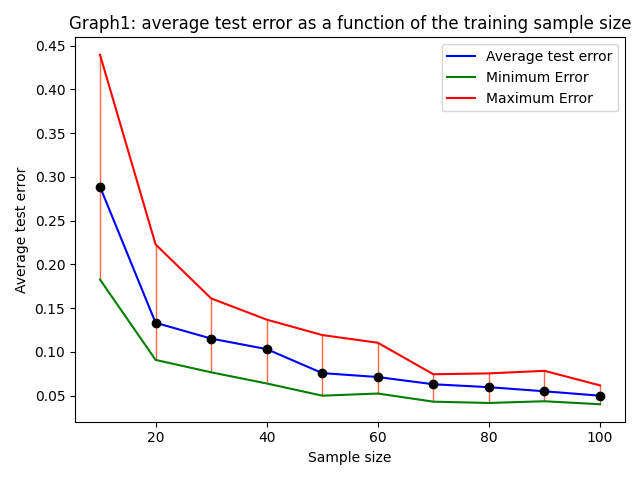
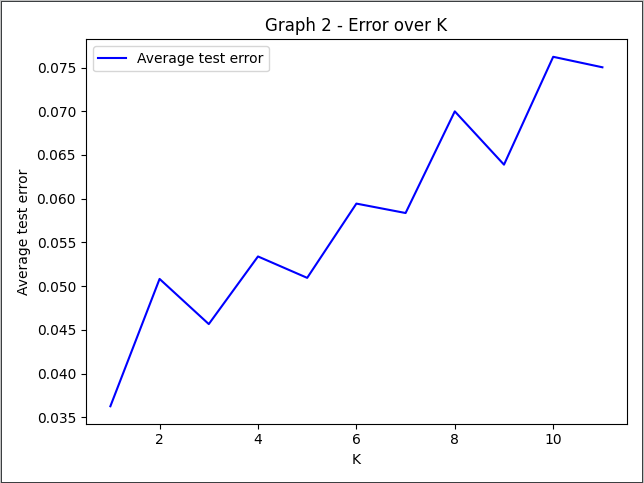
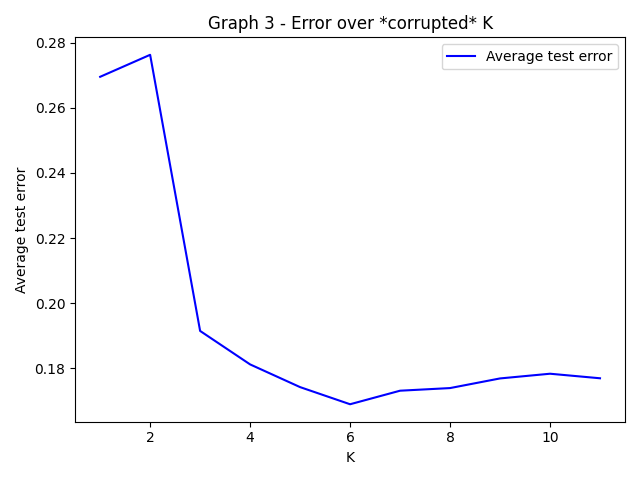
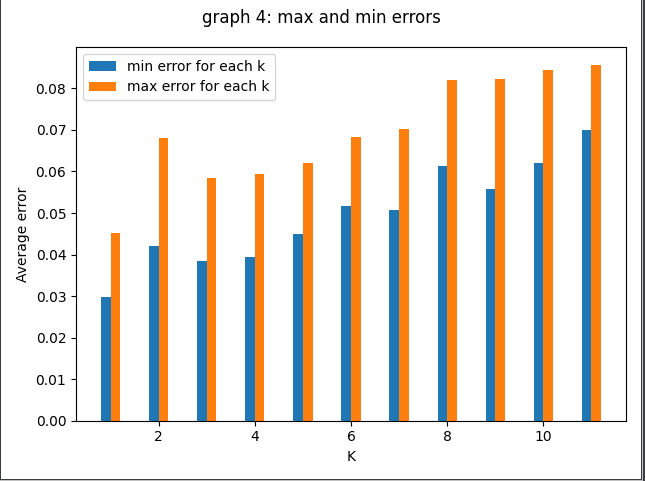
**Introduction to learning and analysis of big data**

