תרגיל בית 3 –מבוא ללמידה

מגישות:

לי בנימין **207502444**

שירה לביא 208607689

**חלק ב׳ - מבוא ללמידה (56 נק׳)**

[Writing Hand on Apple ](https://emojipedia.org/apple/ios-14.6/writing-hand/) **חלק א׳ – חלק היבש (28 נק')**

**k****NN – נעים להכיר**

בחלק זה תכירו אלגוריתם למידה בשם kNN, או בשמו המלא k-Nearest Neighbors, כאשר ה־k הוא למעשה פרמטר!  
יהי סט אימון עם ־דוגמות, , כאשר .  
כלומר הדוגמות הינן וקטורים ־ממדיים והתגיות הינן מדומיין כלשהו, הבעיה היא בעיית קלסיפיקציה (סיווג).  
אם לא נאמר אחרת, הקלסיפיקציה תהיה בינארית, כלומר .  
עבור כל דוגמה בסט האימון, ניתן להסתכל על הכניסה ה־ בווקטור כעל הfeature ה־ של הדוגמה, קרי כל דוגמה מיוצגת על ידי ־ערכים: .  
תהליך ה"אימון" של האלגוריתם הוא טריוויאלי – פשוט שומרים את סט האימון במלואו.  
תהליך הסיווג הוא גם פשוט למדי – כאשר רוצים לסווג דוגמה מסט המבחן מסתכלים על השכנים הקרובים ביותר שלה במישור ה־ממדי מבין הדוגמות בסט האימון, ומסווגים את הדוגמה על פי הסיווג הנפוץ ביותר בקרב השכנים.  
על מנת להימנע משוויון בין הסיווגים, נניח בדרך כלל כי ־אי זוגי, או שנגדיר היטב שובר שוויון.  
אם לא נאמר אחרת, במקרה של שוויון בקלסיפיקציה בינארית, נסווג את הדוגמה כחיובית .

שאלות הבנה

1. (3 נק') כאמור, בתהליך הסיווג אנו בוחרים עבור הדוגמה את הסיווג הנפוץ ביותר של השכנים הקרובים ביותר, אולם עלינו להגדיר את פונקציית המרחק עבור קביעת סט שכנים זה.  
   שתי פונקציות מרחק נפוצות הינן מרחק אוקלידי ומרחק מנהטן.
2. עבור איזה ערכים של נקבל שאין תלות בבחירה בין פונקציות המרחק הנתונות ~~בבחירה פונקציית המרחק~~? (נמקי)

עבור d=1 המרחק האוקלידי שווה למרחק מנהטן ולכן אין תלוית בבחירת פונקציית המרחק.

עבור K=מספר הדוגמאות בסט האימון לא תהיה תלות בפונקציית המרחק שכן הדוגמה תסווג לפי כל הדוגמאות, כלומר כל דוגמה תסווג לחיובית/שלילית בהתאם למספר הדוגמאות בסט האימון שהן רוב.

1. עבור בעיית קלסיפיקציה בינארית תנו דוגמה פשוטה לערכי , סט אימון ודוגמת מבחן בה השימוש בכל אחת מפונקציות המרחק הנ"ל משנה את סיווג דוגמה המבחן.

דוגמה: נבחר d=2 ו-K=1. כלומר לכל דוגמה יש 2 פיצ'רים והסיווג מתבצע על פי השכן הקרוב ביותר.   
סט אימון: סט מבחן:   
נסווג לפי מרחק מנהטן:

הדוגמה תסווג כשלילת כי הנקודה הקרובה ביותר היא .

נסווג לפי מרחק אוקלידי:

הפעם הדוגמה תסווג כחיובית, כי הנקודה הקרובה ביותר היא .

מעתה, אלא אם כן צוין אחרת, נשתמש במרחק אוקלידי.

