IML ex3

אלמוני תמר

1) If we know D, our best predictor would have been assigning the class with higher prob:

$$\forall x \in \mathcal{X} \quad h_D(x) \begin{cases} +1 & Pr(y=1|x) \geq \frac{1}{2} \\ -1 & \text{else} \end{cases}$$

where the probability is over D. This classifier known as the Bayes optimal classifier. show that:

$$h_D = \arg\max_{y \in \{\pm 1\}} Pr(x|y) Pr(y).$$

מעניין גם שמתקיים

$$\arg\max_{y \in \{\pm 1\}} Pr(x|y) \cdot P(y) = \arg\max_{y \in \{\pm 1\}} P(y|x) \cdot P(x)$$

$P(x)$ הוא ההסתברות-על לכן הוא הקבוע ולכן שפיר גם המרחקים בשביטוי שנתן לי כי המקסימום max עם לכן

$$= \arg\max_{y \in \{\pm 1\}} P(y|x) = \arg\max(Pr(y=1|x), Pr(y=-1|x))$$

עכשיו שאנחנו מקבלים הסתברות ל מסלול, כמו שמסלול אחרי עם $x \in X$ ל הסתברות ל הוא מקבל x אם D הוא יותר מתכנס לכיוון אחד מ+ מאחר מסלול כ 1- כמו שמקבלים הם $y \in \{\pm 1\}$ אז

$$1 = Pr(y=1|x) + Pr(y=-1|x) \quad \text{אז} \quad \frac{1}{2} \leq Pr(y=1|x)$$

$$\frac{1}{2} \leq Pr(y=-1|x)$$

לכן לא יכול מקרה הם לא מתקיים אשר מראים את התוצאה

כאן שאם $y=1$ הם $\frac{1}{2} \leq Pr(y=1|x)$ אז אם ה x (בדגם) כ 1. אבל $y=-1$ המקרה $\frac{1}{2} \leq Pr(y=-1|x)$ אז ה x בדגם הוא כ 1- שנתן הכולל בנוסף $h_D(x)$ אז

(2) נניח ש $X=\mathbb{R}^d$, $x|y \sim N(\mu_y, \Sigma)$, mean vector $\mu_y \in \mathbb{R}^d$

covar מטריצה $\Sigma \in \mathbb{R}^{d\times d}$

$$f(x|y) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \cdot \exp\left\{-\frac{1}{2}(x-\mu_y)^T \Sigma^{-1} (x-\mu_y)\right\}$$

כאשר f פונקציה הצפיפות להתפלגות הנורמלית (בעלת השונות)

נרצה לבנות שני הנו יוקס עבור $\mu_{+1}, \mu_{-1}, \Sigma$ אלגוריתם Bayes optimal classifier יהיה

$$h_B(x) = \arg\max_{y\in\{\pm1\}} \delta_y(x)$$

כאשר δ_{+1}, δ_{-1} הם פונקציות $\mathbb{R}^d \to \mathbb{R}$ המקיימות

$$y\in\{\pm1\} \quad \delta_y(x) = x^T \Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T \Sigma^{-1}\mu_y + \ln Pr(y)$$

הסבר: נשים לב שבגלל B אנו מעוניינים בקבל בין ההסתברות הגלובלית $f_N(x|y) = Pr_B(x|y)$ (B מסמן אותה מרחיק שאנו משתמשים Z שלא נכון לגבי לפיצע ו עקיר תוחלת μ_y אותה כאשר f הוא פונקציה הצפיפות מעל התפלגות B (המאוחדת) $Pr(y)$ יקרה. מעניינו) נשים לב שזו שגיאה מעשה הקרן אותו יקרה מהא ההיבטים B, מהנו פונקציה הצפיפות ההתפלגות $f(x|y)$ אלא גם יוקרה מהא $Pr(y)$.

