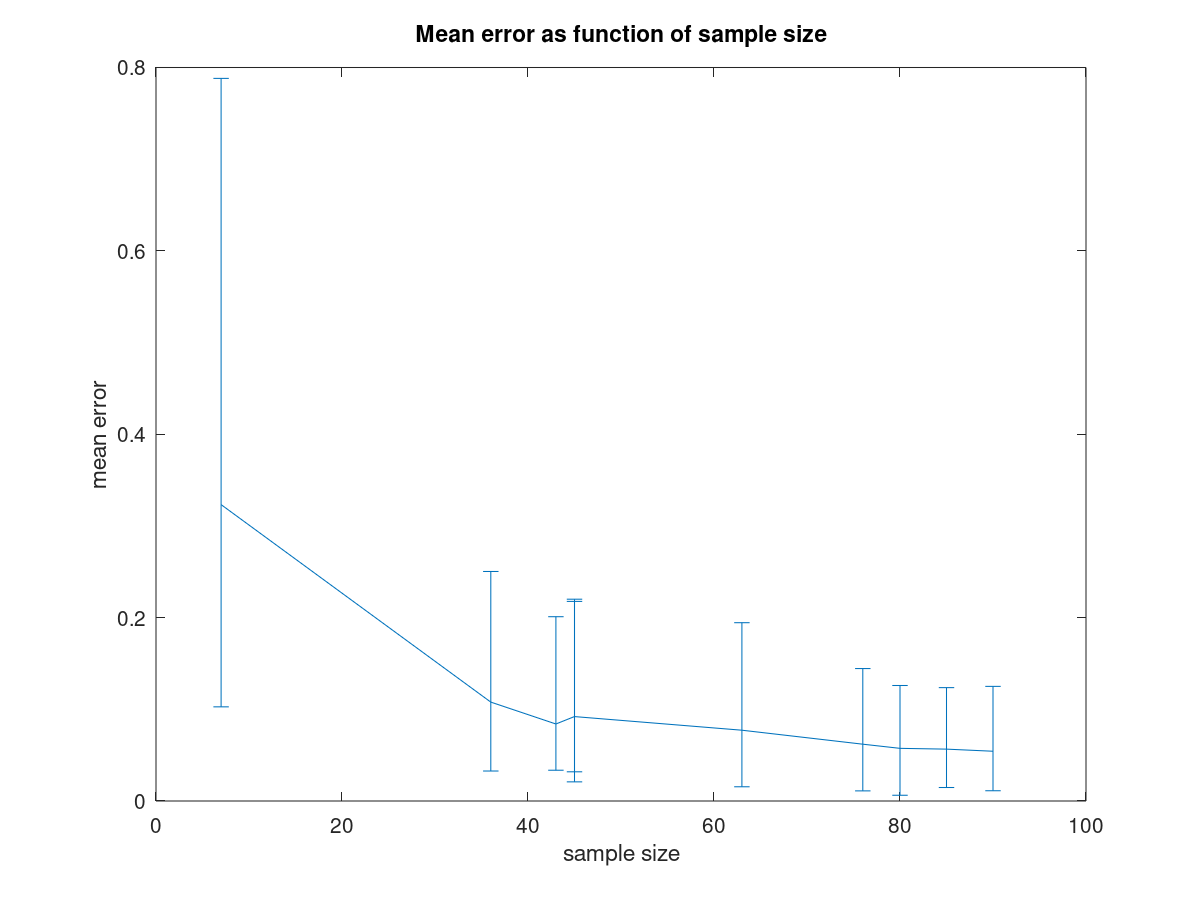
**עבודה 1 – מבוא ללמידה וניתוח של מידע רב**