נתונה קבוצת האימון הבאה, כאשר :



1. (1 נק') איזה ערך של עלינו לבחור על מנת לקבל את הדיוק המרבי **על קבוצת האימון**? מה יהיה ערך זה? (הדוגמא לא יכולה להיות שכנה של עצמה)  
   נבחר k=5 בכדי לקבל את הדיוק המירבי על קבוצת האימון. במצב כזה, כל הדוגמאות בפינה השמאלית העליונה יסווגו כשליליות, וכל הדוגמאות בפינה הימנית התחתונה יסווגו כחיוביות, ולכן הדיוק יהיה 10/14, כלומר יהיו טעויות בסיווג של ארבע נקודות מקבוצת האימון.

1. (1 נק') עבור איזה ערך של נקבל מסווג של קבוצת האימון? קרי כל דוגמת מבחן תקבל את הסיווג הנפוץ של כלל קבוצת האימון?

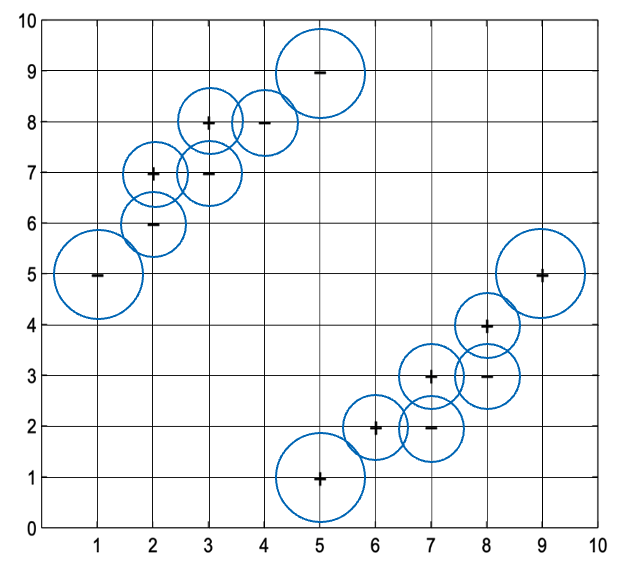
נקבל מסווג כזה עבור K בגודל בקבוצת האימון (פחות אחד אם מניחים כמו בסעיף הקודם שדוגמא לא יכולה להיות שכנה של עצמה). K השכנים הקרובים ביותר יהיו כל הדוגמאות.

1. (2 נק') נמקו מדוע שימוש בערכי גדולים או קטנים מדי יכול להיות גרוע עבור קבוצת הדגימות הנ״ל.

עבור k קטן מידי נוכל לקבל שגיאות על דגימות חריגות. למשל הנקודה (7,2) בדוגמה לעיל היא חריגה כיוון שהיא שלילית בסביבת דוגמאות חיובית, עבור k קטן מידי יכולנו לסווג את הדגימה מימינה כשלילית אף על פי שהיא חיובית, וזה בעצם לתת משקל גדול מידי למבנה הספציפי של קבוצת האימון כולל החריגות שבו (overfitting).

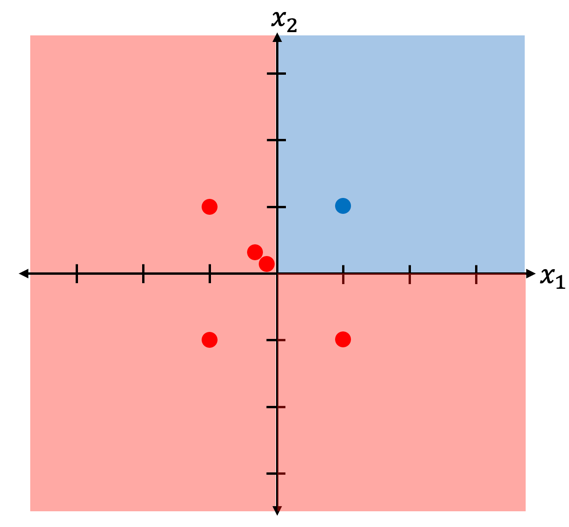
K גדול מידי זה המקרה מהסעיף הקודם, אנחנו צפויים להתחשב גם בשכנים לא קרובים ולתת להם משקל זהה לאלו שקרובים, ובכך אנחנו מאבדים את ההתאמה לדוגמאות והמסווג שנקבל ייתן החלטות כלליות מידי (underfitting).

