תרגיל בית 3 –מבוא ללמידה

206458390 - 315173344

חלק ב’ - מבוא ללמידה (56 נק’)

✍️חלק א’ – חלק היבש (28 נק')

kNN – נעים להכיר

שאלות הבנה

א. (3 נק') כאמור, בתהליך הסיווג אנו בוחרים עבור הדוגמה את הסיווג הנפוץ ביותר של k השכנים הקרובים ביותר, אולם עלינו להגדיר את פונקציית המרחק עבור קביעת סט שכנים זה.
שתי פונקציות מרחק נפוצות הינן מרחק אוקלידי ומרחק מנהטן.

1) עבור איזה ערכים של d, k נקבל שאין תלות בבחירה בין פונקציות המרחק הנתונות? (נמקי)

עבור d=1, אין תלות בבחירה בין פונקציות המרחק לכל k, שכן במימד אחד המרחק על גבי ציר x (מרחק מנהטן בחד-מימד) זהה למרחק האוקלידי.
עבור $k = n$ (n מספר הדגימות), לכל d הבחירה נעשית ע"י "הרוב קובע" ואין משמעות לפונקציית המרחק.
עבור d>1, לכל $k < n$ יש תלות בבחירה שכן:
עבור מימד d=2, בהינתן הנקודות a=(1,1) והנקודה b=(1.5,0), ניתן לראות שעבור הנקודה (0,0) בהינתן פונקציית מרחק אוקלידית השכן הקרוב ביותר אליה הוא a (מרחק $\sqrt{2}$ לעומת 1.5) ובהינתן פונקציית מרחק מנהטן השכן הקרוב ביותר אליה הוא b. ניתן להכליל דוגמה זו לכל מימד d>1 ולכל k<n.

2) עבור בעיית קלסיפיקציה בינארית תנו דוגמה <u>פשוטה</u> לערכי d, k, סט אימון ודוגמת מבחן בה השימוש בכל אחת מפונקציות המרחק הנ"ל משנה את סיווג דוגמה המבחן.

דוגמה:
עבור מודל 1-NN (k=1):
נגדיר $x_i \in \mathbb{R}^2, y_i \in \mathcal{Y}$

$$D = \{(x_1 = (1,1), +), (x_2 = (1.5,0), -)\}$$
$$\mathcal{Y} = \{+, -\}$$

דוגמת מבחן: $x = (0,0)$

<u>לפי פונקציית מנהטן:</u>

$$d_{man}(x_1, x) = 2$$
$$d_{man}(x_2, x) = 1.5$$

השכן הקרוב ביותר הוא x_2 ולכן: $y = y_2 = -$

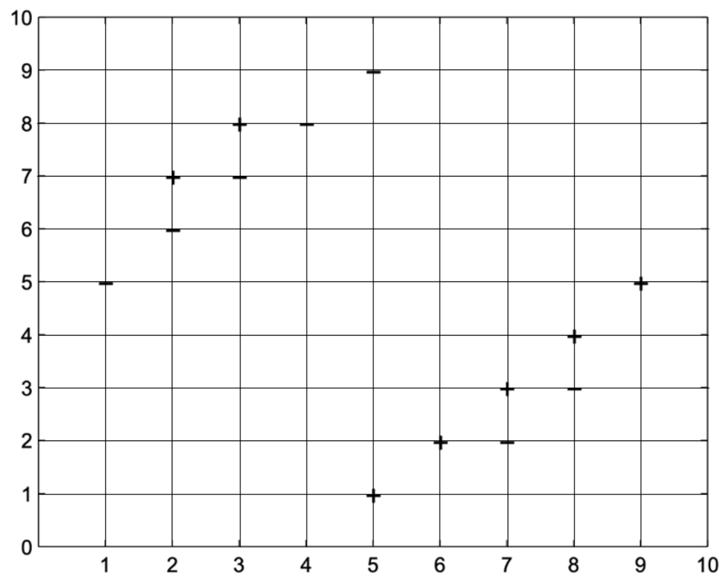
לפי פונקציה אוקלידית:

$$d_{man}(x_1, x) = \sqrt{2} \approx 1.4$$
$$d_{man}(x_2, x) = 1.5$$

השכן הקרוב ביותר הוא x_1 ולכן: $y = y_2 = +$

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

נתונה קבוצת האימון הבאה, כאשר $d = 2$:



3) (1 נק') איזה ערך של k עלינו לבחור על מנת לקבל את הדיוק המרבי **על קבוצת האימון**? מה יהיה ערך זה? (הדוגמא לא יכולה להיות שכנה של עצמה)

עבור k=5, נקבל דיוק של $\frac{10}{14}$ מהנקודות, כלומר 71.43%.

