# שאלה תיאורטית

## שאלה 1

יהי D מאגר אימון עבורו קיימת פונקציה f כלך שקיימות שתי תכונות נומינאליות a, b המקיימות –

נסמן ב את המאגר המתקבל מ D ע"י הסרת התכונה a, וב המתקבל מ D ע"י הסרת

הוכיחו\הפריכו:

1. **T ו בהכרח יסווגו באופן זהה כל דוגמאת מבחן.**

הוכחה:

ההבדל בין עץ לעץ הוא תוספת פיצול על פי תכונה a.

השאלה היא האם תוספת הפיצול על פי תכונה a משנה את הסיווג.

מאחר ו נתון כפונקציה של , ניתן לומר כי לכל ערך של תכונה b נקבל את אותו ערך בתכונה a. אם כך הוספת פיצול בעץ על פי a תתן חלוקה זהה לחלוקה שתתקבל מפיצול בעזרת תכונה b.

יהי דוגמאת מבחן x עבורה , ודוגמאת מבחן y עבורה .  
נניח כי פיצול על פי תכונה a אכן משנה את הסיווג. בהכרח מתקיים ש .

מאחר ונתון כי , ניתן לומר ש , וקיבלנו סתירה.

במילים אחרות אם נוסיף ב  פיצול על פי תכונה , x ו y יהיו בהכרח באותה הקבוצה, ולכן תכונה לא שינתה את הסיווג, ולכן T ו בהכרח יסווגו באופן זהה כל דוגמאת מבחן.

1. **T ו בהכרח יסווגו באופן זהה כל דוגמאת מבחן.**

הפרכה:

ההבדל בין עץ לעץ הוא תוספת פיצול על פי תכונה b.

השאלה היא האם תוספת הפיצול על פי תכונה b משנה את הסיווג.

ניתן דוגמא ל a ו b עבורם נקבל שהסיווג של שתי דוגמאות מבחן ב T יהיה שונה מהסיווג ב **.**

נניח ומטרת המסווג היא סיווג של עופות כאשר הסיווג נעשה על ידי שתי תכונות בלבד, a ו b.

תכונה a היא מספר הרגליים של בעל החיים, ותכונה b היא האם לבעל החיים יש נוצות.

מטרת הסיווג היא הבחנה בין בעל חיים שיכול לעוף לבין בעל חיים שלא.  
נניח והעולם שלנו לא מכיל את החרקים. נרצה לסווג בעל חיים שיש לו כנפיים כבעל יכולת מעוף (על אף שזה לא בהכרח נכון).

הקשר בין a ל b:  
a היא פונקציה שערכה תמיד 2 מאחר ולכל בעלי החיים שאינם חרקים, בעל חיים שהוא בעל מעוף הוא בעל שתי רגליים ולכן .

נתבונן בשתי דוגמאות מבחן, פרה ודרור.  
פיצול על פי תכונה a, מספר הרגליים, יסווג אותם לאותה קבוצה (כאשר סיווגה הסופי של הקבוצה תלוי בערך ה default של הקבוצה על פי דוגמאות המבחן) ולכן הסיווג הסופי שלהם יהיה זהה.  
אם נוסיף פיצול על פי תכונה b נקבל שהם ישוייכו לקבוצות שונות כאשר הפרה תסווג כחסרת יכולת מעוף ואילו הדרור יסווג כבעל חיים בעל יכולת מעוף.

# מימוש עץ החלטה מסוג ID3

## שאלה 2

**מימוש קובץ DT.py:**

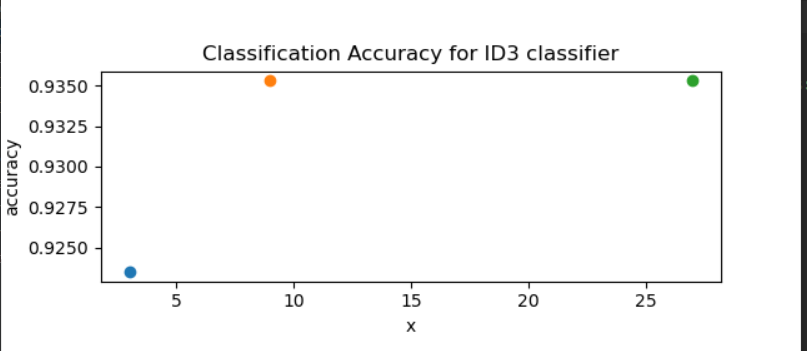
**טעינת ה data, בניית עץ החלטה ID3 לא גזום בעזרת קבוצת האימון, והדפסת דיוק (accuracy) הסיווג של קבוצת המבחן.**

הדיוק שהתקבל, 0.9235, 92.35%.

