מגישים:
עומר גלץ - 318256120
יותם פלדמן - 315717181

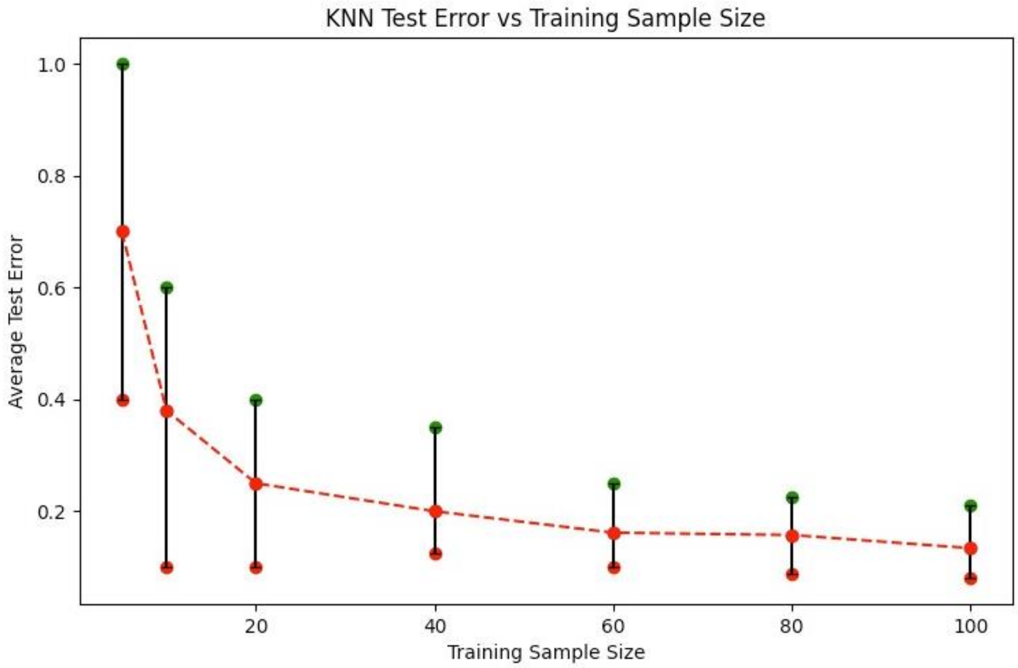
Introduction to Machine Learning

Exercise 1 – Answers

<u>שאלה 1:</u> בקובץ nearest_neighbour.py

<u>שאלה 2:</u>

a. פלט:

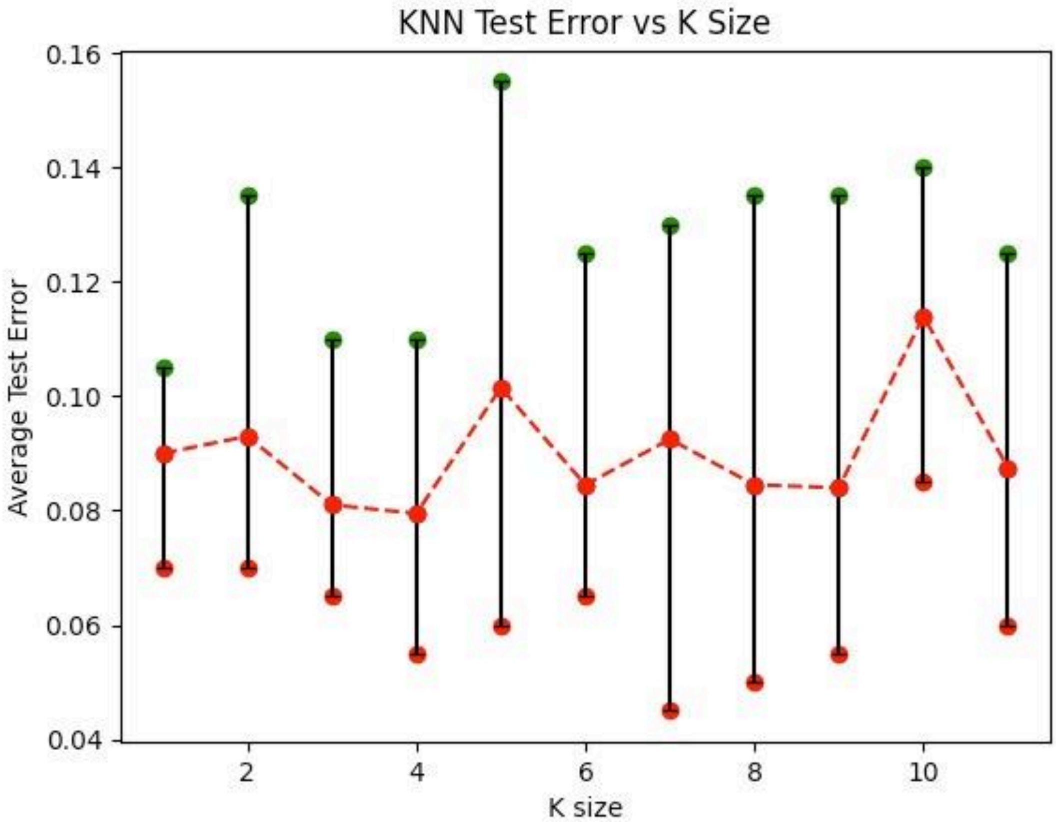


b. ניתן לראות בגרף המצורף שככל שאנו מגדילים את גודל המדגם, השגיאה הממוצעת קטנה. ניתן להסביר זאת מכמה סיבות.
ראשית, ככל שאנו מגדילים את גודל המדגם, ככה אנו יותר "דומים" להתפלגות האמיתית. כלומר, הסיכוי שנקבל מדגם שמחסיר סוגים שונים של לייבלים קטן.
שנית, מכיוון שאנו בודקים עבור הtest_data את המרחק מכל וקטור במדגם, הגדלת המדגם מגדילה את הסיכוי לקבל וקטור קרוב יותר עם תווית נכונה.

c. כן.
אנו עשויים לקבל תוצאות שונות עבור ריצות עם אותו גודל מדגם מכיוון שבכל ריצה בחירת המדגם מתבצעת באופן אקראי ולכן מדידת המרחקים של הtest_data עשויה להתבצע אל מול וקטורים שונים מהמדגם.

d. כן, הפרשי השגיאה (maxError - minError) קטנים ככל שגודל המדגם גדל.
בדומה לסעיף b, תופעה זו מעידה על כך שהמודל הופך ליותר יציב ועקבי כאשר הוא מאומן עם יותר דוגמאות. הגדלת המדגם מקרבת את ההתפלגות האמיתית בצורה טובה יותר, מה שמוביל לחיזוי אמין יותר ופחות משתנה.

e. פלט:



כפי שניתן לראות בגרף, השגיאה המינימלית התקבלה כאשר k=4.
במדגם בגודל 200 כאשר נרצה לבחון דוגמא חדשה בהסתברות גבוהה כבר פגשנו דוגמא קרובה לה בשלב האימון המסווגת נכון.
עבור K קטן, המודל רגיש לרעש בנתונים, אשר עלול להוביל לover fitting (למידה על מס קטן של שכנים הקרובים ביותר).
עבור K גדול, המודל עלול להוביל לunder fitting מהסיבה שהמודל שוקל בחישוב יותר מידי שכנים, שחלקם עשויים להיות די רחוקים מהאינפוט וזה יכול לדלל את ההשפעה של השכנים הקרובים והרלוונטיים ביותר ולהוביל לעלייה בשגיאות.
על כן, כדאי לבחור K מתאים בשביל לאזן את למידת המאפיינים אל מול רגישות יתר לרעש.