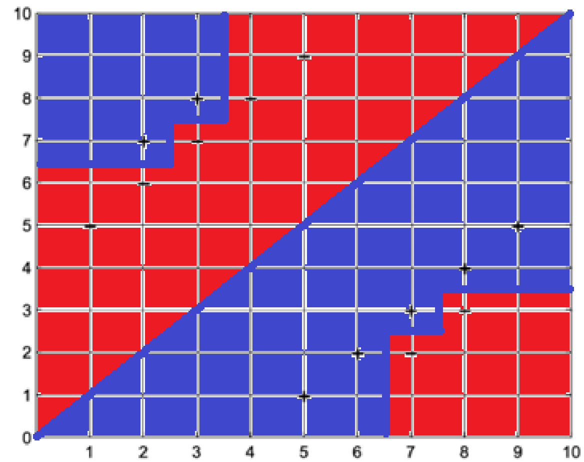
4) (1 נק') עבור איזה ערך של k נקבל מסווג $majority$ של קבוצת האימון? קרי כל דוגמת מבחן תקבל את הסיווג הנפוץ של כלל קבוצת האימון?

עבור k=14.

5) (2 נק') נמקו מדוע שימוש בערכי k גדולים או קטנים מדי יכול להיות גרוע עבור קבוצת הדגימות הנ"ל.

ניתן לראות בדומה הנתונה שהמשולש השמאלי עליון מכיל יותר נקודות שליליות, והמשולש התחתון ימני מכיל יותר נקודות חיוביות. לכן, שימוש ב k גדול מידי יתעלם מתכונה זו ויבחר רק ע"פ הרוב בכל המישור, ושימוש ב k נמוך מידי עלול לסווג ע"פ נקודות שאינן מייצגות (שעלולות להיות outliers או רעש). באופן כללי ניתן לומר שערכי k גדולים מידי יגרמו ל underfitting וערכי k קטנים מידי לoverfitting.

6) (2 נק') שרטט את גבול ההחלטה של 1-nearest neighbor עבור הגרף.
אדום = סיווג –
כחול = סיווג +



<u>השוואה בין מודלי למידה – יש לנמק בקצרה את הפתרונות</u>

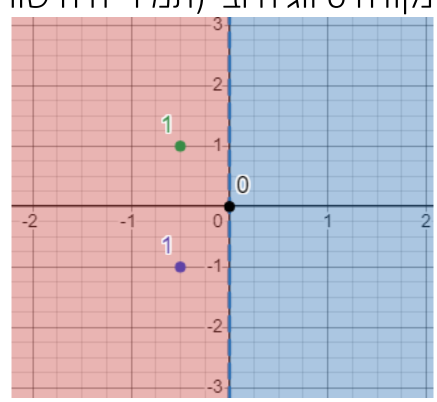
1) (3 נק') הציגו מסווג מטרה $f(x): R^2 \to \{0,1\}$ וקבוצת אימון בעלת לכל היותר 10 דוגמאות כך שלמידת עץ ID3 תניב מסווג אשר עונה נכון עבור כל דוגמת מבחן אפשרית (כלומר יתקבל מסווג המטרה), אך למידת KNN תניב מסווג שעבורו קיימת לפחות דוגמת מבחן אחת עליה הוא יטעה, לכל ערך K שייבחר.

$$f\big((a,b)\big) = \begin{cases} 1 & \text{a} < 0 \\ 0 & o.w \end{cases}$$

* sign(0)=+
(כלומר הסימן ההפוך לסימן הערך לפי ציר x).

$$D = \{\big((-0.5,1),1\big),\big((-0.5,-1),1\big),\big((0,0),0\big)\}$$

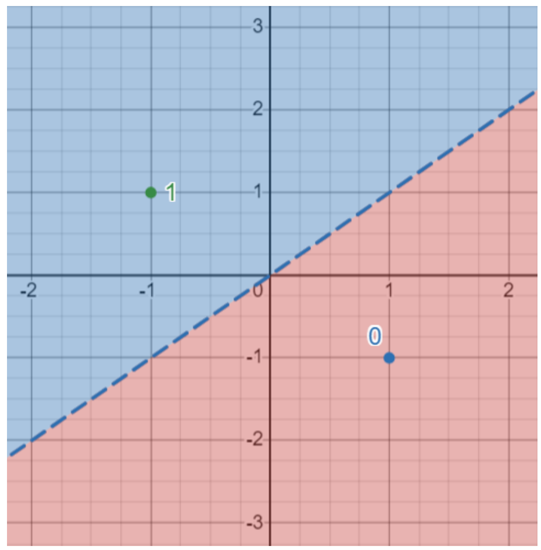
עץ החלטה ID3 (תחת דיסקריטיזציה של מרווחים בגודל 1) יפריד בדיוק לפי פונקציית המטרה (כמו בציור המצורף), שכן רק הפרדה לפי ציר x באופן הזה תיתן אנטרופיה אפס.
לעומת זאת מודל 1-NN לא ייצור הפרדה לינארית מאונכת לציר x, וכל מודל k-NN עם k>1 ייתן לכל נקודה סיווג חיובי (תמיד יהיה שוויון או רוב ל1ים – במקרה של שוויון ייבחר 1).



