**מבחן בקורס מבוא לביואינפורמטיקה**

**מועד' ב, סמסטר א', תשע"י**

שמות המרצים : פרופ' טל פופקו, ד"ר יפתח נחמן, ד"ר דוד בורנשטיין

משך הבחינה : שעתיים

חומר עזר מותר בשימוש : -

חובה לענות על כל השאלות

לכל שאלה תשובה אחת נכונה.

דפי התשובות לבחינה יבדקו ע"י סורק במזכירות הפקולטה.

להזכירכם:

לתשובה נכונה – יש לסמן X במשבצת המתאימה. ניתן להשתמש בעט.שחור

הסימון לא צריך להיות עבה אלא ככתב רגיל.

למחיקה – יש למלא את הריבוע לגמרי מבלי לחרוג לריבועים סמוכים.

אין להשתמש במחק או בטיפקס.

אין לסמן דבר מלבד סימון פרטי הזיהוי בראש הטופס והתשובות בריבועים המיועדים לכך.

את מס' תעודת הזהות יש למלא ב – 9 ספרות (כולל ספרת ביקורת ואפס אם יש בתחילת המספר) לסמן ב – X לפי הדוגמא.

מס' סידורי – למלא ב – 3 ספרות לדוגמא : מס' 9 יסומן 009.

## יש לסמן את התשובות גם בגוף השאלון

שם התלמיד \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

מס' תלמיד \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

מס' סידורי \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**1. סמן את התשובה הנכונה**א. מספר שכני ה SPR גדול ממספר שכני ה NNI עבור עץ חסר שורש עם 5 מינים.

ב. ניתן לעבור מכל עץ עם 71 מינים לכל עץ אחר עם אותו מספר מינים על ידי מספר כלשהו של צעדי NNI.

ג. יש טעם להריץ SIMULATED ANNEALYING עבור עץ עם 5 מינים.
ד. מספר העצים עם שורש עבור 771 מינים קטן ממספר העצים חסרי שורש עבור 770 מינים

**2. העמידו רצף חלבון X של אורגניזם A מול רצף חלבון Y של אורגניזם B בעימוד לוקאלי. בפלט של העימוד המיטבי (עם הציון הגבוה ביותר), עמדה 127 בחלבון X מועמדת מול עמדה 132 בחלבון Y. מה נכון בהכרח?**א. חומצת האמינו בעמדה 127 בחלבון X זהה לחומצת האמינו בעמדה 132 בחלבון Y.

ב. יש אינדלים בפלט של העימוד המיטבי.
ג. יש לפחות mismatch אחד בפלט של העימוד המיטבי.
ד. א+ב.
ה. א'+ג'.
ו. ב'+ג'.
ז. אף תשובה אינה נכונה.

יתכן והעימוד הלוקלי יחזיר משהו כזה

X DDDCVCDDDYD

Y DDDCACDDDYD

כשחומצה V נמצאת בעמדה 127 ברצף X , וחומצה A נמצאת בעמדה 132 ברצף Y, במקרה זה ברור שמסיחים א,ו-ב שגויים. ג׳ שגוי טריוויאלית (למשל אם הרצפים זהים).

**3. במטריצה הבאה של עימוד לוקאלי, מה הציון ל mismatch?**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | C | G | A |
|  | 0 | 0 | 0 | 0 |
| A | 0 | 0 | 0 | 3 |
| C | 0 | 3 | 1 | 1 |
| A | 0 | 1 | 1 | 4 |

א. -3
ב. -2
ג. -1
ד. 0
ה. 1
ו. 2
ז. 3
ח. אף תשובה אינה נכונה.

פתרון: הציון ל MATCH חייב להיות 3. הציון לאינדל חייב להיות -2. הציון ל mismatch יוצא -2.

**4. אורך העימוד הוא מספר העמדות שיש בעימוד. למשל אורך העימוד הבא הוא 5:**

**ACG-T
ACGG-**

**סמן המשפט הנכון ביותר**א. יתכן והאורך של עימוד לוקאלי יהיה גדול מהאורך של כל אחד מהרצפים.

ב. יתכן והאורך של עימוד גלובלי יהיה קטן מהאורך של אחד מהרצפים.
ג. בעימוד גלובלי של שני רצפים עם הפרש אורכים של אחד (נניח רצף א באורך 1000 ורצף באורך 1001), בהכרח יהיה אינדל באורך 1 בדיוק.
ד. בעימוד גלובלי של שני רצפים עם הפרש אורכים של 5 בהכרח יהיה אינדל באורך גדול מ 1.
ה. א'+ג'.
ו. ב'+ד'.
ז. ג'+ד'.
ח. אף תשובה אינה נכונה.