**עבודה 1**

**שאלה 2 :**

1. 
2. כפי שניתן לראות בגרף 1 (הכחול), ככל שגודל המדגם גדל כך השגיאה קטנה. נשים לב כי המדגם נבחר רנדומלית ולכן ככל שהוא גדול יותר כך ההסברות שהוא מייצג טוב יותר את ההתפלגות גדלה (וההסתברות לoverfitting קטנה), ולכן על מדגם גדול יותר נקבל שגיאה ממוצעת קטנה יותר.
3. כפי שניתן לראות בגרף 1 (בירוק ובאדום), התקבלו תוצאות שונות עבור ריצות שונות על גדלי מדגם שווים. ההדים בין התוצאות נובעים מכך שהמדגם שלנו נבחר באופן רנדומלי וכי מבחירה רנדומלית של מדגם אנחנו עלולים לקבל מדגם שלא מייצג באופן טוב את ההתפלגות ואז לקבל שגיאה גדולה (לדוגמה מדגם בו כל התוויות זהות) ואידך אנחנו גם עשויים לקבל מדגם המייצג בצורה טובה את ההתפלגות ואז לקבל שגיאה נמוכה יותר.
4. כפי שניתן לראות בגרף 2, ככל שגודל המדגם גדל, בנוסף לירידת השגיאה הממוצעת, גם ההבדל בין שתי התוצאות קטן. כלומר ההפרש בין התוצאה המקסימלית לתוצאה המינימלית קטן. ניתן להסיק כי מגמה זו נובעת מהגדלת המדגם, שכן בפעולה זו ההסתברות לקבלת מדגם לא מייצג היטב את ההתפלגות (overfitting) קטנה.
5. 

.f



f. כפי שניתן לראות מגרף 2, כאשר k=1 קיבלנו את השגיאה המינימלית, ואילו בגרף 3 ניתן לראות כי השגיאה המינימלית מתקבלת עבור k=6 .

ההבדלים נוסעים מהסיבה שבניסוי השני (בגרף 3) 15% מהתיוגים במדגם ובtest set לא היו נכונים.

כאשר כל התיוגים נכונים, מאחר וקבוצת הטסט שלנו קטנה, מספיק לנו להתבונן בשכן אחד בשביל לקבל בהסתברות גבוהה את התיוג הנכון, אך כאשר חלק מהתיוגים אינם נכונים

ההסתברות לכך שיבחר שכן בעל תיוג "מושחת" (corrupted) גדולה יותר- גם בגלל שחלק מהשכנים (דוגמאות המדגם) לא מתויגים נכון ולא גם שהדוגמאות בטסט עצמו מתויגות לא נכון. כמו כן נשים לב שעבור 6-nn אנחנו לוקחים את התיוג שמופיע הכי הרבה פעמים בין 6 שכנין הקרובים אליו ביותר ובכך מורידים את הסיכוי שהתיוג יקבע בעקבות שכן "מושחת".

**שאלה 3:**

1. תהא והיו כך ש .

מהנתון כי להתפלגות D Bayes-error = 0 , ניתן להספיק כפי שנלמד בכיתה כי ההתפלגות D היא דטרמיניסטית. כמו כן בנוסף נתון כי ו מעל  *ולכן בהכרח .* ידוע גם כי של D *היא c-Liphschitz ולכן כפי שנלמד בכיצה מתקיים .*

*מהנתון נקבל כי ולכן בהכרח אז . נתונה כפונקציית המרחק האוקלידי ולכן .*

1. תהא נקודה במרחב .

מקרה 1: וההתפלגות דטרמיניסטית לכן מהגדרת אלגוריתם הNN הנקודה שתיבחר כקרובה ביותר היא הנקודה עצמה כך ש והאלגוריתם צדק.

מקרה 2: . תהא הנקודה להיות הנקודה הקרובה ביותר ל כך שאלגוריתם הNN ויבחר בה ויחזיר . נשים לב כי מנתוני השאלה בהכרח קיימת נקודה כזו (לכל  *קיימת נקודה כך ש ( בנוסף נשים שהמרחק המקסימלי בין שתי הנקודות הוא לכל היותר . כמו כן, לכן ומסעיף א' נקבל כי () אזי נקבל כי .*

כלומר הראינו כי לכל נקודה  *האלגוריתם מוצא את התווית המתאימה ולכן כנדרש.*

**שאלה 4:**

1. . (לשנות למכפלה קרטזית??)
2. נקבל כי:
3. כלומר התוצאה הטובה ביותר שאנחנו יכולים להשיג ממחלקת ההיפוטזות הנתונה. במקרה הנתון, כאשר מחלקת ההיפוטזות מורכבת מפונקציות קבועות ללייבל מסויים אז ה approximation error יהיה 8%+20% = 28% וזה עבור פונקציה הקבועה

1. אני חושב שלא.. כי לא נתון לי כלום בשביל לחשב את זה... צריך לבדוק עם טלי
2. אנחנו יכולים להשתמש בנוסחה עבור ''D כי התפלגות זו, (בשונה מהתפלגות D), היא דטרמיניסטית.

נשתמש בנוסחה ונקבל כי השגיאה היא:

**שאלה 5:**

1. שגשד
2. להשלים פתרון של סהר.
3. להשלים פתרון של סהר – בגדול בגלל שאמרו שh נבחר מתוך קבוצת היפוטזות שהורצו בERM אז בהכרח השגיאה שלו הנמוכה ביותר ואם בסעיף b ראינו שלכל היותר עבור a באותו התחום השגיאה לכל היותר אפסילון אז בהכרח גם במקרה הזה שהוא השגיאה לכל היותר אפסילון.