## שאלה 3

**גרף המתאר את דיוק עצי ההחלטה הגזומים על קבוצת המבחן כתלות בגודל x עבור הערכים {3, 9, 27}.  
נתחו בקצרה את התוצאות שקיבלתם.**

הגרף שהתקבל:



הדיוק שהתקבל עבור : 0.9235, 92.35%.  
הדיוק שהתקבל עבור : 0.9352, 93.52%.  
הדיוק שהתקבל עבור : 0.9353, 93.53%.

ניתוח התוצאות:

מטרת המסווג היא לדייק בסיווג של אובייקטים לא מסווגים שיגיעו בעתיד.  
התאמת יתר, overfitting, היא מצב שבו הגדלת שגיאת האימון תגרום לירידה בשגיאת המבחן.  
התאמת יתר נובעת בדר"כ מדוגמאות רועשות (כלומר הסיווג שלהן שגוי).

בלמידת עצי החלטה, מתמודדים עם התאמת יתר ע"י הקטנת העצים – מעדיפים עצים קטנים יותר ולא עקביים. אלגוריתם ID3 ללא גיזום הוא אלגוריתם עקבי.

גיזום עץ ההחלטה נעשה כדי להקטין את העץ והחליש את אפקט ה overfitting.

ניתן לראות כי כשהקטנו את העץ, וגזמנו "גיזום מוקדם" על פי מספר דוגמאות מינימלי בצומת, הדיוק גדל (כלומר שגיאת המבחן קטנה) וקיבלנו תוצאות טובות יותר.

ב x=3 העץ גדול יותר מאשר בשני המקרים האחרים, ולא משפר משמעותית את הדיוק ולכן קיבלנו דיוק בסדר גודל של דיוק המסווג ללא הגיזום.  
ב x=9, 27 ניתן לראות מגמת עלייה בדיוק (ביחס ל x=3), כאשר x=27 טוב יותר מ x=9.  
למרות זאת ניתן לראות שהפער ביניהם איננו משמעותי (~0.01%) כך שניתן להסיק שהרמות האחרונות בעץ עם הפרמטר x=9 מצד אחד לא הוסיפו כל כך אינפורמציה בסיווג ולכן ניתן לוותר עליהן, ומצד שני שאפקט ה overfitting פחות מורגש כבר בגיזום זה (מאחר ו x=27 לא הוסיף לדיוק).שאלות תיאורטיות

## שאלה 4

**יהא T עץ החלטה לא גזום, ויהא העץ המתקבל מT באמצעות גיזום מאוחר שבו הוסרה הרמה התחתונה של T.  
הוכיחו\הפריכו: בהכרח קיים כך שהעץ T עם כלל אפסילון-החלטה והעץ עם כלל החלטה הרגיל יסווגו כל דוגמאת מבחן בצורה זהה.**

הפרכה: יהי T עץ עם כלל אפסילון החלטה בעל 3 תכונות רציפות בעלי ערכי הסף v1=1 ערך סף המתקבל מתכונה מס' 1 v2=2 ערך הסף המתקבל מתכונה מס' 2 וv3=2 ערך הסף מתכונה מס' 3(כפי שגם מצוין בגרף). ונקבע שT מסווג את כל הדוגמאות שקטנות/שוות מערך הסף לבן השמאלי ואת הדוגמאות הגדולות מערך הסף לבן הימני.