**מגישים: עומרי אטל 208625103, רעי וייס-ליפשיץ 208347039**

שאלה 2

 סעיף א

סעיף ב

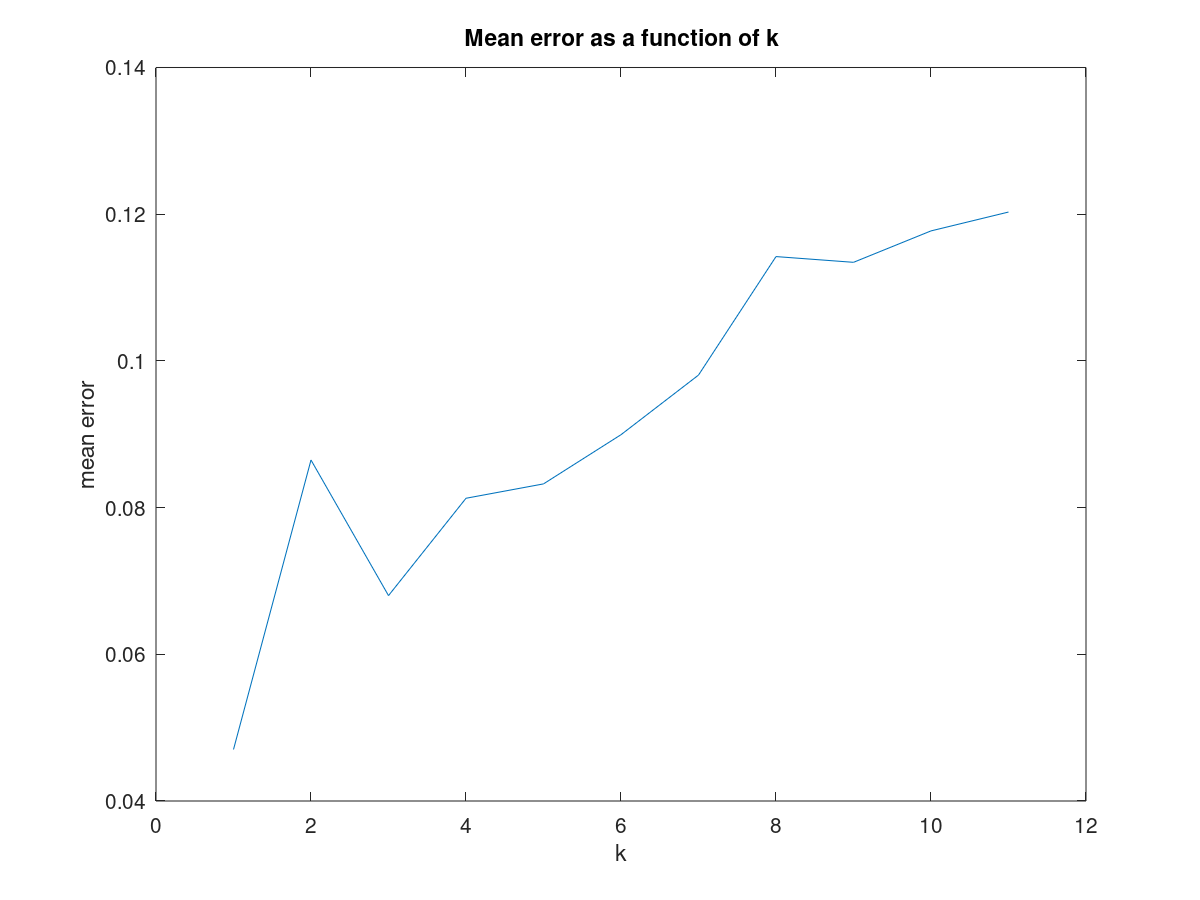
השגיאה הממוצעת יורדת ככל שה sample size גדל.  
הסבר: נשים לב תחילה שתמונות בעלות אותו לייבל יהיו דומות יחסית אחת לשנייה, כלומר המרחק האוקלידי שלהן יהיה קרוב ברוב המקרים. עבור k=1 אנו מריצים את למעשה את שיטת השכן הקרוב. ככל שהמדגם גדל, יש ל-predictor יותר דוגמאות מהלייבלים השונים, כך שלמעשה נוצרים גושים של נקודות בעלי אותו לייבל שגדלים יחד עם המדגם. מכאן, כאשר נפעיל את השכן הקרוב על כל דוגמה מה-test, יש סיכוי יותר גבוה שהדוגמה הזו "תיפול" בגוש הנכון (עם אותו הלייבל).

