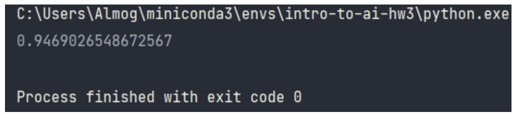
מבוא לבינה מלאכותית ־ תרגיל בית 3

שאלה 1

סעיף 2

לאחר אימון המסווג על קבוצת האימון ובדיקה על ידי קבוצת הבוחן הדיוק המתקבל הינו 0.9469026548672567.



איור 1: דיוק עבור קבוצת הבוחן

שאלה 2

הטענה נכונה. תחילה נשים לב כי נורמליזציית $MinMax$ היא למעשה הטרנספורמציה הבאה:

$$g(v)=\frac{v-v_{min}}{v_{max}-v_{min}}=\frac{1}{v_{max}-v_{min}}\cdot v-\frac{v_{min}}{v_{max}-v_{min}}$$

כאשר v_{max} ו־ v_{min} הם קבועים הנובעים **מסט האימון** המתקבל. נגזור ונקבל $g'(v)=\frac{1}{v_{max}-v_{min}}>0$. כלומר נורמליזציית $MinMax$ היא **טרנספורמציה מונוטונית עולה ממש** ולפיכך מתקיים כי $v_1>v_2$ אם ורק אם $g(v_1)>g(v_2)$.
מכאן נשים לב כי אם $threshold=\frac{v_1+v_2}{2}$ ואם נסמן ב־ $(threshold)'$ את ערך הסף המתקבל מהערכים המנורמלים אז מתקיים:

$$(threshold)'=\frac{g(v_1)+g(v_2)}{2}=\frac{\frac{v_1-v_{min}}{v_{max}-v_{min}}+\frac{v_2-v_{min}}{v_{max}-v_{min}}}{2}=\frac{\frac{v_1+v_2}{2}-v_{min}}{v_{max}-v_{min}}=\frac{theshold-v_{min}}{v_{max}-v_{min}}=g(threshold)$$

כלומר ערך הסף המתקבל מהערכים שעברו נרמול זהה להפעלת הנרמול על ערך הסף, ולכן מתקיים:

$$v_1<threshold<v_2\iff g(v_1)<g(threshold)<g(v_2)$$

כעת, יהי T העץ המתקבל על ידי הפעלת $ID3$ על הדאטה הלא מנורמל ויהי T' העץ המתקבל על ידי הפעלת $ID3$ על הדאטה המנורמל.
נראה באינדוקציה על עומק העץ כי T' זהה ל־ T כלומר יבחר בכל צומת את אותו מאפיין המתאים ב־ T ואת אותו ערך סף (עד כדי נרמול):
__בסיס:__ עומק 0, כלומר u, שהוא הצומת בעומק זה, הינו השורש: אם כמות הדוגמאות המתקבלת בצומת זו היא 0 אז לא תבחר תכונה ולכן הטענה נכונה באופן ריק. אחרת, כמות הדוגמאות גדולה מ־ 0: תהי f התכונה שנבחרה בצומת זה בעץ T על פי פונקציית בחירת המאפיינים ויהיו v_1, v_2 שתי ערכים שונים שהתכונה f יכולה לקבל כך שערך הסף $threshold$ הוא הממוצע שלהם. נשים לב כי כיוון שמדובר בשורש אז הדוגמאות שמגיעות לצומת זה זהות בשני העצים. עוד נשים לב כי ה־ $threshold$ המתאים בעץ T' הוא למעשה הפעלת הטרנספורמציה על ה־ $threshold$ ב־ T, ומשיקולים שהוסברו קודם נסיק כי אותן דוגמאות שהיו קטנות מ־ $threshold$ ב־ T יהיו קטנות מ־ $(threshold)'$ ב־ T' ולכן פיצול הדוגמאות בעצים, כאשר רוצים לחשב את תוספת המידע, יהיה זהה (ולכן גם ההסתברויות). כמו כן, מכיוון שתוספת המידע תלויה גם בסיווג, **שאינו עובר נירמול**, נובע כי תוספת המידע תהיה זהה ולכן אותה תכונה f תבחר עם $(threshold)'=g(threshold)$.
__צעד:__ נניח כי לכל צומת עד עומק k העץ T זהה לעץ T'. יהי u צומת כלשהו בעומק $k+1$ ב־ T'. מהנחת האינדוקציה נובע שעד צומת האב העץ T זהה לעץ T' עד כדי נרמול של ערך הסף, ולכן הדוגמאות המגיעות ל־ u זהות לאלו שמגיעות למקבילו בעץ T. לפיכך נוכל לחזור על שהוסבר בבסיס ולהסיק כי המאפיין שייבחר בשני העצים זהה עם ערכי סף זהים עד כדי נרמול, ובכך סיימנו את ההוכחה.
__מסקנה:__ נרמול של הדאטה אינו משפיע על בניית העץ ולכן גם לא משפיע על הדיוק על קבוצת הבוחן (כמובן בתנאי שקבוצת הבוחן עוברת את **אותה טרנספורמציה מונוטונית עולה ממש** עם הקבועים מסט האימון).