ויהי T:

**V1 = 1**

**V3 = 2**

**V2 = 2**

**-**

**- -**

**+**

**+ +**

T' העץ המתקבל מגיזום העלים של T: (בשוויון בין הדוגמאות עץ ההחלטה מסווג +)

**V1 = 1**

**+ + -**

**- - +**

יהי דוגמא x=[1,2,2] נסמן ערכי התכונות של הדוגמא x:

סיווג x לפי T:

-בפיצול הראשון: , כלומר לכן השורש ימשיך את הסיווג לפי 2 הבנים שלו.

-בפיצול של הבן השמאלי: , כלומר לכן הבן השמאלי ימשיך את הסיווג לפי 2 הבנים שלו.

*-בפיצול של הבן הימני:* , כלומר לכן הבן הימני ימשיך את הסיווג לפי 2 הבנים שלו*.*

*סה"כ T יסווג את x לפי הרב בכל הדוגמאות (בכל צומת מתקיים כלל ההחלטה אפסילון ולכן המסווג מגיע לכל העלים) יש לT 3 דוג' + ו3 דוג' – ולכן זהו שוויון אשר יסווג את x להיות +*

*סיווג לפי T'(T גזום ללא כלל אפסילון החלטה):*

*-בפיצול של השורש: ולכן x1 יסווג לבן השמאלי בT' כלומר לצומת בה 2 דוגמאות – ודוגמא אחת + ולכן x יסווג להיות –*

*כלומר T וT' יסווגו את x שונה .*

## שאלה 5

**יהי מאגר נתונים D, ויהא שנבנה תוך שימוש ב -D ונגזם ע"י הפסקת פיצול צמתים החל מעומק . סטודנט בקורס שם לב שכאשר משתמשים בכלל אפסילון-החלטה ריבוי המסלולים האפשריים עלול להאריך מאוד את שלב הסיווג (inference). הוא הציע לבנות באמצעות D עץ ID3 אחר, נסמנו , אשר נגזם באותו באופן, ונבנה כך: בכל פעם שמגיעים לצומת בעץ המפצל לפי ערכי התכונה עם ערך הסף , כל דוגמאת אימון בצומת המקיימת תועבר לשני הבנים שלו.  
בשלב המבחן, הסיווג יקבע על פי כלל ההחלטה הרגיל (ולא כלל אפסילון-החלטה).  
הוכיחו\הפריכו: העץ עם כלל אפסילון-החלטה והעץ עם כלל ההחלטה הרגיל בהכרח יסווגו כל דוגמאת מבחן באופן זהה.  
הבהרה: שני העצים משתמשים באותו וקטור .**

הפרכה:

תהי x דוגמא שמקיימת את כלל האפסילון בתכונה בלבד, כלומר מתקיים ש .

נניח מתקיים , כלומר הדוגמא תלך לבן עם ערך 1 עבור הסף בעץ הרגיל . נניח שהפיצול הנ"ל מתרחש ברמה ה , כלומר הוא זה שמחלק לעלים. הסיווג של x הוא שלילי.

כמו כן נניח שאם נוציא את x, היחס בסיווגים בעץ המקורי בנקודת הפיצול הנ"ל ברמה ה בבן עם ערך 0 הוא y+1 חיוביים מול y שליליים, ואילו בבן עם ערך 1 הוא מאוזן (כלומר y+1 חיוביים ו y+1 שליליים).

מאחר וסיווגה של x הוא שלילי, בעץ , שבונה את העץ על פי כלל ההחלטה הרגיל, הדוגמא תגיע לבן ערך 1, ותגרום לכך שמספר הדוגמאות החיוביות סה"כ בשני העלים יהיה 2y+2 ומספר הדוגמאות השליליות יהיה y+(y+1)+1=2y+2 כלומר על פי כלל אפסילון החלטה נקבל מאזן, ולכן הסיווג של x יהיה חיובי (זוהי ברירת המחדל במקרה של שוויון).

לעומת זאת, בעץ , הדוגמא x תגיע לשני העלים ותגרום למאזן בבן עם ערך 0 להיות y+1 חיוביים ו y+1 שליילים (הסיווג של x הוא שלילי), ולמאזן בבן עם ערך 1 להיות y+1 חיוביים ו y+2 שליליים.  
כאמור, x יגיע לבן ערך ערך 1 (בעץ זה כלל ההחלטה הוא רגיל). מאחר ושם יש רוב של דוגמאות שליליות (y+2 מול y+1), הסיווג של x יקבע להיות שלילי.