פתרון: א בהכרח ייתכן לדוגמא:

AAACCC-GGGTTTATTT

AAACCCAGGGTTT-TTT

ב' לא יתכן כי בעימוד גלובלי נכללים כל הרצף של A ושל B. ג' לא בהכרח, לדוגמא בעימוד של

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Y | Y | Y | Y | - | - | - | G | G | G | G | A | A | A |
| Y | Y | Y | Y | G | G | G | G | G | G | G | - | - | A |

ד' לא בהכרח – ייתכנו למשל 5 אינדלים לא סמוכים שכל אחד מהם באורך 1.

**5. סמן את המשפט הנכון ביותר לגבי עצים פילוגנטיים:**א. עבור N גדול מ 5 מספר העצים חסרי שורש תמיד יתחלק ב 7.
ב. אם נגדיל את מספר המינים מ 100 ל 101, מספר העצים עם שורש יוכפל פי 197
ג. מספר האפשרויות לשרש עץ חסר שורש עם 1000 מינים גדול ב 1997 ממספר האפשרויות לשרש עץ חסר שורש עם 999 מינים.
ד. א+ב.
ה. א+ג.
ו. ב+ג.

ו. א+ב+ג.

ז. אף תשובה אינה נכונה.

פתרון: מספר העצים חסרי שורש עבור N שווה 6 הוא 1\*3\*5\*7 ועבור N גדול מ 6 זה יהיה הפקטור הזה כפול משהו. מסיח ב' שגוי כי לעץ עם שורש עם 100 מינים יש 198 ענפים. יש 199 מקומות לחבר את המין הבא ולכן מספר העצים יגדל פי 199. ג' שגוי כי מספר העצים גדל ב 2 כל פעם.

6. **סמן את המשפט הנכון ביותר**:

1. באלגוריתם שלFITCH , זמן החישוב פרופורציונלי למספר הקודקודים הפנימיים כפול מספר העמדות.
2. באלגוריתם שלFITCH , כשמספר המינים הוא 20 ובעמדה מסוימת ל 18 מינים יש את אותה אות (למשל C) ולשני המינים יש את אותה אות(למשל A) אז ציון הפרסימוניה של העמדה הוא 2 ללא תלות בעץ.
3. כשמספר המינים הוא 12, יהיו תמיד בדיוק (לא פחות ולא יותר) 21 עצים בעלי שורש עם אותו ציון פרסימוניה.
4. באלגוריתם שלFITCH , בעץ עם 8 רצפים, בעמדה מסוימת ציון הפרסימוניה יכול להיות 8.

פתרון: א נכון: עבור כל קודקוד פנימי אנו עושים איחוד או חיתוך ואת זה אנו צריכים לעשות עבור על עמדה. ב' שגוי כי אם 2 המינים יוצרים תת-עץ, הציון יהיה 1. ג' שגוי כי יתכן שכל העצים יהיו עם אותו ציון (למשל שכל הרצפים זהים זה לזה). ד' שגוי כי הציון המירבי עבור 8 רצפים הוא 7.

7. **מהו ציון הפרסימוניה של העץ הבא?**



**כשהרצפים הם**

|  |  |
| --- | --- |
| **GDLLCK** | **Human** |
| **GDMLCM** | **Chimp** |
| **GDNLCA** | **Gorilla** |
| **ADLLCT** | **Horse** |
| **GDLLAK** | **Cow** |

א. 7
ב. 8
ג. 9
ד. 10
ה. 11
ו. 12
ז. 13

ח. אף תשובה אינה נכונה

פתרון: עמדה 1: **1** שינויים. עמדה 2: **0** שינויים, עמדה 3: **2** שינויים. עמדה 4: **0** שינויים. עמדה 5: **1** שינויים. עמדה 6: **3** שינויים. סה"כ: 7 שינויים.

8. **מהו היחס בין מספר העצים בעלי שורש האפשריים עם N=71 רצפים למספר העצים חסרי שורש עם N=70 רצפים?**א. 70
ב. 71
ג. 137
ד. 139
ה. 139/137

ו. 139\*137

ז. אף תשובה אינה נכונה.

פתרון: מספר העצים בעלי שורש הוא 1\*3\*5\*…\*(2n-3) . מס העצים חסרי שורש הוא: 1\*3\*5\*…\*(2n-5)

ולכן יש לנו:

1\*3\*…\*(2\*71-3)/1\*3\*…\*(2\*70-5) = 1\*3\*…\*139/1\*3\*…\*135 = 137\*139

9. מה מהבאים נכון לגבי אלגוריתם הסיווג KNN (classification):

א. ב-KNN זמן האימון תלוי במספר הנקודות.