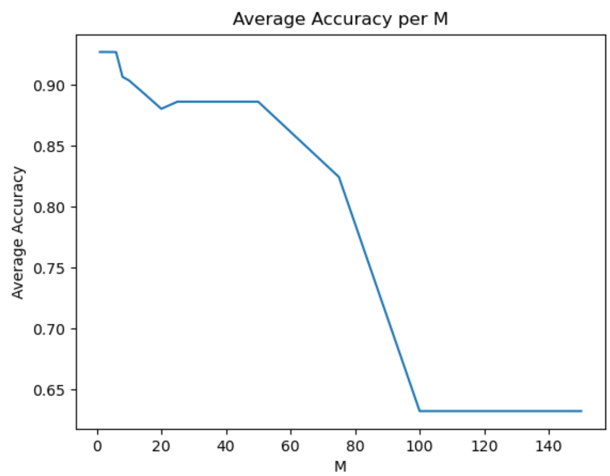
* הערה: ניתן אישור מהמתרגל האחראי על התרגיל לחרוג במעט מ־ 20 שורות

שאלה 3

סעיף 1

גיזום של עץ החלטה מאפשר לקבל עץ סיווג קטן יותר, אשר בתקווה יוכל למזער את תופעת ה־ $Over-Fitting$. מטרתו של גיזום עץ החלטה היא למנוע התאמת יתר של המסווג לדאטה עליו אומן (סט האימון), כלומר תיתכן שגיאה גדולה יותר עבור דאטה זה, אך נוכל לגרום לכך שכל החלטה, אשר מתבצעת בעלים, תסתמך על יותר דאטה ובכך אולי תגרום לדיוק גבוה יותר על **דוגמה חדשה שלא נראתה קודם לכן**.

סעיף 3



איור 2: דיוק ממוצע כתלות בערך של M

ניתן לראות שהגרף הנ"ל הינו במגמה יורדת, כלומר **באופן כללי** ככל ש־ M יותר קטן כך הדיוק הממוצע גבוה יותר. בפרט ניתן לראות כי הדיוק הממוצע המקסימלי מתקבל עבור $M=1$ והוא 0.9469026548672567. מהתבוננות בגרף אנו רואים כי הגיזום למעשה רק פוגע ביכולת ההכללה של המסווג, אולי מכיוון שסט האימון קטן יחסית (כ־ 350 דוגמאות בלבד) ולכן העץ "קטן מספיק" גם ללא גיזום. ביתר פירוט, ייתכן שבעת בניית העץ מספר הדוגמאות מגיע "מהר יחסית" מתחת ל־ M עבור M שבערך גדול מ־ 5 ולכן בסופו של דבר נוצר עץ "קטן מדי", ואילו עבור M שבערך קטן מ־ 5 אין השפעה כי (אולי) כמות הדוגמאות בכל צומת גדולה מ־ 5 או שווה 0, ולכן העץ שנוצר זהה לעץ שנוצר ללא גיזום.

סעיף 4

כפי שניתן לראות בסעיף הקודם, הערך של M הנותן דיוק מקסימלי הינו $M=1$, כלומר ללא גיזום כלל, ולכן הדיוק ישאר 0.9469026548672567 כפי שהיה מקודם.

שאלה 4

סעיף 1

עבור הרצת אלגוריתם $ID3$ עם גיזום מוקדם ועם פונקציית ההפסד המוזכרת בשאלה ערך ה־ M המתקבל הינו $M = 1$ ־ כלומר ללא גיזום.
לאחר אימון המודל עם ערך ה־ M שנמצא על כל סט האימון, השגיאה המתקבלת על סט הבוחן הינה 0.02123893805309735.