1. (2 נק') שרטט את גבול ההחלטה של 1-nearest neighbor עבור הגרף.

לכל נקודה נבחר את הסיווג של השכן שחותך את המעגל שסביבה בחפיפה הכי גדולה.

**השוואה בין מודלי למידה – יש לנמק בקצרה את הפתרונות**

1. (3 נק') הציגו מסווג מטרה וקבוצת אימון בעלת לכל היותר 10 דוגמאות כך שלמידת עץ ID3 תניב מסווג אשר עונה נכון עבור כל דוגמת מבחן אפשרית (כלומר יתקבל מסווג המטרה), אך למידת KNN תניב מסווג שעבורו קיימת לפחות דוגמת מבחן אחת עליה הוא יטעה, לכל ערך K שייבחר.

נגדיר את מסווג המטרה הבא: לכל דוגמה  *המסווג יקיים:*

נבחן את סט האימון הבא:

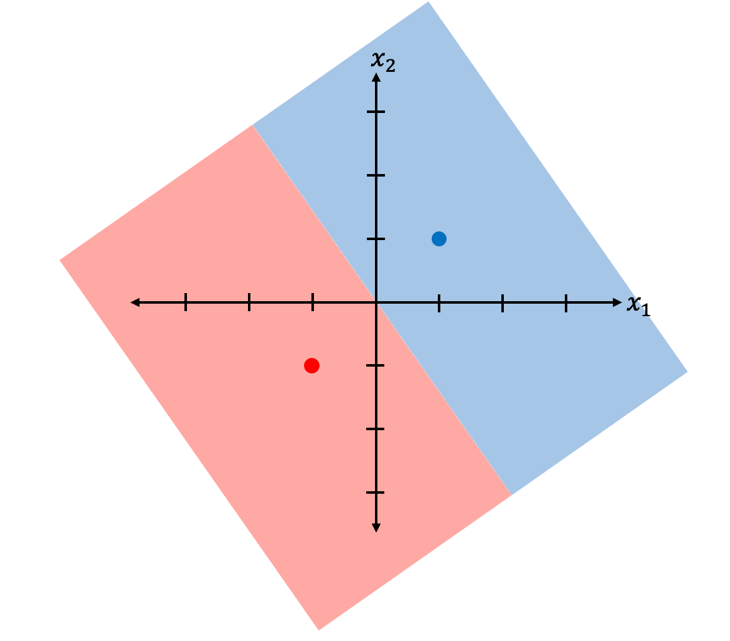
|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
| 0 |  |  |
| 0 |  |  |
| 0 |  |  |
|  |  |  |
|  |  |  |

*עץ ההחלטה שמפצל לפי האם ולאחר מכן לפי הפיצ'ר השני, האם הוא עץ אפשרי לקבל, זה בעצם עץ שמתנהג לגמרי כמו מסווג המטרה ולכן יהיה צודק על כל דוגמת מבחן. נראה שקיימת דוגמת מבחן עליה* KNN עם *סט האימון הזה* *יטעה.*

*עבור דוגמת המבחן מסווג המטרה יסווג את הדוגמה כ-1, עץ ההחלטה גם יסווג כ-1 ולכן יענה נכון. לעומת זאת,* KNN *יסווג את הדוגמה כ-0 לכל* k *ולכן יטעה עליה:*

* *עבור* k=1:  *השכן הקרוב ביותר מקבוצת האימון הוא , שמסווג כ-0 ולכן הסיווג שייבחר הוא 0.*
* *עבור* k=2: *שני השכנים הקרובים ביותר הם , שמסווגים 0 ולכן הסיווג שייבחר הוא 0.*
* *עבור כל* k *גדול מזה, תמיד תהיה במקסימום רק נקודה אחת שמסווגת 1 (כי יש רק אחת כזו בקבוצת האימון) ולכן הנקודה הזו תמיד תהיה במיעוט בכל קבוצת שכנים והסיווג שייבחר תמיד יהיה 0.*

