Intro to AI

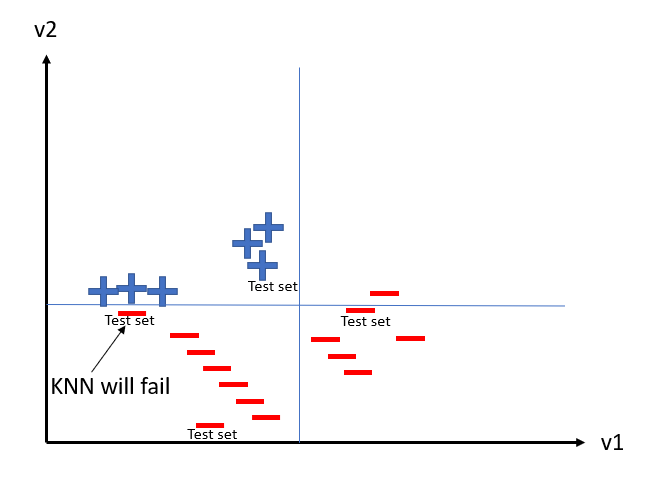
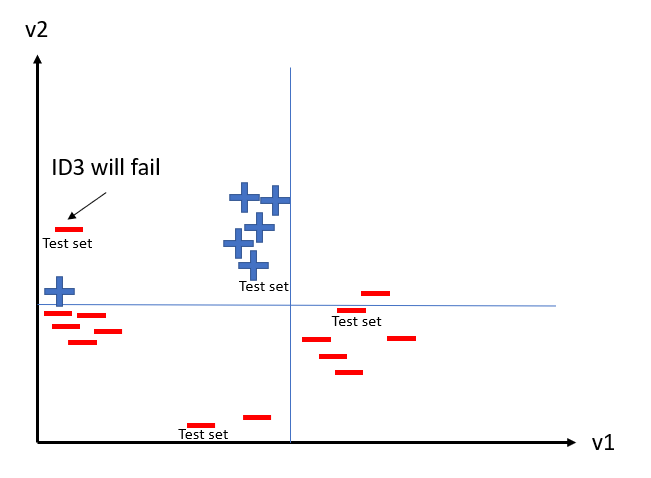
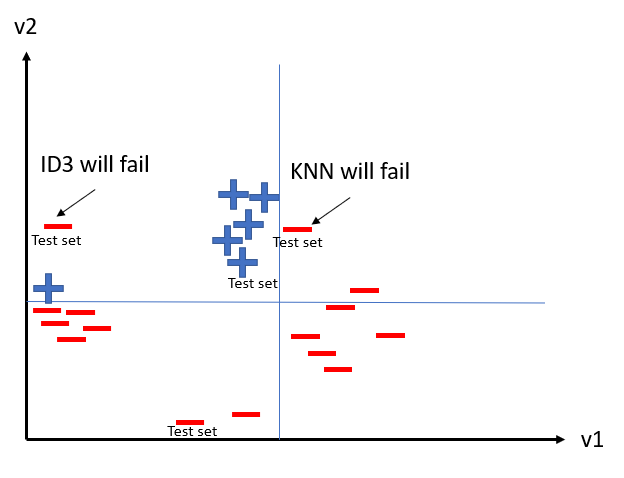
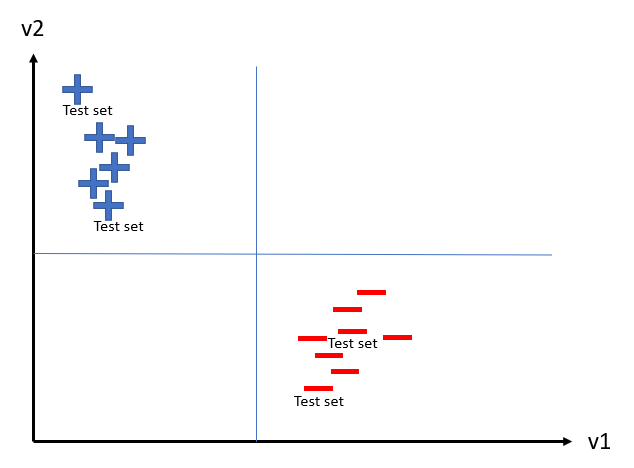
**Homework Exercise 3 – Dry**