סעיף 2

באלגוריתם ה־ $ID3$ בכל פעם שפונקציית מציאת המאפיין נקראת המאפיין הנבחר הוא זה שממקסם את תוספת המידע (Information Gain).
מכיוון שאנו רוצים למזער **באופן ספציפי את פונקציית ה־ $Loss$ הנתונה**, נציע את השינוי הבא **לפונקציית בחירת המאפיין** אשר תעזור לשפר את ערך ה־ $Loss$ עבור המסווג כולו:
במקום להחזיר את המאפיין שממקסם את תוספת המידע ־ נשמור רשימה של מועמדים אשר לבסוף נבחר אחד מהם בצורה חכמה על סמך פונקציית ה־ $Loss$ הנתונה. העיקרון מאחורי רעיון זה דומה לעיקרון שלמדנו ב־ $A^*\epsilon$ בו מסתכלים על "חלון" של מועמדים.
נסביר כיצד הפונקציה תפעל:
נסמן ב־ $Candidates$ את רשימת המועמדים, כאשר לכל $candidate \in Candidates$ אנו שומרים את מספר המאפיין, ערך הסף ותוספת המידע שהצליח להשיג, כלומר:

$$candidate = (FeatureID, Threshold, IG)$$

כמו כן, נסמן ב־ IG_{max} את תוספת המידע של המועמד בעל תוספת המידע המקסימלית ברשימת המועמדים כרגע.
בכל שלב בריצת הפונקציה כאשר נמצא מועמד $(FeatureID, Threshold, IG)$ נבצע את כל מה שצריך מבין הבאים:

- אם $IG_{max} < IG$ אז נעדכן את IG_{max}.
- אם מתקיים $\left|\frac{IG}{IG_{max}} - 1\right| < \epsilon$ נוסיף אותו לרשימת המועמדים.

כמו כן, אם ביצענו עידכון ל־ IG_{max} אז נבצע את הסינון הבא:
לכל $candidate \in Candidates$ נבדוק האם מתקיים: $\left|\frac{candidate.IG}{IG_{max}} - 1\right| < \epsilon$. אם כן, נשאיר את המועמד ברשימת המועמדים, אחרת נסיר אותו מהרשימה. נשים לב שבכך אנו למעשה שומרים את כל המועמדים שתוספת המידע שהצליחו להשיג "רחוקה" מ־ IG_{max} בלכל היותר ϵ אחוז:

$$\left|\frac{candidate.IG}{IG_{max}} - 1\right| < \epsilon \iff 1 - \epsilon < \frac{candidate.IG}{IG_{max}} < 1 + \epsilon \iff IG_{max} \cdot (1 - \epsilon) < candidate.IG < IG_{max} \cdot (1 + \epsilon)$$

לבסוף, לאחר שעברנו על כל המאפיינים וערכי ה־ $Threshold$ שלהם וכעת שיש בידינו רשימה של מועמדים בעלי תוספת המידע המקסימלית עד כדי ϵ, נפעל באופן הבא:

- לכל מועמד $(FeatureID, Threshold, IG)$ נבצע:
  - נפריד בין הדוגמאות בסט האימון שעבורם ערך המאפיין גדול שווה לסף ולאלו שעבורם ערך המאפיין קטן מערך הסף.
  - לכל אחת מתתי הקבוצות הנ"ל נבחר את הסיווג על סמך הרוב $(Majority)$ ונחשב את ה־ $Loss$ עבורם.
  - לבסוף נסכום את ה־ $Loss$ עבור שתי תתי הקבוצות הנ"ל באופן ממושקל על פי גודל תת הקבוצה.
- לבסוף, נבחר את המאפיין וערך הסף של המועמד אשר ממזער את השגיאה הממושקלת הנ"ל.

בכך למעשה אנו מבטיחים כי המאפיינים וערך הסף שנבחר יקיימו את הבאים:

1. תוספת המידע המתקבלת עבורם היא לכל היותר רחוקה ב־ ϵ אחוז מתוספת המידע המקסימלית האפשרית.
2. מבין המאפיינים שעומדים בדרישת ה־ ϵ המאפיין שנבחר ממזער את פונקציית ה־ $Loss$ (תחת ההנחה של $LookAhead$ של אחד, כלומר כאילו כל אחד מילדיו הם עלים).