1. (3 נק') הציגו מסווג מטרה וקבוצת אימון בעלת לכל היותר 10 דוגמאות כך שלמידת מסווג KNN עבור ערך K מסוים תניב מסווג אשר עונה נכון עבור כל דוגמת מבחן אפשרית (כלומר יתקבל מסווג המטרה), אך למידת עץ ID3 תניב מסווג אשר עבורו קיימת לפחות דוגמת מבחן אפשרית אחת עליה הוא יטעה.

נגדיר את מסווג המטרה הבא: לכל דוגמה  *המסווג יקיים:*

נבחן את סט הדוגמאות הבא:

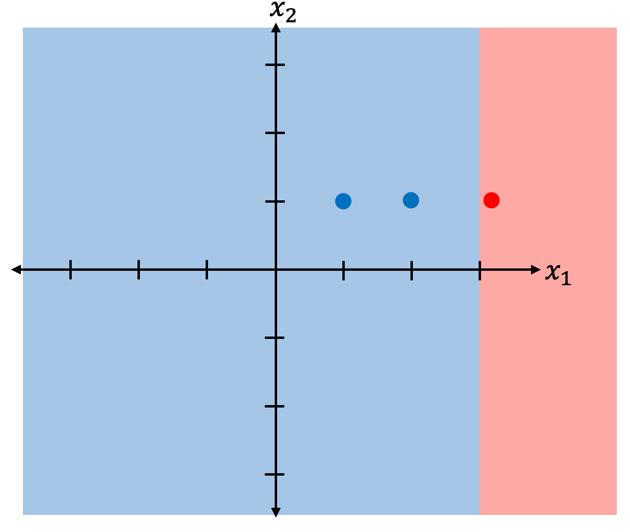
|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
| 0 |  |  |

*עץ החלטה שאפשר יהיה לקבל יהיה בעל פיצול בודד לפי או ולכן הוא יטעה למשל על הדוגמה .*

*לעומת זאת,* KNN *יהיה צודק לכל נקודת מבחן. כל נקודה שעבורה תהיה יותר קרובה לנקודה ולכן תסווג כ1, ואם תנאי זה לא מתקיים עבורה היא תהיה קרובה יותר לנקודה ותסווג כ0. כלומר, לכל נקודה ה* KNN*יסווג נכון.  
נראה מדוע כל נקודה שעבורה שעבורה תהיה יותר קרובה לנקודה וכל נקודה שלא מקיימת את התנאי תהיה יותר קרובה ל :*

אם מתקיים אז ולכן הנקודה קרובה יותר ותיבחר. אם מתקיים אז קורה המצב ההפוך ו והנקודה קרובה יותר ולכן תיבחר.

1. (3 נק') הציגו מסווג מטרה וקבוצת אימון בעלת לכל היותר 10 דוגמאות כך שלמידת מסווג KNN עבור ערך K מסוים תניב מסווג אשר עבורו קיימת לפחות דוגמת מבחן אפשרית אחת עליה הוא יטעה, וגם למידת עץ ID3 תניב מסווג אשר עבורו קיימת לפחות דוגמת מבחן אחת אפשרית עליה הוא יטעה.

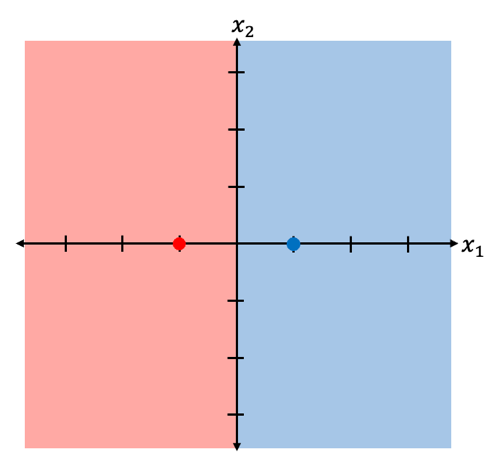
נגדיר את מסווג המטרה הבא: לכל דוגמה  *המסווג יקיים:*