**Winter 2021-22**

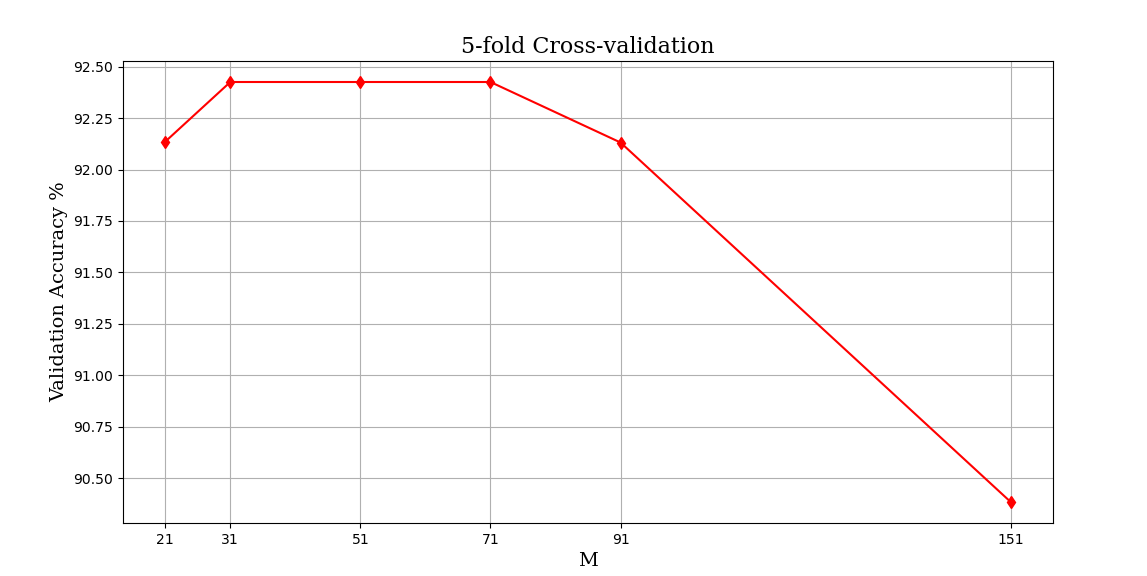
**Ophir Khill 206629966**

**Gil Kapel 305323776**

**חלק ב'**

1. **הוכחה:**
   1. בהינתן דאטה כלשהו עם תכונות רציפות ותיוגים בינאריים, הפעלה של פונקציית נירמול MINMAX, על הדאטה אינה משפיעה על דיוק של מסווג ID3, הנלמד על קבוצת האימון והנבחן על קבוצת המבחן וזאת מכיוון שאלגוריתם זה מחלק בכל צומת את כל האיברים לפי פונקציה סקלרית ולכן אם נפעיל את פונקציית הנירמול minmax גם על ערכי הפונקציה הנ"ל נקבל כי החלוקה למחלקות תישאר זהה ומכיוון ש – ID3 פועל עם הפונקציה הנ"ל נקבל עץ זהה לאחר הנירמול.
2. KNN מול ID3:
   1. ID3 מוגדרת ע"י הקווים הכחולים (קטן מv1 מסויים וגדול מv2 אחר, עבור נקבל כי KNN ייכשל בחץ המסומן - דגימה מקבוצה הבחינה תסווג כחיובית כחולה למרות שהיא שלילית ואדומה, אך מכיוון שערך v2 שלה קטן מההפרדה בid3 העץ יצליח לחזות אותה בתור דגימה אדומה.
   2. עבור K=3 במקרה הנ"ל חזאי KNN יחזה את כל הנקודות כפי שהן מכיוון שליד כל דגימת ניסוי יש לפחות 3 דגימות מאותו סוג שהכי קרובות אליה. נשים לב כי עבור סט האימון בבניית עץ ע"י ID3 נקבל אותו סיווג כמו בסעיף הקודם (הנק' האדומה השמאלית עליונה לא בסט אימון ולכן לא תשפיע על בניית העץ) אך עץ זה לא ינבא נכון בהפעלת החיזוי על הדגימה הנ"ל – ינבא כי הדגימה חיובית כחולה ולא שלילית אדומה.
   3. בקבוצה זו, כפי שהוסבר בסעיף 1.3 ID3 ייכשל בדגימה האדומה- שמאלית עליונה. ניתן לראות כי הוספנו עוד דגימת ניסוי אדומה שלילית בה KNN ייכשל עבור k=3 וזאת מכיוון שדגימה זו נמצאת בקרבת דבוקה של 3 דגימות חיוביות ורחוקה מדגימות אדומות שליליות ולכן תסווג ע"י KNN כדגימה כחולה חיובית למרות שהיא דגימה אדומה שלילית. לכן שני המסווגים ייטעו עבור קבוצת אימון וניסוי זאת.