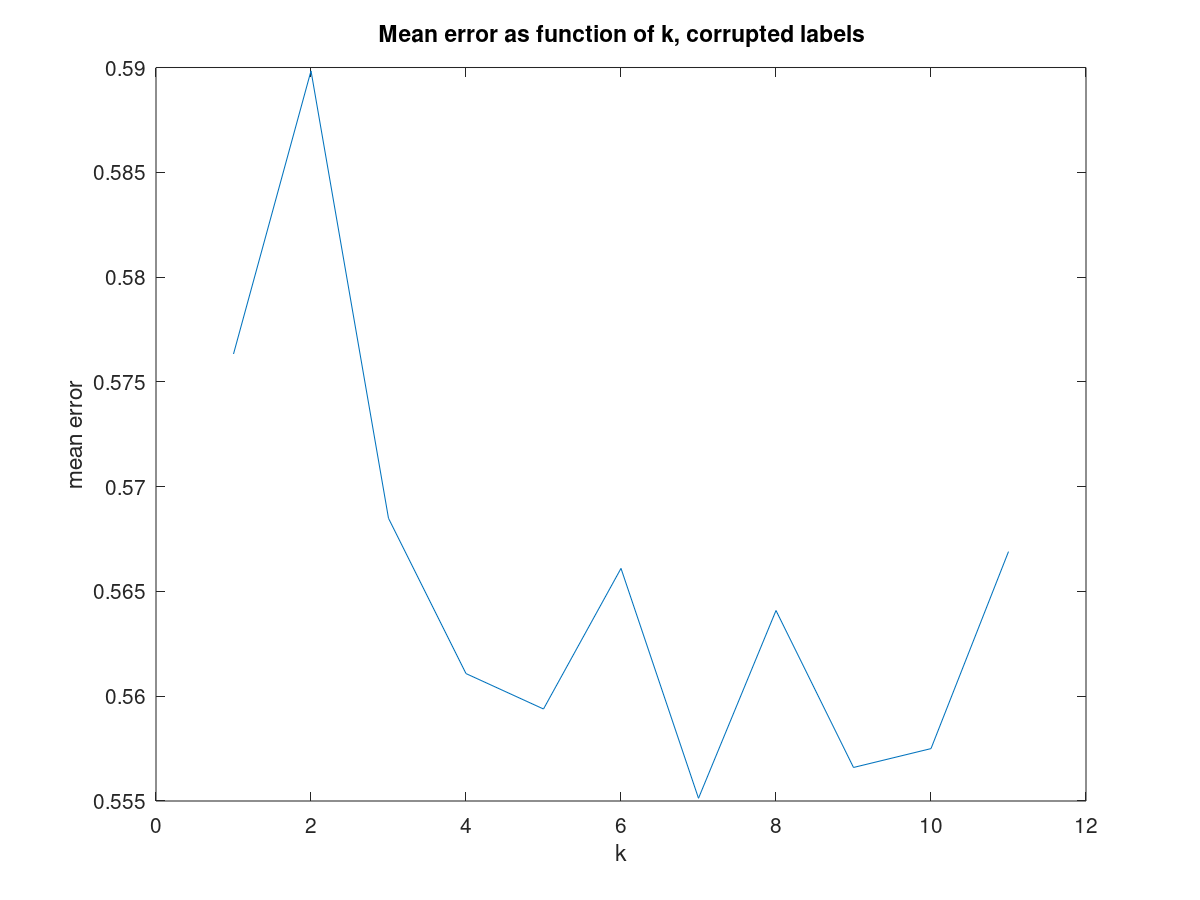
סעיף ג

כן, משום שב-gensmallm אנו יוצרים מדגם רנדומלי מה-train samples , כולכן בהרצות שונות יהיה לנו מדגמים שונים, מה שמשפיע על השגיאות.

סעיף ד

טענו בסעיף ב שככל שהמדגם גדל כך לרוב השגיאות קטנות, ולכן בפרט זה נכון עבור השגיאה המקסימלית והמינימלית. בנוסף, ניתן לראות שהשגיאה המינימלית יחסית קבועה וחסומה על ידי 0, ומשום שהשגיאה הממוצעת יורדת, השגיאה המקסימלית חייבת לקטון.

סעיף ה

סעיף ו

סעיף ז

ערכים אופטמיליים:

* ניסוי ראשון:
* ניסוי שני:

ההסבר לניסויים:

* בניסוי הראשון, אנו מגדילים את k בכל איטרציה, ככל שאנחנו מגדילים את k ככה יש יותר סיכוי שה- k-nn יתחשב בשכנים מאוד רחוקים ל-input שקיבלנו, ובפרט שכנים עם תווית שגויה.
* בניסוי השני, השגיאה הרבה יותר גבוהה, אבל נשים לב שככל ש-k עולה ככה השגיאה קטנה עד לנקודה מסוימת. מכיוון שה-training sample הושחת יש לו דוגמאות לא נכונות, אבל ככל שנגדיל את k נתחשב ביותר שכנים ולכן הסיכוי שניתקל בשכנים "אמיתיים" גדל ואז השגיאה תקטן עבור test samples לא מושחתים. מכיוון שקיימים test מושחתים אז טווח השגיאות גבוה בהרבה בהשוואה לניסוי הראשון, כי עבור test sample מושחת הסיכוי שהמדגם ייתן לו התווית הנכונה (שהיא למעשה שגויה) קלוש.