נבחן את סט הדוגמאות הבא:

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

*עץ החלטה שאפשר יהיה לקבל יהיה בעל פיצול אחד לפי כאשר כל מי שמקיים תנאי זה יסווג כ1 וכל מי שלא יסווג כ0. עץ כזה יטעה על הדוגמא , שאמורה להיות מסווגת לפי מסווג המטרה אבל תסווג כ. בנוסף, דוגמת המבחן תסווג כ על ידי מסווג* KNN *עם* k=3 *שכנים (זה בעצם סיווג רוב) וגם זה יהיה טעות.*

1. (3 נק') הציגו מסווג מטרה וקבוצת אימון בעלת לכל היותר 10 דוגמאות כך שלמידת מסווג KNN עבור ערך K מסוים תניב מסווג אשר עונה נכון עבור כל דוגמת מבחן אפשרית (כלומר יתקבל מסווג המטרה), וגם למידת עץ ID3 תניב מסווג עונה נכון עבור כל דוגמת מבחן אפשרית (כלומר יתקבל מסווג המטרה).

נגדיר את מסווג המטרה הבא: לכל דוגמה  *המסווג יקיים:*

נבחן את סט הדוגמאות הבא:

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | 0 |  |
|  |  |  |

*עץ החלטה שאפשר יהיה לקבל יפצל לפי , וזה יהיה בדיוק כמו מסווג המטרה ולכן יסווג כל דוגמא נכון.*

*גם מסווג* KNN *יסווג כל דוגמא נכון עבור* k=1, *מכיוון שכל דוגמא שנמצאת ברביע השני והשלישי (כל דוגמא עבורה תהיה יותר קרובה לנקודה מאשר לנקודה ולכן תסווג כ0, וכל דוגמא שעבורה תהיה קרובה יותר לנקודה ולכן תסווג כ1.*

**מתפצלים ונהנים**

(7 נק') כידוע, בעת סיווג של דוגמת מבחן על ידי עץ החלטה, בכל צומת בעץ אנו מחליטים לאיזה צומת בן להעביר את דוגמת המבחן על ידי ערך סף שמושווה לfeature של הדוגמה. לפעמים ערך הסף קרוב מאוד לערך הfeature של דוגמת המבחן. היינו רוצים להתחשב בערכים "קרובים" לערך הסף בעת סיווג דוגמת מבחן, ולא לחרוץ את גורלה של הדוגמה לתת־עץ אחד בלבד; לצורך כך נציג את האלגוריתם הבא:  
  
יהיו עץ החלטה , דוגמת מבחן , ווקטור המקיים .  
כלל אפסילון־החלטה שונה מכלל ההחלטה הרגיל שנלמד בכיתה באופן הבא:   
נניח שמגיעים לצומת בעץ המפצל לפי ערכי התכונה , עם ערך הסף .  
אם מתקיים  אזי ממשיכים **בשני** המסלולים היוצאים מצומת זה, ואחרת ממשיכים לבן המתאים בדומה לכלל ההחלטה הרגיל. לבסוף, מסווגים את הדוגמה *בהתאם לסיווג הנפוץ ביותר של הדוגמאות הנמצאות בכל העלים אליהם הגענו במהלך הסיור על העץ (במקרה של שוויון – הסיווג ייקבע להיות ).*

יהא עץ החלטה לא גזום, ויהא העץ המתקבל מ־ באמצעות גיזום מאוחר שבו הוסרה הרמה התחתונה של (כלומר כל הדוגמות השייכות לזוג עלים אחים הועברו לצומת האב שלהם).  
הוכיחו\הפריכו: **בהכרח** קיים ווקטור כך שהעץ עם כלל אפסילון־החלטה והעץ עם כלל ההחלטה הרגיל יסווגו כל דוגמת מבחן ב בצורה זהה.