ב. KNN מחלק את הנקודות לשני תתי מרחבים.

ג. K הוא תמיד אי-זוגי.

ד. אף אחת מהתשובות לא נכונה.

10. הגרף הבא מייצג מדידה של רמות ביטוי של שני גנים עבור רקמות בריאות, המוגדרות חיוביות ומסומנות בלבן ורקמות חולות, המוגדרות שליליות ומסומנות בשחור.
**עיגולים מסמלים רקמות שמשמשות ל-training וריבועים כאלו שמשמשות ל-testing.**

אם השתמשו במסווג מסוג KNN מה תהיה רמת ה-Accuracy עבור K=5?

1. 1/5
2. 1/3
3. 1/2
4. 2/5
5. 2/3
6. 3/5
7. 4/5
8. 1

*פתרון: מתוך חמשת הריבועים ב-test set, רק הלבן ברבע הימני התחתון יסווג לא נכון (כי לפחות שלושת העיגולים הקרובים אליו שחורים), עבור כל האחרים שלושת העיגולים הקרובים ביותר הם בצבע של הריבוע.*

11. עבור אותם נתונים מהשאלה הקודמת, מה תהיה רמת ה-Precision עבור K=5?

1. 1/5
2. 1/3
3. 1/2
4. 2/5
5. 2/3
6. 3/5
7. 4/5
8. 1

*פתרון: שני ריבועים יקבלו סיווג חיובי (עליון ימני ותחתון שמאלי) ושניהם אכן חיוביים – לכן 1.*

12. מה מהבאים נכון לגבי רשתות נוירונים (Neural networks):

1. בשלב האימון ברשת נוירונים משתנה מספר הנוירונים ברשת.
2. בשלב האימון ברשת נוירונים משתנים המשקלים ברשת.
3. מספר הנוירונים ברשת נוירונים ללא שכבות נסתרות תלוי במספר הפיצ'רים (features).
4. אף אחת מהתשובות לא נכונה.

*פתרון: באימון רשת נוירונים משתנים המשקלים על הקשתות השונות. מספר השכבות בשכבת הקלט (השכבה הראשונה) הוא כמספר ה-features.*

13. מה מהבאים נכון לגבי(SVM) Support vector machine :

1. בSVM- אין שלב של אימון (Lazy algorithm).
2. SVM יכול ליצור הפרדה של המרחב לשניים.
3. SVM יכול ליצור תבניות הפרדה מורכבות על-ידי העלאת הבעיה למימד גבוה יותר.
4. אף אחת מהתשובות לא נכונה.

*פתרון ב-SVM יש תקופת אימון שהיא קביעת הקו/מישור/על-מישור מפריד*

14. מה נכון עבור הנתונים המופיעים בתרשים?

1. ניתן לבצע סיווג טוב של הנתונים ע"י עץ החלטה.
2. ניתן לבצע סיווג טוב של הנתונים ע"י רשת נוירונים, אבל רק עם שכבות נסתרות.
3. ניתן לבצע סיווג טוב של הנתונים ע"י KNN.
4. אף אחת מהתשובות לא נכונה.

15. מה מהבאים נכון?

1. אם רוצים לקבץ נתונים ל-K קבוצות ניתן להשתמש באלגוריתם UPGMA.
2. אם רוצים לקבץ נתונים ל-K קבוצות ניתן להשתמש באלגוריתם KNN.
3. אם רוצים לקבץ נתונים ל-K קבוצות ניתן להשתמש באלגוריתם K-means.
4. אף אחת מהתשובות לא נכונה.

*פתרון: UPGMA משרה קיבוץ לכל K, ניתן "לחתוך" את העץ כך שיווצרו K קבוצות. KNN הוא אלגוריתם סיווג ולא מחלק את הנתונים לקבוצות. K-means מקבץ את הנתונים ל-K קבוצות.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| g7 | g6 | g5 | g4 | g3 | g2 | g1 |  |
| 29 | 28 | 23 | 25 | 22 | 8 | 0 | g1 |
| 27 | 33 | 32 | 43 | 26 | 0 |  | g2 |
| 23 | 6 | 34 | 54 | 0 |  |  | g3 |
| 34 | 24 | 33 | 0 |  |  |  | g4 |
| 23 | 22 | 0 |  |  |  |  | g5 |
| 27 | 0 |  |  |  |  |  | g6 |
| 0 |  |  |  |  |  |  | g7 |

16. אילו מהמשפטים נכון לגבי UPGMA בהינתן מטריצת המרחקים הבאה:

1. בחלוקה ל-6 קבוצות g1 ו-g2 יהיו באותה קבוצה.
2. בחלוקה ל-6 קבוצות g6 ו-g3 יהיו באותה קבוצה.
3. בחלוקה ל-2 קבוצות g1 ו-g2 יהיו באותה קבוצה.
4. אף אחת מהתשובות לא נכונה.