קיבלנו שהסיווג של x בעץ יהיה חיובי, ואילו בעץ שלילי.

# מימוש עץ החלטה המשתמש בכלל אפסילון החלטה

## שאלה 6

מימוש קובץ DT\_epsilon.py.

הדיוק שהתקבל: 0.9335, 93.35%.

# שאלה תיאורטית

## שאלה 7

הפרכה:

יהי מאגר הדוגמאות הבאים:

מאגרי דוגמאות אלו מכילים תכונה אחת בעלת ערכים ממשיים וסיווג בינארי T,F

מחיתוך/איחוד קבוצות אלו יתקבלו הקבוצות הבאות:

*נבנה את הקבוצה :*

*לכל נחלק למקרים*

*אם : אז לכן (כיוון שאין חזרות בקבוצות 1,2 => x מופיע פעמיים ב*

*אם : (ושייך לאיחוד) אז*

לכן => x מופיע פעם אחת ב

ואם אז ולכן x לא יופיע ב כלל.

הקבוצה המתקבלת:

*תהי דוגמא x=[2] כלומר x דוגמא בעלת ערך תכונה של 2*

*סיווג x לפי : שלושת הדוגמאות הקרובות ביותר לדוגמא בקבוצת האיחוד*

הם : {(1,T),(2,F),(4,T)} ולכן הדוגמא תסווג ל T

סיווג x לפי  *: שלושת הדוגמאות הקרובות ביותר לדוגמא בקבוצת החיתוך*

הם: {(1,T),(2,F),(8,T)} ולכן הדוגמא תסווג ל T

אזי לפי הטענה גורר שגם

סיווג x לפי : שלושת הדוגמאות הקרובות ביותר לדוגמא בקבוצת החיבור

הם: {(2,F),(2,F),(1,T)} *הדוגמא x תסווג ל F בסתירה לטענה ולכן הטענה אינה נכונה.*

# מימוש KNN ובחינת השפעה של פרמטר K

## שאלה 8

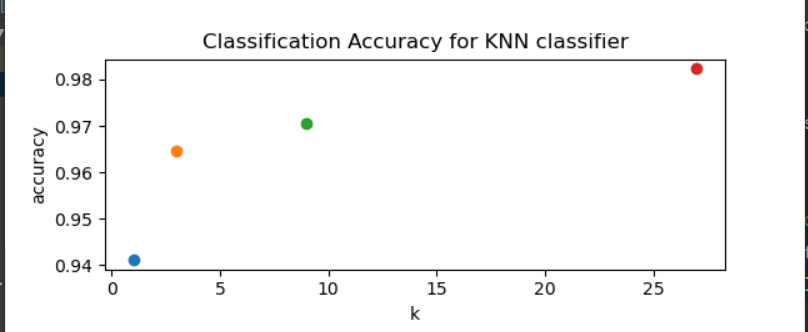
**קובץ פייתון KNN.py אשר טוען את ה data, מנרמל אותו, בונה מסווג KNN סטדנרטי עם k=9, ומדפיס את דיוק הסיווג על קבוצת המבחן.**

הדיוק שהתקבל: 0.9706, 97.06%.

## שאלה 9

**הציגו גרף המתאר את דיוק מסווג ה KNN על קבוצת המבחן כתלות בגודל k עבור הערכים {1, 3, 9, 27}.  
נתחו בקצרה את התוצאות שקיבלתם.**

הגרף שהתקבל:



הדיוק שהתקבל עבור k=1: 0.9412, 94.12%.  
הדיוק שהתקבל עבור k=3: 0.9647, 96.47%.  
הדיוק שהתקבל עבור k=9: 0.9706, 97.06%.  
הדיוק שהתקבל עבור k=27: 0.9824, 98.24%.