הפרכה.

נתבונן בעץ הבא:

(+++,--)

+++

--

לאחר גיזום העץ וקבלת , נישאר עם שתי דוגמאות חיוביות ודוגמא אחת שלילית ולכן העץ יסווג את כל דוגמאות המבחן כחיוביות.

נניח בשלילה כי קיים וקטור כך שהעץ עם כלל אפסילון־החלטה והעץ עם כלל ההחלטה הרגיל יסווגו כל דוגמת מבחן ב בצורה זהה.

עבור דוגמת מבחן *העץ T מקיים*  *ולכן נמשיך בתהליך הסיווג לפי כלל ההחלטה הרגיל ונעבור לבן הימני. הדוגמה תסווג כמינוס. לעומת זאת, כאמור העץ יסווג את הדוגמא כפלוס.*

*הגענו לסתירה להנחת השלילה לכן לא קיים וקטור כנ"ל.*

**חלק ב׳ - היכרות עם הקוד  
רקע**

חלק זה הוא רק עבור היכרות הקוד, עבורו עליו במלואו ווודאו כי הינכם מבינים את הקוד.  
בחלק של הלמידה, נעזר ב 𝑑𝑎𝑡𝑎𝑠𝑒𝑡, הדאטה חולק עבורכם לשתי קבוצות: קבוצת אימון train.csv וקבוצת מבחן test.csv.   
ככלל, קבוצת האימון תשמש אותנו לבניית המסווגים, וקבוצת המבחן תשמש להערכת ביצועיהם.

בקובץ utils.py תוכלו למצוא את הפונקציות הבאות לשימושכם:  
 load\_data\_set, create\_train\_validation\_split, get\_dataset\_split   
אשר טוענות/מחלקת את הדאטה בקבצי ה־csv למערכי np.array (קראו את תיעוד הפונקציות).

הדאטה של ID3 עבור התרגיל מכיל מדדים שנאספו מצילומים שנועדו להבחין בין גידול שפיר לגידול ממאיר. כל דוגמה מכילה 30 מדדים כאלה, ותווית בינארית **diagnosis** הקובעת את סוג הגידול (0=שפיר, 1=ממאיר). כל התכונות (מדדים( רציפות. העמודה הראשונה מציינת האם האדם חולה (M) או בריא (B). שאר העמודות מציינות כל תכונות רפואיות שונות של אותו אדם (התכונות מורכבות ואינכם צריכים להתייחס למשמעות שלהן כלל).

תיקיית :

* תיקיה זו אלו מכילה את קבצי הנתונים עבור .

קובץ :

* קובץ זה מכיל פונקציות עזר שימושיות לאורך התרגיל, כמו טעינה של וחישוב הדיוק.
* בחלק הבא יהיה עליכם לממש את הפונקציה . קראו את תיעוד הפונקציות ואת ההערות הנמצאות תחת התיאור TODO.

קובץ :

* קובץ בדיקה בסיסי שיכול לעזור לכם לבדוק את המימוש.

קובץ :

* קובץ זה מכיל 3 מחלקות שימושית לבניית עץ שלנו.
  + המחלקה : מחלקה זו מממשת הסתעפות של צומת בעץ. היא שומרת את התכונה ואת הערך שלפיהם מפצלים את הדאטה שלנו.
  + המחלקה : מחלקה זו מממשת צומת בעץ ההחלטה.   
    הצומת מכיל שאלה ואת שני הבנים כאשר הוא הענף בחלק של הדאטה שעונה על שאלת הצומת   
    (הפונקציה של ה מחזירה ).  
    ו־ הוא הענף בחלק של הדאטה שעונה על שאלת הצומת   
    (הפונקציה של ה מחזירה ).
  + המחלקה : מחלקה זו מממשת צומת שהוא עלה בעץ ההחלטה. העלה מכיל לכל אחד מהמחלקות בדאטה את מספר הדוגמאות בעלה עבור כל מחלקה (למשל: ).