*פתרון: בחלוקה ל-6 נעשה רק איחוד אחד של הזוג הקרוב ביותר: g6 ו-g3. הקיבוץ הבא יהיה בין g1 ו-g2 ולכן בחלוקה לחמש קבוצות או פחות g1 ו-g2 יהיו באותה קבוצה.*

 17. מוטיב רשת X מופיע c פעמים ברשת בקרת השיעתוק של שמר. כדי לבדוק האם הוא מופיע **מעל** המצופה באקראי, ייצרו 10,000 רשתות מעורבבות אקראית מהרשת המקורית. באף אחת מהן לא התקבל המוטיב בדיוק c פעמים. ב-200 מתוכן X מופיע מעל c פעמים, וב-50 מתוך אותן 200, X מופיע מעל 2\*c פעמים. מהו ה P-value לשאלה ששאלו?

א. 0.01

ב. 0.02

ג. 0.02\*c

ד. לא ניתן לדעת

ה. 0.0004

*פתרון: ב-2% מהרשתות האקראיות מופיע המוטיב c פעמים ומעלה, לכן ההסתברות לקבלת c או יותר באקראי היא 0.02.*

18. נתון מוטיב רשת בקרת שיעתוק מסוג C2: גן X מדכא את שיעתוק Y ו Z, וגן Y מפעיל את שיעתוק Z (כאשר רמת Y גבוהה מסף מסוים). X מופעל על-ידי אות Sx בתגובה מיידית. פונקצית בקרת השיעתוק על Z היא: ON כאשר X לא פעיל ו-Y גבוה מהסף, OFF בכל צירוף אחר.

במצב ההתחלתי X כבוי,Z,Y ברמתם המקסימלית. מדליקים את האות Sx לזמן ארוך, ואז מכבים אותו.

הפלט Z מראה את התכונות הבאות ביחס לאות Sx:

א. עיכוב בתגובה להדלקת אות, עמידות כנגד הפסקה קצרה (off pulse) באות.

ב. עיכוב בתגובה להדלקת אות, עמידות כנגד on pulse של אות.

ג. עיכוב בתגובה לכיבוי אות, עמידות כנגד הפסקה קצרה (off pulse) באות.

ד. עיכוב בתגובה לכיבוי אות, עמידות כנגד on pulse של אות.

19. מי ממוטיבי הרשת הבאים יכול להביא למצב של תנודות מחזוריות:

א. משוב שלילי

ב. משוב חיובי

ג. משוב דיכוי הדדי (mutual inhibition)

ד. Incoherent (I1) feed forward loop.

ה. C1 coherent feed forward loop.

*פתרון: משוב שלילי הוא המוטיב היחיד שיכול (בפרמטרים מסוימים) להביא לתנודות מחזוריות.*

20. לרשת scale-free יש את התכונות הבאות, בהשוואה לגרף רנדומי עם אותו מספר קדקדים ואותה דרגה ממוצעת:

א. יותר hubs (קדקדים עם דרגה גבוהה בהרבה מהממוצע).

ב. עמידות טובה יותר לכשל קדקד שנבחר על ידי תוקף זדוני.

ג. מרחק קצר יותר בין שני קדקדים אקראיים.

ד. עמידות טובה יותר לכשל 1% מהקדקדים שנבחרו באקראי.

*פתרון: בגרף רנדומי יש התפלגות דרגות נורמלית בקירוב סביב הממוצע, בעוד ברשת scale-free* יש התפלגות חוק חזקה, ולכן יש יותר קדקדים עם דרגה גבוהה. ריבוי הhubs ומבנה הרשת גורמים לתופעת ה-small world, כלומר מרחק קצר יחסית בין זוגות אקראיים של קדקדים. כמו כן, 1% מהקודקודים שנבחרו באקראי יכילו בעיקר קודקודים עם דרגה נמוכה מאד, ולכן קישוריות הרשת כנראה תיפגע פחות.