שאלה 3

סעיף א

סעיף ב

נגדיר בה"כ ש- lettuce = 1 ו- carrot = 0.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | 1 |
|  |  |
|  |  |

כלומר, .

סעיף ג

|  |  |  |
| --- | --- | --- |
| probability | preferred food | age |
| 0 | carrot | 7 |
| 60% | lettuce | 7 |
| 15% | carrot | 13 |
| 25% | lettuce | 13 |

סעיף ד

גם פה נגדיר בה"כ ש- lettuce = 1 ו- carrot = 0.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | 1 |

כלומר,

סעיף ה

למדנו בהרצאה ש-

כאשר

בשאלה זו, . לכן:

עלינו לחשב את לכל .

*ולכן סה"כ נקבל:*

*לסיכום, תוחלת שגיאת אלגוריתם ה memorize כאשר הוא:*

שאלה 4

סעיף א

נסמן

*כך ש- , ולכן*

*סעיף ב*

*נשים לב ש: לפי כלל בייס, ולכן לכל מתקיים ש-*

*סעיף ג*

*נסמן*

*נמצא את כפונקציה של . נשים לב ש- כאשר . כלומר הינו סכום של משתנים מקריים שמתפלגים ברנולית עם הסתברות . לכן, , וכידוע . מכאן - .*

*סעיף ד*

*סעיף ה*

*תחילה נבחין ש-*

*סעיף ו*

*יהי ,*

*משום ש- אז מתקיים שלכל אכן .*

*אז נתבונן במקרה ש- . נגדיר . נשים לב ש- . לכן . אזי נקבל ש-*

*כנדרש.*

*שאלה 5*

*סעיף א*

*יהיו כך ש- וגם . משום ש- בהכרח מתקיים ש- התפלגות דטרמיניסטית, כלומר שלכל קיים יחיד בתמיכה של . לכן מתכונת דטרמיניסטיות של מתקיים שלכל , מקיימת ש- הוא 0 או 1. נשים לב שמשום ש- מתקיים , ולכן  
 . נתון ש- עבור היא פונקציה על פי המרחב האוקלידי ולכן בפרט עבור מתקיים , אז סה"כ לפי  
 נקבל .*

*סעיף ב*

*נסמן . צריך להוכיח ש- , כלומר צריך להראות שלכל , . יהיו כך ש- .  
נניח בשלילה ש- , כלומר השכן הקרוב ביותר של – x במדגם הוא בעל תווית , נסמנו . כלומר .*

*אבחנה: המרחק . לפי הנתון בו לכל קיים עבורו , ובפרט עבור , ולכן מכיוון ש- מכסה את כל נובע ש- , שכן קוטר הכדור הינו .*

*לפי האבחנה, , אבל ולכן, . לפי ההנחה, היא התווית של ו- היא התווית של , כך ש- . אך , בסתירה להיות פונקציה לפי סעיף א.*

*(\*) הבהרה: הטענה שהוכחנו בסעיף א מתבססת על כל מדגם , ולכן אם נוכל להוסיפו למדגם ובכך אילוץ המרחק על יתקיים.*

*שאלה 6*

*סעיף א*

*נטען ש , נוכיח זאת על ידי שנראה שלכל קיימת , ולכל   
קיים . ינבע ש- תכיל לפחות את כל ההיפותזות של ולכן ינבע ששגיאת האפרוקסימציה לפי היא הקטנה ביותר.*

* *תהא , אז נגדיר . נשים לב ש- ולכן . לפי הגדרת , , וגם משום ש- נמצא באינדקס אי-זוגי ב-   
   נובע ש-*
* *תהא , אז נגדיר , נשים לב ש-*

*, ולכן . משום ש- באינדקס אי-זוגי ב- מתקיים:*

*סעיף ב*

*יהי זוגי כך ש- . יהי ומדגם בגודל .*

*נסמן ב - את ההיפותזה שחזרה מאלגוריתם nearest neighbor ביחס למדגם S. נראה תחילה ש-  
 . נסמן . נראה שקיימת היפותזה כך ש- . נגדיר .*