קובץ :

* קובץ זה מכיל את המחלקה של שתצטרכו לממש חלקים ממנה, עיינו בהערות ותיעוד המתודות.

קובץ :

* קובץ הרצת הניסויים של ID3, הקובץ מכיל את הניסויים הבאים, שיוסברו בהמשך:

**חלק ג׳ – חלק רטוב ID3 (28 נק')**

עבור חלק זה מותר לכם להשתמש בספריות הבאות:

All the built in packages in python, sklearn, pandas ,numpy, random, matplotlib, argparse, abc, typing.

**אך כמובן שאין להשתמש באלגוריתמי הלמידה, או בכל אלגוריתם או מבנה נתונים אחר המהווה חלק מאלגוריתם למידה אותו תתבקשו לממש.**

1. (3 נק') השלימו את הקובץ utils.py ע"י מימוש הפונקציה .   
   קראו את תיעוד הפונקציה ואת ההערות הנמצאות תחת התיאור.  
   מימשנו  
   (הריצו את הטסטים המתאימים בקובץ *לוודא שהמימוש שלכם נכון).*  
   שימו לב! בתיעוד ישנן הגבלות על הקוד עצמו, אי־עמידה בהגבלות אלו תגרור הורדת נקודות.  
   בנוסף, שנו את ערך ה בתחילת הקובץ מ־ למספר תעודת הזהות של אחד מהמגישים.
2. (10 נק') **אלגוריתם ID3:**
   1. השלימו את הקובץ ID3.py ובכך ממשו את אלגוריתם כפי שנלמד בהרצאה.   
      שימו לב שכל התכונות רציפות. אתם מתבקשים להשתמש בשיטה של חלוקה דינמית המתוארת בהרצאה. כאשר בוחנים ערך סף לפיצול של תכונה רציפה, דוגמאות עם ערך השווה לערך הסף משתייכות לקבוצה עם הערכים הגדולים מערך הסף. במקרה שיש כמה תכונות אופטימליות בצומת מסוים בחרו את התכונה בעלת האינדקס המקסימלי.   
      כלל המימוש הנ"ל צריך להופיע בקובץ בשם , באזורים המוקצים לכך.  
      (השלימו את הקוד החסר אחרי שעיינתם והפנמתם את הקובץ ואת המחלקות שהוא מכיל).

מימשנו

* 1. ממשו את שנמצאת ב

[Writing Hand on Apple ](https://emojipedia.org/apple/ios-14.6/writing-hand/)*והריצו את החלק המתאים ב* ציינו בדו"ח את הדיוק שקיבלתם.

קיבלנו דיוק של 96.12%

1. **גיזום מוקדם.**

פיצול צומת מתקיים כל עוד יש בו יותר דוגמאות מחסם המינימום ,𝑚כלומר בתהליך בניית העץ מבוצע "גיזום מוקדם" כפי שלמדתם בהרצאות. שימו לב כי פירוש הדבר הינו שהעצים הנלמדים אינם בהכרח עקביים עם הדוגמאות. לאחר סיום הלמידה (של עץ יחיד), הסיווג של אובייקט חדש באמצעות העץ שנלמד מתבצע לפי רוב הדוגמאות בעלה המתאים.

* 1. (2 נק') [Writing Hand on Apple ](https://emojipedia.org/apple/ios-14.6/writing-hand/)הסבירו מה החשיבות של הגיזום באופן כללי ואיזה תופעה הוא מנסה למנוע?

גיזום עוזר למנוע overfitting. בבניית עץ החלטה נוצר מצב שהעץ תואם לדוגמאות האימון שלנו ונרצה לקבל מודל שמכליל באופן טוב ולהימנע מתופעת ה-overfitting. הגיזום עוזר לנו כיוון שהוא מאפשר לנו להתעלם מהדוגמאות החריגות ולהכליל את המודל.