   4. ניתן לראות בקבוצה זו כי עבור נקבל כי KNN הינו מסווג מטרה ויסווג נכון את כל דגימות הניסוי (מעל 8 נקבל כי כל דגימת ניסוי תקבל תוצאה אדומה מכיוון שזהו הרוב של הדגימות) וניתן לראות בקלות את סיווג העץ ע"י ID3 המאפשר הפרדה מלאה בין קבוצות הדגימות ולכן גם כאן תהיה דגימה נכונה לכל דגימות הניסוי.
3. שאלה 3:
   1. מכיוון שבקבוצת האימון יש 5 דגימות עם תווית '0' ו5 דגימות עם תווית '1' נקבל כי ערך הדיוק של מסווג הרוב הינו 50% מכיוון שאין רוב לאף תווית המסווג ייבחר רנדומלית בפילוג אחיד ומכיוון שיש 50% דגימות מאותו סיווג הוא יסווג ב50% סיווג נכון וב50% סיווג שגיאה.
   2. ע"י הרצת 2-fold validation כאשר נחלק את שני הקבוצות ל – 2 ע"פ החלוקה הנתונה יתקבל שקבוצה 1 : תחזה תמיד 1 (Majority classifier)  
      קבצוה 2: תחזה תמיד 0 (Majority classifier)  
      מכיוון שבקבוצה 1 80% מהדגימות בעלות תווית "1" אז כאשר נבדוק אותה לפי המסווג של קבוצה 2 נקבל דיוק של 20%  
      כנ"ל על קבוצה 2 (20% דיוק)  
      ולכן כאשר נעשה ממוצע על שניהם נקבל שערך הדיוק הוא 20%

**חלק ג'**

1. שאלה 4 – רטוב
2. שאלה 5 – רטוב
   1. הדיוק המתקבל הינו: 95.6%
3. שאלה 6 –
   1. גיזום עץ נועד למנוע תופעת Over fitting הנובעת מכך שבנינו את העץ ע"י קבוצת דגימות אימון ספציפית שאינה מכילה את כל הדגימות בעולם. אם לא נגזום את העץ נקבל מסווג מושלם עבור קבוצת האימון אך כנראה שהסתברות השגיאה עבור דגימה שאינה מקבוצת האימון תגדל. באופן כללי ע"פ עקרון Occam's Razor נרצה לבחור בפתרון הפשוט ביותר הנותן תוצאה טובה על קבוצת האימון אך לאו דווקא התוצאה המיטבית.
   2. רטוב
   3. 
   4. ניתן לראות כי עבור גיזום גדול מ70 או קטן מ30 קיבלנו ירידה בדיוק סט הוולידציה ולכן נבחר את פרמטר הגיזום להיות 51. הדיוק המתקבל עבור **סט** **הניסוי** לאחר הגיזום ע"ב פרמטר הגיזום 51 הינו 97.35% - ניתן לראות כי הגיזום אכן שיפר את תוצאות הדיוק אשר התקבלו בשאלה 5.1 - מ95.6 ל97.3 אחוז דיוק בסיווג.