סעיף 3

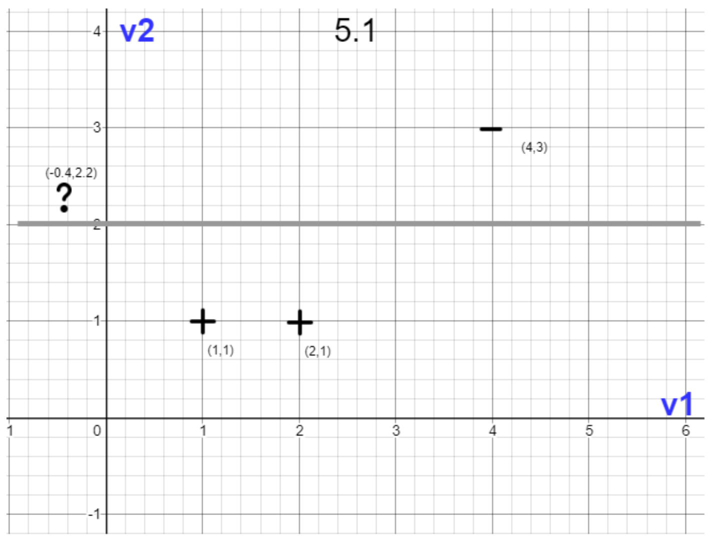
נשים לב שבאלגוריתם זה ϵ הינו פרמטר אותו נצטרך לכוונן. את הכיוונון נעשה על ידי הגרלה של ערכים שונים (קטנים יחסית) והרצת $Cross - Validation$ לכל אחד מהם.
לאחר ביצוע הפעולה הנ"ל השגיאה המתקבלת עבור סט האימון ועבור ה־ ϵ הטוב ביותר היא: 0.001769911504424779.

שאלה 5

הבהרות לסעיפים הבאים:

- מסווגי המטרה מסומנים באפור.
- סיווג " $-$ " מתאים לערך 0.
- סיווג " $+$ " מתאים לערך 1.

סעיף א



איור 3: סט אימון ומסווג מטרה

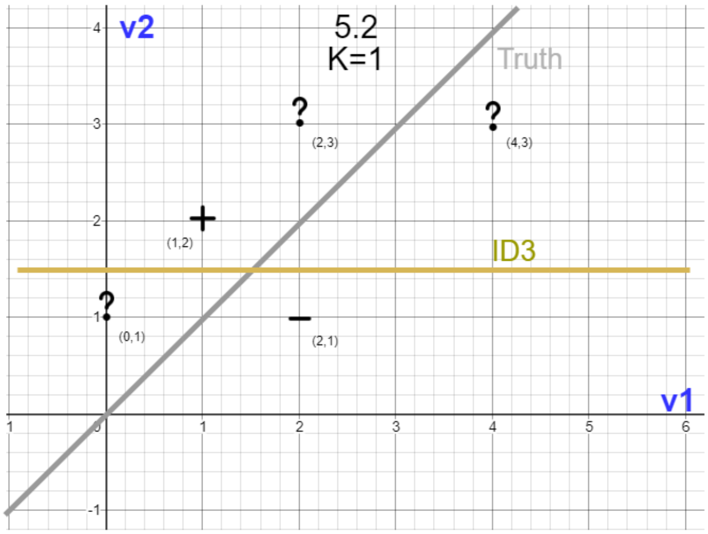
הבהרה: מסווג המטרה מסומן באפור ומתלכד עם המסווג הנוצר מ־ $ID3$.
נשים לב כי מקסום תוספת המידע מתקבל עבור v_1 עם $threshold = 3$ או עבור v_2 עם $threshold = 2$, כאשר שניהם יוצרים הפרדה מקסימלית ולכן ייבחר v_2 כי יש לו (למאפיין) אינדקס גדול יותר. לפיכך המסווג יהיה:

$$h_{truth}\left(x_i\right) = h_{ID3}\left(x_i\right) = \begin{cases} 1 & v_{2,i} < 2 \\ 0 & 2 \leq v_{2,i} \end{cases}$$