שאלה 3

נתון ש- $Y=\{0,1\}$, D התפלגות מעל $X \times Y$ כאשר ל- D יש $Bayes\text{-}error=0$, כלומר D דטרמיניסטית

לכן $\eta_y(x)\in\{0,1\}$ לכל $x,y\in X\times Y$.

$\eta_y(x)$ מוגדרת כ- C-Lipschitz , כלומר לכל $x_1,x_2\in Supp(D)$, $|\eta(x_1)-\eta(x_2)|\le C\cdot\rho(x_1,x_2)$.

(a) נתון כי $S\sim D^m$, יהי $(x_1,y_1),(x_2,y_2)\in S$ כך ש- $y_1\neq y_2$.

$(x_1,y_1),(x_2,y_2)\in S$ אז $x_1,x_2\in Supp(D)$.

$Y=\{0,1\}$
$y_1\neq y_2$
D דטרמיניסטית

לפי הגדרה

$$|\eta(x_1)-\eta(x_2)| = 1 \le C\cdot\rho(x_1,x_2) \iff \frac{1}{C}\le\rho(x_1,x_2) = \|x_1-x_2\|$$

בסה"כ קיבלנו: $\frac{1}{C}\le\|x_1-x_2\|$.

(b) נניח בשלילה כי $err(f_S^m,D)\neq 0$ אז קיימת $(x,y)\in S$ כך ש- $f_S^m(x)\neq y$.

$x\in X$ לכן קיים כדור $B^*\in\mathcal{B}$ כך ש- x נמצא בכדור ($x\in B^*$).

נתון כי לכל $B\in\mathcal{B}$ קיים $(x,y)\in S$ כך ש- $x\in B$ אז בפרט עבור B^* נקבל כי קיימת $(x',y')\in S$ כך ש- $x'\in B^*$.

קיבלנו כי $x',x\in B^*$ ולכן $\rho(x,x')\le\frac{2}{3c}$ (כאשר $\frac{2}{3c}$ הוא קוטר הכדור B^*)

נבחין כי $\frac{2}{3c}\le\frac{1}{c}$ ולפי סעיף (a) , $\rho(x',x)\ge\frac{1}{c}$, בסתירה.

שאלה 4:

(a) $\mathcal{X}=\{(x_1,x_2) \mid x_1\in[0,42],\ x_2\in[0,5]\}$, $\mathcal{Y}\in\{black, white\}$

(b) $\eta_b(8,4)=0.42 \quad \eta_w(8,4)=0.06 \Leftarrow h_{bayes}(8,4)=black$

$\eta_b(15,1)=0.21 \quad \eta_w(15,1)=0.07 \Leftarrow h_{bayes}(15,1)=black$

$\eta_w(15,2)=0.24 \Leftarrow h_{bayes}(15,2)=white$

(c) ראינו בכיתה כי $error(h,D)=\sum_{x\in\mathcal{X}} Pr(X=x)\left(1-\eta_{h_{bayes}(x)}(x)\right)$ לכן:

$$error(h,D) = Pr(8,4)\cdot\left(1-\eta_{h_{bayes}(8,4)}(8,4)\right) + Pr(15,1)\cdot\left(1-\eta_{h_{bayes}(15,1)}(15,1)\right) + Pr(15,2)\cdot\left(1-\eta_{h_{bayes}(15,2)}(15,2)\right)$$

ע"פ (b)

$$= Pr(8,4)\cdot\left(1-\eta_{black}(8,4)\right) + Pr(15,1)\cdot\left(1-\eta_{black}(15,1)\right) + Pr(15,2)\cdot\left(1-\eta_{white}(15,2)\right)$$

$\eta_y(x)=\frac{Pr(Y=y, X=x)}{Pr(X=x)}$

$$= 0.48\cdot\left(1-\frac{0.42}{0.48}\right) + 0.28\cdot\left(1-\frac{0.21}{0.28}\right) + 0.24\cdot\cancel{\left(1-\frac{0.24}{0.24}\right)}^{0} = \underline{\underline{0.13}}$$

Approximation error : $\quad \mathrm{err}_{\mathrm{app}} := \min_{h\in\mathcal{H}} \mathrm{err}(h,\mathcal{D})$.

