבאות (ייתכן שנובעות גם שורות נוספות):

a1, b1, c1

a1, b2, c2

1. מה ניתן לומר על הNF של הרלציה? נמקו היטב. (10 נק')

Relations:

A not->B, A not->C, A not->(B,C).

B not->A, B not->C, B not->(A,C).

C not->A, C not->B, C not->(A,B).

(A,B) not->C, (A,C) not->B, (B,C) not->A

* (A,B,C) is the candidate key. Then all vars are prime, Then obviously 2,3,3.5 holds.

4NF not holds since some rows arise from the others.

1. עבור כל אחת מהשורות להלן, ציינו האם **ייתכן** שהיא ברלציה (כיתבו "ייתכן"), לא ייתכן שהיא ברלציה (כיתבו "לא ייתכן"), או שהיא מוכרחת להיות ברלציה כיוון שהיא נובעת מכל הנתונים (כיתבו "נובעת"): (9 נק')

a2, b2, c2 \_\_\_ייתכן\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

a2, b1, c2 \_\_\_\_ייתכן\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

a2, b1, c1 \_\_\_\_\_ייתכן\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3. XML – XPath (10 נק')

נתון מסמך xml המתאר נתונים על חולי קורונה ומבודדים.

להלן דוגמא למבנה ותכולת הxml-

<corona>

<isolations>

<isolated>

<id>335598740</id>

<full\_name>”Tom Segev”</full\_name>

<age>36</age>

<gender>male</gender>

<start\_date>01/10/2021</start\_date>

<end\_date>14/01/2021</end\_date>

<place>”חדר ושירותים נפרדים. בבית גרים אנשים נוספים”</place>

</isolated>

**........**

</isolations>

<confirmed\_positive\_cases>

<case>

<id>775521547</id>

<name>”Jacky Levi”</name>

<age>56</age>

<gender>male</gender>

<start\_date>10/10/2021</start\_date>

<is\_symptomatics>”True”</is\_ symptomatics>

</case>

**.......**

</confirmed\_positive\_cases>

</corona>

כתבו שאילתא **אחת** ב-XPath שתחזיר את שמות החולים והמבודדים שגילם גדול מ-40 (מטופלים שמופיעים גם ברשימת החולים וגם ברשימת המבודדים יופיעו פעמיים בתוצאה)

//isolations/isolated[age>40]|//confirmed\_positive\_cases/case[age>40]

4. mongoDB (15 נק')

נתון collection של נתונים ב- mongoDB המאחסן נתוני מתחסנים לפי ערים בארץ.

שם ה-collection : Vaccinated.

להלן דוגמא למסמך ב-collection:

{

\_id: ObjectId("50a8240b927d5d8b5891743c"),

id: 0335598750

city: "Ariel",

first\_name: “Yael”,

last\_name: “Cohen”,

gender: “female”,

age: “39”

first\_dose\_date: new Date("May 04, 2021"),

second\_dose\_date: new Date("May 25, 2021"),

is\_recovered: “True”

}

כתבו שאילתא ב-mongoDB שתחזיר לכל עיר את מס' המחוסנים (חיסון ראשון) שהם גם מחלימים וגם מעל גיל 30.

db.Vaccinated.aggregate([

{

$match:{

is\_recovered:'True',

age:{$gt:30}

}

},

{

$group:

{

\_id:"$city", count:{$count:{}}

}

}

]);

5. Java Streams (15 נק')

כיתבו באמצעות שימוש ב-Java Streams בלבד פקודה המדפיסה את כל המספרים הראשוניים מ-2 ועד X נתון. **אין** לבנות פונקציית עזר, **אין** להשתמש בפונקציה קיימת לבדיקת ראשוניות (כגון: isProbablePrime) אין צורך להתחשב ביעילות. (רמז: העזרו ב-IntStream.rangeClosed(a,b) המחזירה stream המכיל את כל המספרים מ-a ועד b).