כמו כן, ניתן לראות כי הדגימה "?" מסט הבוחן הממוקמת ב־ $(-0.4, 2.2)$ אמורה להיות שלילית שכן היא מעל הישר המפריד, ומתקיים:

- עבור $K = 1$ הדגימה הקרובה ביותר היא $(1, 1)$ שסימנה חיובי ולכן תסווג כחיובית ־ שגיאת סיווג.
- עבור $K = 2$ הדגימות הקרובות ביותר הן $(1, 1)$ ו־ $(2, 1)$ שסימנן חיובי ולכן תסווג כחיובית ־ שגיאת סיווג.
- עבור $K = 3$ הדגימות הקרובות ביותר הן כל סט האימון, כאשר רובו מסומן חיובי, ולכן תסווג כחיובית ־ שגיאת סיווג.

סעיף ב



איור 4: סט אימון, מסווג מטרה ומסווג מ־ $ID3$

נבחר $K=1$. כמו כן, בתמונה הנ"ל ניתן לראות את מסווג המטרה (Ground Truth) בקו אפור המקיים:

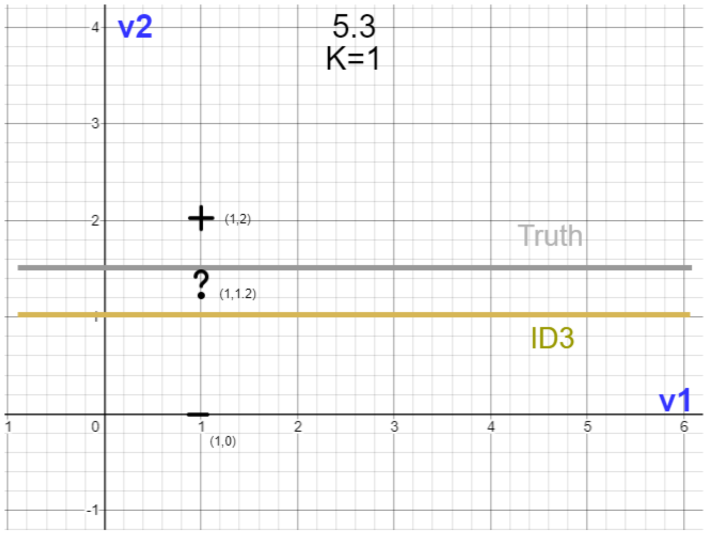
$$h_{truth}\left(x_i\right)=\begin{cases}1 & v_{1,i}<v_{2,i}\\ 0 & v_{2,i}\leq v_{1,i}\end{cases}$$

כמו כן, עבור סט האימון הנ"ל, ישר זה הוא גם המקום הגאומטרי של כל הנקודות שמרחקן זהה משתי הדגימות בסט האימון, ולכן אם דגימה מסוימת מסט הבוחן נמצאת מעל ישר זה, למשל ה־ "?" הנמצא ב־ $(2,3)$, אז היא קרובה יותר ל־ " + " שנמצא ב־ $(1,2)$ ולכן תסווג כחיובית לפי $NN-1$. באופן דומה אם היא מתחת לישר זה, למשל ה־ "?" הנמצא ב־ $(4,3)$, אז היא קרובה יותר ל־ " − " שנמצא ב־ $(2,1)$ ולכן תסווג כשלילית לפי $NN-1$. עוד נציין כי אם היא נמצאת **על הישר** אז מרחקיה שווים ולכן תסווג כ־ " − " (כי מסתכלים קודם על הדגימה מסט האימון שעבורה v_1 מקסימלי). מכאן נסיק כי כל דוגמה שתתקבל מסט בוחן כלשהו תסווג נכון על ידי $NN-1$.
מצד שני, עבור סט האימון הנ"ל אלגוריתם ה־ $ID3$ יחזיר את המסווג הבא (כל מאפיין יאפשר הפרדה מלאה אך האלגוריתם יבחר את המאפיין עם האינדקס הגדול יותר):

$$h_{ID3}\left(x_i\right)=\begin{cases}1 & 1.5\leq v_{2,i}\\ 0 & v_{2,i}<1.5\end{cases}$$

לכן, דוגמה הנמצאת ב־ "?" שממוקם ב־ $(0,1)$ תסווג על ידי העץ כשלילית בעוד שהיא למעשה חיובית.