2) (3 נק') הציגו מסווג מטרה $f(x): R^2 \to \{0,1\}$ 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.

$$f\big((a,b)\big) = \begin{cases} 1 & b \geq a \\ 0 & o.w \end{cases}$$
$$D = \{\big((1,-1),0\big), \big((-1,1),1\big)\}$$

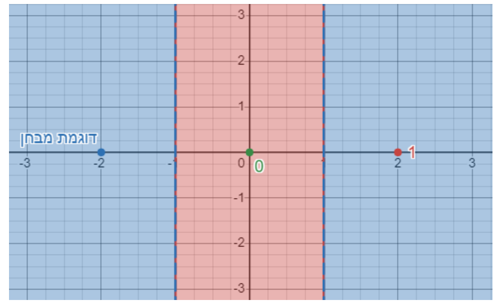
מסווג 1-NN (k=1) יסווג לפי כלל ההחלטה (במקרה של שוויון מסווג 1) משום שההחלטה נעשית ע"י מרחק אוקלידי בין שתי נקודות האימון).
לעומת זאת, עץ החלטה מבוסס ID3 יחלק באופן שמקביל לאחד הצירים ויגיע לאנטרופיה אפס בשני העלים ויעצור. כלל ההחלטה שלא זהה ל f ולכן קיימת בהכרח לפחות נקודת מבחן אחת עבורה הסיווג שגוי.



3) (3 נק') הציגו מסווג מטרה $f(x): R^2 \to \{0,1\}$ 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.

$$f(a,b) = \begin{cases} 0 & -1 \leq a \leq 1 \\ 1 & o.w \end{cases}$$
$$D = \{\big((0,0),0\big), \big((2,0),1\big)\}$$

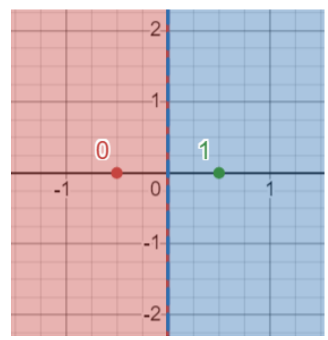
ניתן לראות כי במקרה זה מודל 1-NN ייצור קו הפרדה ב x=1, מודל עץ החלטה ID3 (תחת דיסקריטיזציה של מקטעים בגודל 1) ייצור קו הפרדה מקביל לציר y בין 0 ל 2. שניהם יסווגו את נקודת המבחן כ-0, אך מסווג המטרה יסווג אותה כ-1.



4) (3 נק') הציגו מסווג מטרה $f(x): R^2 \rightarrow \{0,1\}$ 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).

$$f(a,b) = \begin{cases} 0 & a < 0 \\ 1 & o.w \end{cases}$$
$$D = \{((0.5,0),1),((-0.5,0),0)\}$$

מודל 1-NN ייצור הפרדה לפי ציר ה y, כאשר במקרה של שוויון במרחקים בין שני שכנים בוחר את השכן החיובי.
מודל עץ החלטה ID3 עם דיסקריטיזציה יבחר את החלוקה הראשונה לפי ציר x באפס (האופציה היחידה לקבל אנתרופיה אפס) ויסיים. שני המודלים זהים למסווג המטרה כנדרש.



מתפצלים ו~~נהנים~~

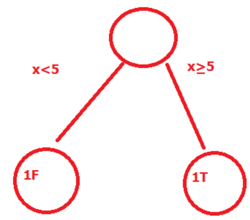
(7 נק') כידוע, בעת סיווג של דוגמת מבחן על ידי עץ החלטה, בכל צומת בעץ אנו מחליטים לאיזה צומת בן להעביר את דוגמת המבחן על ידי ערך סף v שמושווה לfeature של הדוגמה. לפעמים ערך הסף קרוב מאוד לערך הfeature של דוגמת המבחן. היינו רוצים להתחשב בערכים "קרובים" לערך הסף בעת סיווג דוגמת מבחן, ולא לחרוץ את גורלה של הדוגמה לתת־עץ אחד בלבד; לצורך כך נציג את האלגוריתם הבא:

יהיו עץ החלטה T, דוגמת מבחן $x \in \mathbb{R}^d$, ווקטור $\varepsilon \in \mathbb{R}^d$ המקיים $\forall i \in [1,d]: \varepsilon_i > 0$.
כלל אפסילון־החלטה שונה מכלל ההחלטה הרגיל שנלמד בכיתה באופן הבא:
נניח שמגיעים לצומת בעץ המפצל לפי ערכי התכונה i, עם ערך הסף v_i.
אם מתקיים $|x_i - v_i| \le \varepsilon_i$ אזי ממשיכים **בשני** המסלולים היוצאים מצומת זה, ואחרת ממשיכי לבן המתאים בדומה לכלל ההחלטה הרגיל. לבסוף, מסווגים את הדוגמה x בהתאם לסיווג הנפוץ ביותר של הדוגמאות הנמצאות בבכל העלים אליהם הגענו במהלך הסיור על העץ (במקרה של שוויון – הסיווג ייקבע להיות $True$).