*המקרה הכללי:  
נגדיר להיות המספר ה-j בגודלו ב- אז נגדיר .*

*אבחנה: , וגם אם אז בהכרח .*

*נגדיר את לפי לכל באופן הבא:*

*אם אז נגדיר ולכל נגדיר כך ש- מסודרים בסדר ולה.*

*אם אז לכל , נגדיר כך ש- מסודרים בסדר עולה.*

*אם אז נגדיר ולכל נגדיר אם , ואם אז לכל .*

*אם אז לכל נגדיר אם , ואם אז לכל .*

*נשים לב, שאם אז לא מוגדר , ולכן ניתן להגדיר אותו כרצוננו בלי לפגוע בערכים קיימים. אבל, אם וגם אז בהכרח לפי האבחנה, ולכן במקרה זה לא נוסיף את ו- תישאר ב- .*

*אבחנה: לכל מתקיים ש- אי-זוגי אם ורק אם לכל ,  
 .*

*כעת נותר להראות ש- . יהי . נראה ש- . נסמן ב- את השכן הקרוב ביותר של במדגם S. נניח בה"כ ש- , אז .*

*נניח ש- מסודרים בסדר גודל עולה, ולהם סדרת לייבלים . והוא נמצא בתת-סדרה שכל ערכי שלה הם 1. נתבונן ב- וב- , שהן התווית הראשונה והאחרונה בתת הסדרה הזו בהתאמה.*

*אם אז בהכרח . לכן . נחלק לתתי מקרים:*

* *אם אז ולכן ולכן ו- .*
* *אם אז ולכן. לכן . אם אז x קרוב יותר ל- מאשר ובפרט לא השכן הכי קרוב שלו, בסתירה להנחה. לכן , ואז יתקיים .*

*אם אז בהכרח , אזי , ולפי אופן הגדרת עבור מתקיים ש-j אי-זוגי. בנוסף, נטען ש- , באופן דומה למקרה שבו ו- (שכן אחרת x יותר קרוב ל- מאשר ל - , ובפרט לא קרוב יותר ל- ). נחלק לתתי מקרים:*

* *, אז לפי הגדרת מתקיים ש- ומשום ש- באינדקס אי זוגי וכי וגם , אז .*
* *, זהו מקרה זהה ל- ו- , ובפרט , ולכן .*

*נותר להראות ש- , ואז יקיים את תנאי ה ERM עבור מחלקת ההיפותזות . למדנו בהרצאה בעל שגיאה 0 על המדגם כאשר לכל תווית במדגם S קיימת תווית יחידה נטען שהדבר מתקיים עבור כבשאלה בהסתברות 1. יהי , נראה שלכל מתקיים שההסתברות שבמיקום ה-i נבחר שכבר היה במדגם היא 0. נסמן ב- מ"מ שמייצג את הבחירה ה-i במדגם.*

*לכן ההסתברות שנראה שתי דוגמאות זהות היא 0 ולכן לפי מה שנלמד בהרצאה, היפותזת ה nearest neighbor תיתן שגיאה 0 על המדגם ותקיים את תנאי ה ERM עבור.*

*סעיף ג*

*יהי זוגי. אנחנו נראה שקיים מדגם בגודל , כך לא מקיימת את תנאי ה ERM עבור . נראה ש- , ואז לא מקיים את הדרוש.   
נגדיר: , כך שלכל ,  
 . נניח בשלילה שקיימת היפותזה כך ש- . נראה שלכל שתוויתו 1 במדגם S, קיים טווח אחד ויחיד שיכיל רק את עבור אי-זוגי. נבחין שקיימים במדגם S דוגמאות שתוויתם 1, וב-f קיימים טווחים שמכילים x יחיד. מעיקרון שובך היונים, קיים טווח שיכיל 2 איברים, בסתירה.*

*יהי שתוויתו 1 במדגם S, לכן כי היפותזה זו לא טועה על המדגם. אז קיים טווח כך ש- . נניח בשלילה שקיים בה"כ כך ש- . אם , הדבר לא אפשרי שכן אז בסתירה לכך שהתווית שלו 0 במדגם. אם אז לפי הגדרת S קיים כך ש- וגם ולכן גם , ואז בסתירה לכך שהתווית שלו היא 0.*

*הוכחנו שלא קיימת כך ששגיאתה על המדגם היא 0. בנוסף, לפי סוף סעיף ב, גם עבור S בגודל זה, משום שכל , ולכן בהכרח*