סעיף ג



איור 5: סט אימון, מסווג מטרה ומסווג מ־ $ID3$

נבחר $K=1$. כמו כן, בתמונה הנ"ל ניתן לראות את מסווג המטרה (Ground Truth) בקו אפור המקיים:

$$h_{truth}\left(x_i\right)=\begin{cases}1 & 1.5\leq v_{2,i}\\ 0 & Otherwise\end{cases}$$

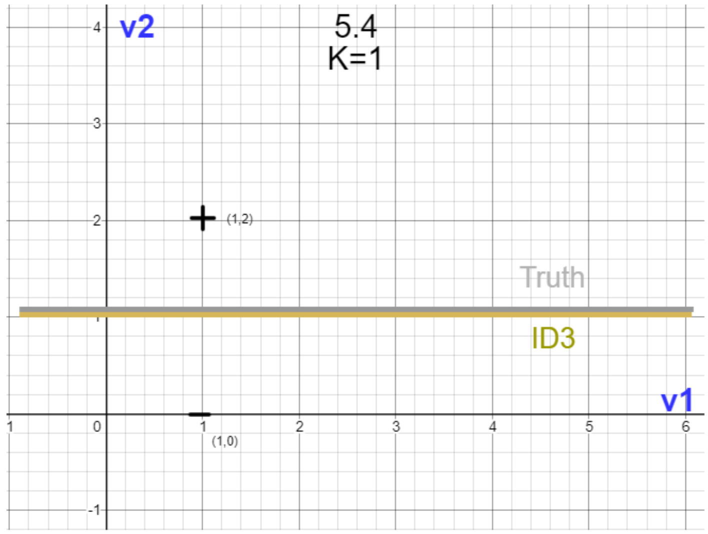
כמו כן, עבור סט האימון הנ"ל נקבל שהמסווג המתקבל מ־ $ID3$ הינו:

$$h_{ID3}\left(x_i\right)=\begin{cases}1 & 1\leq v_{2,i}\\ 0 & v_{2,i}<1\end{cases}$$

כעת, עבור הדגימה מסט הבוחן הנמצאת ב־ "?" הממוקם ב־ $(1,1.2)$ מתקיים:

- תסווג ב־ " $+$ " על ידי h_{ID3} כי $1\leq 1.2=v_{2,i}$ למרות שהיא בפועל אמורה להיות " $-$ " - שגיאת סיווג.
- תסווג ב־ " $+$ " על ידי $NN-1$ כי הדגימה הקרובה מסט האימון היא ה־ " $+$ " הנמצא ב־ $(1,2)$ למרות שהיא בפועל אמורה להיות " $-$ " - שגיאת סיווג.

סעיף ד



איור 6: סט אימון, מסווג מטרה ומסווג מ־ $ID3$

נבחר $K = 1$. כמו כן, בתרשים הנ"ל ניתן לראות כי מתקיים:

$$h_{truth}\left(x_i\right) = h_{ID3}\left(x_i\right) = \begin{cases} 1 & 1 \leq v_{2,i} \\ 0 & v_{2,i} < 1 \end{cases}$$

נשים לב שעבור דגימה מסוימת מסט בוחן כלשהו מתקיים אחד מהבאים:

- אם היא נמצאת מעל ישר זה אז היא חיובית (Ground Truth) והיא מקיימת $1 < v_{2,i}$ ולכן תסווג ב־ " $+$ " הן על ידי h_{ID3} והן על ידי $1 - NN$ (כי אז היא קרובה יותר ל־ " $+$ " שנמצא ב־ $(1,2)$).
- אם היא נמצאת מתחת לישר זה אז היא שלילית (Ground Truth) והיא מקיימת $v_{2,i} < 1$ ולכן תסווג ב־ " $-$ " הן על ידי h_{ID3} והן על ידי $1 - NN$ (כי אז היא קרובה יותר ל־ " $-$ " שנמצא ב־ $(1,0)$).
- אם היא נמצאת **על** הישר אז היא חיובית (Ground Truth) והיא מקיימת $v_{2,i} = 1$ ולכן תסווג ב־ " $+$ " הן על ידי h_{ID3} והן על ידי $1 - NN$ (כי אז המרחקים שווים וגם v_1 שווים ולכן נתייחס קודם לדגימה שעבורה v_2 מקסימלי וזו דגימה חיובית).