יהא T עץ החלטה לא גזום, ויהא T' העץ המתקבל מ־T באמצעות גיזום מאוחר שבו הוסרה הרמה התחתונה של T (כלומר כל הדוגמות השייכות לזוג עלים אחים הועברו לצומת האב שלהם).
הוכיחו\הפריכו: **בהכרח** קיים ווקטור ε כך שהעץ T עם כלל אפסילון־החלטה והעץ T' עם כלל ההחלטה הרגיל יסווגו כל דוגמת מבחן ב$\mathbb{R}^d$ בצורה זהה.

הפרכה:

נביט על דוגמה פשוטה בממד יחיד (x ו- ϵ הם סקלרים).
יהיה עץ החלטה T עם חלוקה יחידה בשורש ע"פ ה-threshold הבא: $x \geq 5$. בכל אחד מהעלים דוגמה יחידה, בעלה הימני T ובעלה השמאלי F.



ניתן לראות כי לאחר גיזום, נישאר עם עלה בודד כך שכל דוגמה מקבלת את הסיווג T (לפי הסיווג של רוב הנקודות, במקרה הזה עם שובר שוויון).

נניח בשלילה שקיים ϵ שעבורו T מסווג כל דוגמת מבחן ב $\mathbb{R}^1$ בצורה זהה ל T', כלומר את כל הדוגמאות ב-T. נבחן את דוגמת המבחן $(5-2\epsilon)$, היא רחוקה ביותר מ ϵ מה *threshold* ולכן תסווג ל F בסתירה להנחת השלילה.

חלק ג' – חלק רטוב ID3 (28 נק')

1. .

2. (10 נק') אלגוריתם ID3:

a. .

b. ממשו את $basic\_experiment$ שנמצאת ב $ID3\_experiments.py$ TODO והריצו את החלק המתאים ב $main$ ציינו בדו"ח את הדיוק שקיבלתם. ✍

Test Accuracy: 94.17%

3. גיזום מוקדם.

✍ **a.** (2 נק') הסבירו מה החשיבות של הגיזום באופן כללי ואיזה תופעה הוא מנסה למנוע?

גיזום מונע מאיתנו לקבל עץ שהוא overfitted לקבוצת האימון, בכך להוריד את הדיוק של המודל על קבוצת האימון במטרה להגדיל את הדיוק על קבוצת המבחן.

b. .

c. (8 נק') **שימו לב, זהו סעיף יבש ואין צורך להגיש את הקוד שכתבתם עבורו.**
בצעו כיוונון לפרמטר M על קבוצת האימון.

✍ i. השתמשו בתוצאות שקיבלתם כדי ליצור גרף המציג את השפעת הפרמטר M על הדיוק. צרפו את הגרף בדו"ח. (לשימושכם הפונקציה $util\_plot\_graph$ בתוך הקובץ $utils.py$).

✍ ii. הסבירו את הגרף שקיבלתם. לאיזה גיזום קיבלתם התוצאה הטובה ביותר ומהי תוצאה זו?

ניתן לראות כי ערך m האופטימלי הוא 5, כלומר התוצאה האופטימלית על קבוצת הואלידציה מתקבלת כאשר מגבילים את כמות הדוגמאות הדרושות על מנת לבצע חלוקה של צומת ל-5 דוגמאות.

אחוז הדיוק שהתקבל הוא:92.90%

✍ **d.** (2 נק’) השתמשו באלגוריתם ID3 עם הגיזום המוקדם כדי ללמוד מסווג מתוך **כל** קבוצת האימון ולבצע חיזוי על קבוצת המבחן.
השתמשו בערך ה-M האופטימלי שמצאתם בסעיף c. (ממשו $best\_m\_test$ שנמצאת ב $ID3\_experiments.py$ והריצו את החלק המתאים ב $main$). ציינו בדו"ח את הדיוק שקיבלתם. האם הגיזום שיפר את הביצועים ביחס להרצה ללא גיזום?

הגיזום שיפר את הביצועים על קבוצת המבחן, לפני הגיזום קיבלנו 94.17% הצלחה על קבוצת המבחן ולאחר הגיזום 97.09%. השיפור נובע ממודל פחות overfitted לקבוצת האימון.