**חלק ד'**

1. אלגוריתם KNN בוחן את התווית הנפוצה ביותר מבין k הדגימות השכנות אליו במרחב של דגימות האימון ובוחר בתווית זו. **יתרונות האלגוריתם** הינם 1) זמן אימון מהיר ופשוט למימוש ו2) ניתן להתאים פונקציות מרחק שונות ולקבל מגוון מודלים למגוון בעיות סיווג (לדוגמא פונקצית דימיון בין שני אובייקטים ללא תכונות אמפיריות). **חסרונות האלגוריתם** הינם 1) זמן סיווג ארוך מאוד על חשבון זמן אימון קצר ו2) רגישות לתכונות לא רלוונטיות - תכונות שאינן משפיעות על סיווג של דוגמא מסויימת ישפיעו בכל מקרה על מיקומו במרחב ולכן על אופציה לשכנים אחרים מהסיווג ללא התכונה הנ"ל.
2. שאלה 8:
   1. קבוצת תת הקבוצות הינן קבוצת החזקה - קבוצה בגודל - ולכן זהו מס' תתי הקבוצות X.
   2. ישנן תתי קבוצות בגודל b וזאת מכיוון שכל תת קבוצה יכולה להכיל b מספרים מתוך הקבוצה הראשית D.
3. שאלה 9:
   1. על מנת למצוא את כמות וסוג המאפיינים האידיאלים לסיווג ע"י KNN נאמן את המודל ע"י דגימות האימון בלבד, אותן נחלק לדגימות ולידציה ודגימות אימון (מתוך דגימות האימון הכלליות – 80% לאימון המודל ו20% לוולידציה)
   2. גודל תת הקבוצה הממוצע הינו 5 (מתוך 8 מאפיינים במידע הגולמי) אך מכיוון שאנו בונים את המודל ע"ב סט האימון והוולידציה באופן רנדומלי ומכיוון שאנחנו עובדים עם אלגוריתם חמדן לא בהכרח נקבל את גודל זה בכל בניית מסווג מחדש. בנוסף, ניתן לראות כי בממוצע נקבל ציון טוב יותר מהציון עם כל המאפיינים אך ייתכן מצב בו נקבל ציון פחות טוב, להערכתנו הסיבה לשגיאה זו הינה בעקבות אי קורלציה בין סט האימון לסט הניסוי – אנו בוחרים את המסווג הטוב ביותר עבור סט הוולידציה שאינו בהכרח הטוב ביותר עבור סט הניסוי.
   3. האלגוריתם הממומש הינו אלגוריתם חמדן הבוחר בכל איטרציה איזה מאפיין להוריד מסט המאפיינים אשר מביא את התוצאה הטובה ביותר מבין כל המאפיינים.

האלגוריתם יבצע היחלצות ממינימום מקומי ע"י random restart המבוצע מס' פעמים ויחזיר את סט המאפיינים האופטימלי (הסט אשר מביא את הציון הגבוהה ביותר עבור סט הוולידציה).

עקרון האלגוריתם:

* + 1. נבצע את האלגוריתם הבא על כל ערכי b האפשריים:
    2. נאתחל מערך של **כל המשתנים** features.
    3. נבצע לולאה ובכל איטרציה נבחר מאפיין אחד ממערך features אותו נוריד (המאפיין שהורדתו תייצר את הציון הגבוהה ביותר) :
       1. נבצע אימון מבוסס k-fold על סט האימון
       2. נבנה מסווג KNN ע"ב סט האימון החלקי
       3. נבצע חיזוי לכל הדגימות בסט ולידציה וניתן ציון למסווג
       4. נבחר את המסווג הטוב ביותר (לאחר ביצוע k-fold - kפעמים ע"ב data מעורבל).
    4. נבצע בדיקה זה 10 פעמים כאשר נשמר את הרנדומליות של הבדיקות – random restart.
    5. נחזיר את המערך features אשר הביא את הציון הטוב ביותר בכל האיטרציות (בוחר את כמות המאפיינים ואת המאפיינים האפקטיביים ביותר לסיווג)