שאלה 6

סעיף 1

בשאלה הנ"ל אנו נדרשים לממש את האלגוריתם $knn-decision-tree$ ולשם כך אנו נדרשים למצוא ערכים "טובים" לפרמטרים N, K, p.
על מנת למצוא ערכים אלו ביצעתי $5-Cross-Validation$ מספר פעמים, כאשר בכל פעם בחרתי את כל הפרמטרים N, K, p באקראי, וזאת מכיוון שזמן הריצה לבדיקת פרמטרים גדולה יחסית ולכן מעבר על הערכים האפשריים בצורת $Grid$ לדעתי פחות מתאימה.
תוצאת הדיוק המתקבלת על סט הבוחן הינה 0.9646017699115044.

שאלה 7

סעיף 1

נציע את השיפור הבא:
נציג פרמטר חדש של טמפרטורה שיסומן ב־ T אותו נכוונן בהמשך, ונבצע את הסיווג על סמך הועדה של K העצים הקרובים ביותר ל־ x_i כאשר כל עץ מקבל משקל בהתאם למידת הקירבה של הסנטרואיד שלו ל־ x_i ועל סמך T. כלומר בהינתן דגימה x_i מסט הבוחן שלא נראתה קודם לכן נקבע את סיווגה באופן הבא:

$$\hat{y_i} = h\left(x_i\right) = sign\left(\sum_{\text{j}\in\text{k\_closest\_trees\_to\_x\_i}} w_j \cdot predict\left(\text{tree\_j}, x_i\right)\right)$$

כאשר:

$$predict\left(\text{tree\_j}, x_i\right) = \begin{cases} 1 & \text{if tree j predicts } x_i \text{ is `sick`} \\ -1 & \text{if tree j predicts } x_i \text{ is `healty`} \end{cases}$$

$$w_j = \frac{e^{\frac{d(x_i,j)}{T}}}{\sum_{l\ \in\text{k\_closest\_trees\_to\_x\_i}} e^{\frac{d(x_i,l)}{T}}}$$

$$d\left(x_i, l\right) = \begin{cases} -m & m < K,\ tree_l \text{ is the `m` closest to } x_i (\text{starting from } 0) \\ -\infty & Otherwise \end{cases}$$

בכך למעשה נוכל לשקלל את התוצאות של K העצים הקרובים ביותר, כאשר ניתן משקל גדול יותר לעצים שהסנטרואיד שלהם יותר קרוב ל־ x_i.
עוד נציין כי הטמפרטורה, (נציין כי $T > 0$), שולטת על רמת ה"עדיפות" שניתן לכל עץ. כך למשל, עבור T גדול מאוד נקבל:

$$w_j = \frac{e^{\frac{d(x_i,j)}{T}}}{\sum_{l\ \in\text{k\_closest\_trees\_to\_x\_i}} e^{\frac{d(x_i,l)}{T}}} \approx \frac{e^0}{\sum_{l\ \in\text{k\_closest\_trees\_to\_x\_i}} e^0} = \frac{1}{K}$$

כלומר עבור T גדול נקבל שהמשקלים זהים - כלומר העץ הקרוב ביותר יקבל עדיפות אפסית אל מול שאר $K-1$ העצים הקרובים.
באופן דומה עבור T קרוב ל־ 0 נקבל:

$$w_j = \frac{e^{\frac{d(x_i,j)}{T}}}{\sum_{l\ \in\text{k\_closest\_trees\_to\_x\_i}} e^{\frac{d(x_i,l)}{T}}} \approx \frac{\mathbb{I}\left[j \text{ is the closest tree to } x_i\right]}{0 + ... + 1 + ... + 0} =$$

$$= \mathbb{I}\left[j \text{ is the closest tree to } x_i\right] = \begin{cases} 1 & j \text{ is the closest tree to } x_i \\ 0 & j \text{ is not the closest tree to } x_i \end{cases}$$

כלומר עבור T קטן מאוד אנו נותנים עדיפות "אינסופית" לעץ הקרוב ביותר - כלומר העץ הקרוב ביותר הוא היחיד שמשפיע.

סעיף 2

באופן דומה לשאלות הקודמות חיפוש הפרמטרים N, K, p, T בוצע על ידי הגרלת ערכים רנדומליים ובדיקתם בעזרת $Cross - Validation$.
תוצאת הדיוק המתקבלת על סט המבחן הינה 0.9911504424778761.