(d) נשים לב כי ישנן רק שתי תוויות אפשריות לכן נסיק כי:

$\mathcal{H}=\{h_1,h_2\}$ $\quad h_w(x)=white$, $h_b(x)=black$

נחשב את השגיאה עבור שתי הפונ' h_1,h_2 וניקח את המינימלית מביניהן להיות שגיאת הקירוב:

$$err(h_b,D) = \sum_{x\in\mathcal{X}} Pr(X=x)\left(1-\eta_{h_b(x)}(x)\right) =$$

$$= Pr(8,4)\cdot\left(1-\eta_{h_{b(8,4)}}(8,4)\right) + Pr(15,1)\cdot\left(1-\eta_{h_{b(15,1)}}(15,1)\right) + Pr(15,2)\cdot\left(1-\eta_{h_{b(15,2)}}(15,2)\right)$$

$$= Pr(8,4)\cdot\left(1-\eta_{black}(8,4)\right) + Pr(15,1)\cdot\left(1-\eta_{black}(15,1)\right) + Pr(15,2)\cdot\left(1-\eta_{black}(15,2)\right)$$

$$= 0.48\cdot\left(1-\frac{0.42}{0.48}\right) + 0.28\cdot\left(1-\frac{0.21}{0.28}\right) + 0.24\cdot(1-0) = 0.13+0.24 = 0.37$$

$$err(h_w,D) = \sum_{x\in\mathcal{X}} Pr(X=x)\left(1-\eta_{h_w(x)}(x)\right) =$$

$$= Pr(8,4)\cdot\left(1-\eta_{h_{w(8,4)}}(8,4)\right) + Pr(15,1)\cdot\left(1-\eta_{h_{w(15,1)}}(15,1)\right) + Pr(15,2)\cdot\left(1-\eta_{h_{w(15,2)}}(15,2)\right)$$

$$= Pr(8,4)\cdot\left(1-\eta_{white}(8,4)\right) + Pr(15,1)\cdot\left(1-\eta_{white}(15,1)\right) + Pr(15,2)\cdot\left(1-\eta_{white}(15,2)\right)$$

$$= 0.48\cdot\left(1-\frac{0.06}{0.48}\right) + 0.28\cdot\left(1-\frac{0.07}{0.28}\right) + 0.24\cdot\cancel{\left(1-\frac{0.24}{0.24}\right)}^{0} = 0.63$$

קיבלנו כי שגיאת הקירוב היא $err(h_b,D)=0.37$.

המשך שאלה 4:

$$\mathbb{E}_{S\sim\mathcal{D}^m}[\mathrm{err}(\hat{h}_S,\mathcal{D})] = \frac{k-1}{k}\sum_{x\in\mathcal{X}} p_x(1-p_x)^m.$$

(e)

נתונים: $m=3$, $k=2$ (black, white)

$$E[err(h, D'')] = \frac{1}{2}\cdot\sum_{x\in X} P_x(1-P_x)^m$$
$$= \frac{1}{2}\cdot\left[P_{(8,4)}(1-P_{(8,4)})^3 + P_{(9,5)}(1-P_{(8,5)})^3 + P_{(9,2)}(1-P_{(9,2)})^3 + P_{(11,6)}(1-P_{(11,6)})^3\right] =$$
$$= \frac{1}{2}\cdot\left[0.08(1-0.08)^3 + 0.15(1-0.15)^3 + 0.47(1-0.47)^3 + 0.3(1-0.3)^3\right] = \underline{\underline{0.163}}$$