* 1. (3 נק') **עדכנו** את המימוש בקובץ כך שיבצע גיזום מוקדם כפי שהוגדר בהרצאה.   
     הפרמטר מציין את המספר המינימלי בעלה לקבלת החלטה, קרי יבוצע גיזום מוקדם אם ורק אם מספר הדוגמות בצומת קטן שווה לפרמטר הנ"ל.   
     מימשנו
  2. (8 נק') **שימו לב, זהו סעיף יבש ואין צורך להגיש את הקוד שכתבתם עבורו.**בצעו כיוונון לפרמטר על קבוצת האימון:

1. בחרו לפחות חמישה ערכים שונים לפרמטר .

2. עבור כל ערך, חשבו את הדיוק של האלגוריתם על ידי  על קבוצת האימון בלבד.   
כדי לבצע את חלוקת קבוצת האימון ל- קבוצות יש להשתמש בפונקציה [sklearn.model\_selection.KFold](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html" \l "sklearn.model_selection.KFold" \o "sklearn.model_selection)עם הפרמטרים ,   
ו־ אשר שווה למספר תעודת הזהות של אחד מהשותפים.

להלן התוצאות שקיבלנו על קבוצת האימון:

|  |  |
| --- | --- |
| Valid accuracy | M |
| 96.14 | 10 |
| 93.24 | 20 |
| 91.63 | 30 |
| 90.66 | 40 |
| 90.66 | 50 |
| 90.66 | 70 |
| 90.66 | 80 |

* + 1. [Writing Hand on Apple ](https://emojipedia.org/apple/ios-14.6/writing-hand/)השתמשו בתוצאות שקיבלתם כדי ליצור גרף המציג את השפעת הפרמטר על הדיוק.   
       צרפו את הגרף בדו״ח. (לשימושכם הפונקציה בתוך הקובץ ).



* + 1. [Writing Hand on Apple ](https://emojipedia.org/apple/ios-14.6/writing-hand/) הסבירו את הגרף שקיבלתם. לאיזה גיזום קיבלתם התוצאה הטובה ביותר ומהי תוצאה זו?

בגרף ניתן לראות כי התוצאה הטובה ביותר על סט הולידציה הייתה עבור פרמטר סף לגיזום של 10 דוגמאות. באופן כללי הגרף מציג את אחוז הדיוק שהתקבל על סט הולידציה כתלות בפרמטר הגיזום, והגיוני לראות שאחוזי הדיוק יורדים ככל שגוזמים יותר כי המודל הופך להיות פחות ספציפי, פחות מותאים באמת לאופי הדוגמאות שלנו. ככל שגוזמים יותר נוצר עץ רדוד יותר עם פחות כוח החלטה באמת, זה יותר מצב של *underfitting*.

* 1. [Writing Hand on Apple ](https://emojipedia.org/apple/ios-14.6/writing-hand/)(2 נק׳) השתמשו באלגוריתם ID3 עם הגיזום המוקדם כדי ללמוד מסווג מתוך **כל** קבוצת האימון ולבצע חיזוי על קבוצת המבחן.   
     השתמשו בערך ה־ האופטימלי שמצאתם בסעיף c. (ממשו שנמצאת ב  *והריצו את החלק המתאים ב ).* ציינו בדו"ח את הדיוק שקיבלתם. האם הגיזום שיפר את הביצועים ביחס להרצה ללא גיזום?

מימשנו את הפונקציה. הערך הטוב ביותר שקיבלנו הוא 10 ועבורו קיבלנו על סט המבחן דיוק של 96.12. דיוק זה זהה לדיוק ללא גיזום, וזה כנראה נובע מזה שכמות הדוגמאות בעלים כבר גדולה מ10 ואז הפרמטר שלנו לא השפיע.

בחנו גם את ערך גיזום של 40 בשביל לראות איך זה משפיע על דוגמאות המבחן, ואכן עבורו קיבלנו תוצאה של 98.06. דיוק זה גבוהה מהדיוק ללא הגיזום, וזה מצביע על כך שעבור 10 או לא לגזום בכלל מקבלים עץ שהוא מותאם מידי לדוגמאות האימון, כלומר זה מקרה של overfitting.