21. חולה אושפז לאחר שנחשף לקרינת UV שהשביתה 1% אקראיים מהגנים בתאי הריאה שלו, כולל הגן X. בביה״ח נדבק בנגיף שהשבית hub מרכזי Y ברשת בקרת השיעתוק של תאי הריאה. נגדיר **קישוריות** של קודקוד בתור מספר הקודקודים אליהם ניתן להגיע במסלול באורך כלשהו מאותו קודקוד. בהסתברות גבוהה:

א. השבתת X פגעה בקישוריות הממוצעת ברשת יותר מהשבתת Y.

ב. המרחק הממוצע של X משאר קדקדי הרשת גדול יותר מהמרחק הממוצע של Y משאר קדקדי הרשת.

ג.Y כנראה נוסף לרשת לפני X.

ד. קרינת ה-UV פגעה בקישוריות הממוצעת ברשת יותר מתקיפת הנגיף.

*פתרון*: נתון ש-Y הוא hub (כי הוא נבחר על ידי תוקף זדוני), וסביר של-X קישוריות נמוכה (כי הוא אחד מתוך 1% שנבחרו באקראי), ולכן פגיעה ב-Y תשפיע על קישוריות הרשת יותר. כמו כן סביר ש-X נוסף לרשת מזמן. גם ברמת הפגיע הכוללת, אמרנו שאפילו פגיעה של 5% אקראיים מהרשת לא תפגע קשות בקישוריות.

22. באילו מקרים למדידות single cell RNA seq **לא יהיה** יתרון על מדידות microarrays?

א. זיהוי תאים נדירים בתוך אוכלוסיית תאים גדולה.

ב. זיהוי שלבי תהליך הקורה בקצב שונה בין תא תא באוכלוסית תאים שהתחילו באותו מצב.

ג. זיהוי שלבי תהליך הקורה בקצב זהה בכל התאים באוכלוסית תאים שהתחילו באותו מצב.

ד. זיהוי הרכב התאים בריקמה הטרוגנית.

*פתרון: אם תהליך קורה בקצב זהה באוכלוסיית תאים, הרי שכל התאים נמצאים באותו המצב בנקודת זמן מסוימת, ולכן אין יתרון למדוד את פרופילי הביטוי שלהם בנפרד.*

23. אנו חוקרים תאי גידול סרטן שד. מדדנו פרופילי ביטוי mRNA מ-10000 תאים בודדים שנלקחו מ-4 גידולים שונים בשלבים שונים של המחלה. X, Y הם גנים שביטויים עולה תוך כדי המחלה.

מה **לא נכון** לאמר?

א. אם PCA יפריד את התאים ל-4 שלבים שונים במחלה, כנראה שגם tSNE יעשה זאת.

ב. בעזרת PCA אפשר לתאר את התאים במערכת צירים חדשה, שכל ציר בה הוא איזושהי קומבינציה לינארית של רמות כל הגנים.

ג. אם נריץ פעמיים את tSNE עם אותם פרמטרים, נקבל את אותה התוצאה.

ד. Monocle יוכל לזהות שגן X משתנה לפני גן Y לאורך שלבי המחלה.

*פתרון: tSNE הוא אלגוריתם איטרטיבי סטוכסטי, ולכן שתי ריצות עם פרמטרים זהים יגיעו לתוצאות שונות.*

**24**. בשיטת ה-DropSeq לכל bead מחוברים כ-108 פרובים. בכל אחד מהם יש ברקוד לתא (BC) באורך 12 נוקלאוטידים, ומזהה מולקולרי ייחודי (UMI) באורך 8 נוקלאוטידים. אילו טענות נכונות?

א. מספר מולקולות הmRNA לגן X בתא המזוהה על ידי BC1 מוערך על ידי מספר ה-UMIים השונים שרוצפו מחוברים למקטע של X עם 1BC .

ב. מספר מולקולות הmRNA לגן X בתא המזוהה על ידי BC1 מוערך על ידי מספר העותקים של X עם 1BC שרוצפו, ללא קשר ל-UMI בכל אחד מהעותקים.

ג. ניתן להבחין בניסוי אחד בין 412 תאים לכל היותר.

ד. ניתן להבחין בניסוי אחד בין420 תאים לכל היותר.

*פתרון: מספר התאים הניתנים להבחנה הוא לכל היותר מספר ה-bc האפשריים , כלומר 412.*

*לקבלת מספר עותקי הmRNA בתא, אנחנו סופרים את ה-UMI השונים, ולא משנה לנו כמה עותקים התקבלו מכל UMI.*