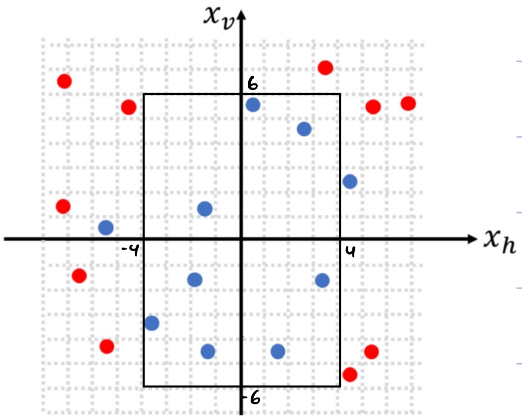
נשים לב כי להתפלגות D'' יש תוויות דטרמיניסטיות (בשונה מהתפלגות D) ולכן ההסתברות לטעות תלויה ב-M_S שלמדנו בכיתה (missing mass - ההסתברות שנקבל x שלא נצפה באימון).

בהסתברות לא דטרמיניסטית D, האלג' ה-Memorize לא יכול לחזות בצורה אמינה את התוצאות של הדוגמאות בהן צפה (יתכן כמה תוויות עבור דוגמאות בודדות), על כן, שימוש בנוסחה של תוחלת השגיאה של אלג' ה-Memorize לא ניתנת לשימוש בתצורה אשר נלמדה בכיתה.

שאלה 5:

$S=\{(0.5,5.5),(2.7,4.5),(4.5,2.5),(-1.5,1.2),(-5.5,0.5),(-1.9,-1.5),(-3.5,-3.2),(-1.3,-4.5),(3.3,-1.6),(1.5,-4.5),$

$(3.5,7),(5.5,5.5),(6.9,5.6),(-7.1,6.5),(-4.5,5.5),(-7.2,1.5),(-6.5,-1.5),(-5.5,-4.2),(5.3,-4.5),(4.5,-5.5)\}$



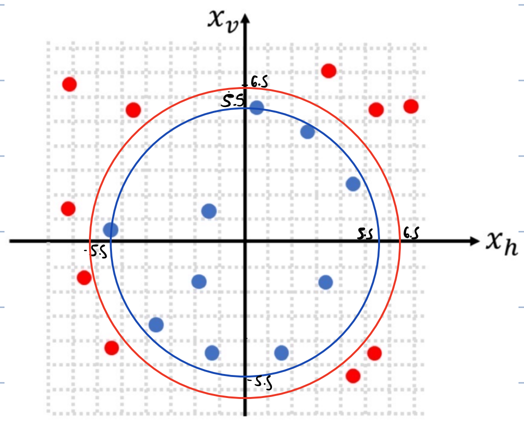
(a) לפי הנתונים, הספים האופטימלים יהיו $T_h=4$, $T_v=6$

$H_{rec}:=\{f^{rec}_{T_h,T_v} \mid T_h=4, T_v=6\}$

לכן אלג' ERM ימצא לנו את המלבן המורכב מהצלעות $X_h=4$, $X_v=6$

$\widehat{err}(h,S)=\frac{1}{20}\cdot\sum_{i=1}^{20}\mathbb{I}[h(x_i)\neq y_i]=\frac{1}{20}\cdot 2=\frac{1}{10}$

שתי הנק' הכחולות שמחוץ לסף שבגרף



(b) נגדיר את g להיות משוואת מעגל: $g(X_h,X_v)=\sqrt{X_h^2+X_v^2}$

$H_g:=\{f_{g,T} \mid T\in[5.5,6.5]\}$

- ניתן לראות כי עבור $T\in[5.5,6.5]$, H_g יחזיר לנו שגיאה אמפירית השווה ל-0 בהינתן המדגם הנוכחי.
- ניתן להניח כי H_g תספק פרדיקטור בעזרת ERM שטוב יותר מאשר H_{rec} מאחר וקביעת הסף ע"י רדיוס המעגל מתאימה יותר להנחת החקלאי אשר בודק את מרחק התפוחים מגזע העץ, לכן כאשר נקבל תפוח חדש שאינו במדגם נרצה לחזות את סיווגו בהתאם למרחק שלו מגזע העץ ללא תלות בזווית מהגזע.