IntStream.rangeClosed(2, x).parallel()

.filter(

n->{

int sqrt = (int)Math.floor(Math.sqrt(n));

return !IntStream.rangeClosed(2, sqrt)

.filter(m->n%m == 0).findAny().isPresent();

}

)

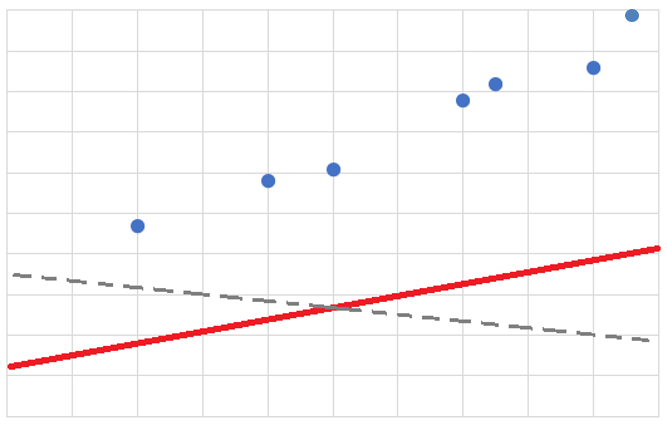
.forEach((e)->{

System.out.print(""+e+',');

});

6. Linear regression (18 נק')

נתון הגרף הבא המתאר דאטה (באמצעות הנקודות) עם פיצ'ר בודד (בציר ה-x) וה target label (בציר ה-y), וכן שני מודלים לינארים: הקו הרציף והקו המרוסק. הניחו שהקצה השמאלי למטה הוא ראשית הצירים (0,0).



ענו על השאלות הבאות. נמקו את תשובותיכם!

1. איזה מהמשפטים הבאים נכון בנוגע לLoss (פונקציית הטעות) של שני המודלים: (6 נק')
   1. הLoss של המודל של הקו המרוסק נמוך יותר מזו של הקו הרציף.
   2. הLoss של המודל של הקו הרציף נמוך יותר מזו של הקו המרוסק.
   3. לא ניתן לדעת איזה Loss נמוך יותר כיוון שאין מספרים על הצירים.
   4. הLoss של שני המודלים זהה, שכן כל הנקודות נמצאות מעל הקווים.

\_\_את ה loss מודדים ע״י ממוצא ריבועי המרחקים של הנקודות מהפונקציה בהתחשבות בערכי הy בלבד לכן אפילו שאין מספרים, ניתן לראות כי סכום המרחקים מהקו המרוסק גדול מסכום המרחקים מהקו הרציף ומכאן שגם פונקציית הloss של המודל המרוסק היא גבוהה יותר. לכן התשובה היא b

1. נסתכל על המודל של הקו **המרוסק**: מה ניתן לומר על הw של המודל של **הקו המרוסק**, ומה יקרה לה באיטרציה הבאה של Gradient descent? (6 נק')
   1. הw הנוכחית שלילית והיא תגדל (כלומר תהיה פחות שלילית או תהפוך חיובית).
   2. הw הנוכחית שלילית והיא תקטן (כלומר תהיה יותר שלילית).
   3. הw הנוכחית חיובית והיא תגדל.
   4. הw הנוכחית חיובית והיא תקטן (כלומר תהיה קטנה יותר או תהפוך שלילית).
   5. לא ניתן לדעת האם הw תקטן או תגדל לפי הנתונים בשאלה, הדבר תלוי במה שקרה באיטרציה הקודמת.

הפונקציה יורדת ולכן w שלילי. (אגב, נקודת החיתוך עם ציר הy חייובית לכן b חייובי). כאשר נציב ערך x של נקודה בפונקציית הנגזרת של הloss לפי w, נקבל תוצאה שלילית שכן ולכן גם הממוצע יהיה שלילי. מw נחסיר את הממוצע השלילי הזה כלומר נגדיל את w ! התשובה היא a

1. נסתכל על המודל של הקו **הרציף**: מה ניתן לומר על הנגזרת של הLoss בנקודה w, b התואמים את אלו של המודל של הקו הרציף. זכרו שבGradient descent אנו תמיד מתקנים בצעד בכיוון **הפוך** לנגזרת. (6 נק')
   1. לא ניתן לומר דבר כיוון שאין מספרים על הצירים.
   2. הנגזרת החלקית לפי w בנקודה w,b תהיה חיובית, ואילו הנגזרת החלקית לפי b בנקודה w,b תהיה שלילית.
   3. הנגזרת החלקית לפי w בנקודה w,b תהיה שלילית, ואילו הנגזרת החלקית לפי b בנקודה w,b תהיה חיובית.
   4. הנגזרת החלקית לפי w בנקודה w,b תהיה חיובית, וכן הנגזרת החלקית לפי b בנקודה w,b תהיה חיובית.
   5. הנגזרת החלקית לפי w בנקודה w,b תהיה שלילית, וכן הנגזרת החלקית לפי b בנקודה w,b תהיה שלילית.

נזכר בפונקציית הנגזרת החלקית:

נשים לב ש ובנוסף ולכן

אזי שתי הנגזרות החלקיות יהיו שליליות ולכן התשובה היא e.