ניתוח התוצאות:

עבור k=1, קיבלנו את הדיוק הנמוך ביותר. הסיבה היא שהסיווג של כל דוגמאת מבחן התבצע על פי הדוגמא היחידה שהכי קרובה לה. שוב, ניתן לראות את אפקט ה overfitting.

ככל שהרשנו לעצמנו להתחשב במגוון של דוגמאות, הסיווג היה יותר אמין, והצלחנו לכסות על השגיאות הנובעות מסטיות שונות בין דוגמאות האימון לדוגמאות המבחן (תכונות לא רלוונטיות, דוגמאות רועשות וכו').

במקרה זה, העובדה ש k=27 נתן את התוצאה הטובה ביותר, מצביע אכן על כך שדוגמאות האימון שלנו אינן "מושלמות", או שיש תכונות לא רלוונטיות (ככל שמיצענו יותר קיבלנו תוצאה טובה יותר).

# שאלה תיאורטית

## שאלה 10

**נתונה קבוצת האימון הבאה, המורכבת מ 6 דוגמאות בעלות תכונה רציפה אחת –**



**נניח כי אנו בונים שני מסווגים: עץ ID3 לא גזום עם כלל אפסילון-החלטה, ומסווג KNN עם k=3.  
בשאלה זו נניח כי כאשר עץ ההחלטה קובע ערך לפיצול, הוא בוחר את ממוצע הערכים של שתי הדוגמאות הרלוונטיות. כמו כן, בשלב ההחלטה- אם יש שוויון, עץ ההחלטה בוחר סיווג בצורה רנדומלית. האם קיים ערך של המבטיח ששני המסווגים יסווגו את כל דוגמאות האימון בצורה זהה?  
אם כן, מצאו ערך כזה והראו זאת. אם לא, הסבירו למה לא.**

קיים אפסילון כנ"ל.

עץ ההחלטה של מסווג ה ID3 יראה בצורה הבאה:

**V = 9**

**+ + + +  
- -**

**V = 2.5**

**+ + +**

**+  
- -**

**- -**

**+**

תוצאות סיווג ה KNN עם k=3:

לכל נקבל סיווג חיובי.  
ב יש שוויון ולכן על פי נתוני השאלה הסיווג יהיה רנדומי.  
לכל נקבל סיווג שלילי.

עבור דוגמאות המבחן הסיווגים של דוגמאות המבחן יהיו באופן הבא:  
סיווג חיובי: עבור הדוגמאות עם הערכים 8, 10, 12, 13.  
סיווג שלילי: עבור הדוגמאות עם הערכים 2, 3.

מסווג ה ID3 לא גזום עם כלל אפסילון-החלטה:

על מנת שגם 2 וגם 3 יקבלו סיווג שלילי, על שתיהן להגיע לשני הבנים בפיצול השני, כלומר אפסילון לכל הפחות חצי.

נרצה לגרום לכך שרק הדוגמאות 3 ו 2 לא יגיעו לבן הימני בפיצול הראשון, זה יעבוד עבור אפסילון קטן מ 6 (כיון שזהו המרחק בין 9 שהוא ערך הסף ל 3 שהיא הקרובה אליו מבין השתיים).

הערך הכי קטן של אפסילון שניתן לבחור על מנת שכל הדוגמאות 8, 10, 12, 13 יקבלו סיווג חיובי הוא 1. הדוגמאות הכי קרובות ל 9 הן 8 ו 10, ומרחקן 1. עבור הדוגמאות 10, 12, 13 נקבל סיווג חיובי גם אם לא יהיו מספיק קרובות לערף הסף (כלומר עבור ערך אפסילון קטן) מאחר והן גדולות מ 9 ולכן יגיעו לבן הימני בפיצול הראשון ויקבלו סיווג חיובי כנדרש.  
עבור 8, אם נבחר אפסילון קטן מאחת, נגיע לבן הימני בפיצול השני (--) וסיווגה יהיה שלילי.

לסיכום, עבור נוכל להבטיח ששני המסווגים יסווגו את כל דוגמאות האימון בצורה זהה.

# מימוש המסווג המשולב

## שאלה 11

קובץ פייתון KNN\_epsilon.py.

הדיוק שהתקבל: 0.9353, 93.53%.