מהמון אם כדי הקרן:

$$h_B(x) = \arg\max Pr(x|y)Pr(y) = \arg\max_{y\in\{\pm1\}} f(x|y)Pr(y) =$$

ln(x) הוא פונ' מונוטונית עולה, לכן אם נפעל נשמר מקסימום $y\in\{\pm1\}$

$$= \arg\max \ln(f(x|y)\cdot Pr(y)) = \arg\max \ln(f(x|y)) + \ln(Pr(y)) =$$

כאשר $\arg\max_{y\in\{\pm1\}} \ln(Pr(y)) = \arg\max Pr(\ln(y))$ כי לא תלוי בערך x אשר לכן לגבי $y\in\{\pm1\}$

$$= \arg\max \ln\left[\frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left\{-\frac{1}{2}(x-\mu_y)^T \Sigma^{-1}(x-\mu_y)\right\}\right] + \ln(Pr(y)) =$$

(A)

$$\arg\max_{y\in\{\pm1\}} \left[\ln\left(\frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}}\right) + \ln\left(\exp\left\{-\frac{1}{2}(x-\mu_y)^T \Sigma^{-1}(x-\mu_y)\right\}\right)\right] + \ln(Pr(y)) =$$

נשים לב ש (A) הוא קבוע כאשרכים שם Σ הוא מטריצה הפיכה ו לכן לא נשפיע על הערך המקסימלי $y\in\{\pm1\}$

$$\arg\max_{y\in\{\pm1\}} \left[-\frac{1}{2}(x-\mu_y)^T \Sigma^{-1}(x-\mu_y)\right] + \ln(Pr(y))] =$$

$$= \arg\max_{y\in\{\pm1\}} \left[\underbrace{-\frac{1}{2}x^T\Sigma^{-1}x}_{\text{לא תלוי ב } y} - \frac{1}{2}\mu_y^T\Sigma^{-1}x + \underbrace{\frac{1}{2}x^T\Sigma^{-1}\mu_y}_{\text{transpose}} - \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y\right] + \ln(Pr(y)) =$$

$$\arg\max_{y\in\{\pm1\}} \left(x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y\right) + \ln(Pr(y)) = \delta_y$$

נדרש

$y \in \mathbb{R}^d,\ y_i \in \{\pm 1\}$

(3) בהינתן נתוני אימון קבלו, אותם נוכל לחלק ל $\Pr(y), \Sigma, \mu_{-1}, \mu_1$
לפי הנוסחאות בהרצאה. Training set $S = (x_1, y_1), \dots, (x_m, y_m)$ אותם נקבל לכל אחד.
אז קיבלנו המטריצה $\Pr(y), \Sigma, \mu_{-1}, \mu_1$ בהסתמך על S

הסבר הסתברות

$$\Pr(y) = \frac{1}{m} \sum_{y \in \{\pm 1\}} \{\#i \mid y_i = y\} = \begin{cases} \Pr(y=1) = \frac{1}{m} \{\#i \mid y_i = 1\} \\ \Pr(y=-1) = \frac{1}{m} \{\#i \mid y_i = -1\} \end{cases}$$

הסבר תוחלת

נסמן ב $\#y_i$ את הקבוצה $\{\#i \mid y_i = y\}$

$$\hat{\mu}_1 = \frac{1}{\#y_1} \sum_{i=1}^{m} \{(x_i) \text{ s.t. } y_i = 1\}$$

$$\hat{\mu}_{-1} = \frac{1}{\#y_{-1}} \sum_{i=1}^{m} \{x_i \text{ s.t. } y_i = -1\}.$$

נשים לב שעבור כל דגימה x_i בתוחלת אם הדגם y_i שלו נמצא ב $y=1$ אז נוסיף אותו לחישוב תוחלת μ_1 אם הדגם $y_i = -1$ נוסיף אותו לחישוב תוחלת μ_{-1}

הסבר מטריצת covariance (הנקודה 2 עמ' 32)

$$\hat{\Sigma} = \frac{1}{m} \sum_{y=\pm 1} \sum_{\substack{i=1 \\ s.t. \\ y_i = y}}^{m} (x_i - \hat{\mu}_y)(x_i - \hat{\mu}_y)^T$$

(4) מסקנה

ה classifier אותו נוכל לבנות בשלב הבא:

(א) לפי שיטת בייס אם ההסתברות גדולה מ 0.5

(ב) לפי שיטת כל 1 מספר שבשלב ההיסטוגרמה הוא בין. אבל יש בעיה לדוגמא כאשר n גדול אז הוא הנוסחה (ב) לכן, נפתרון הבעיה ע"י ההבדל בין 2 (לא מסקנה) נבדוק אם המרחק בין 1 (הנ"ל) גדול מ ההתפלגות מסקנה). (כי אין כמו הנ"ל אבל נפתרון לבעיה שלא 2 שקיים יכול להיות אז נקבל הסתברות השייכת לעיר אחרת, כי הוא גדול יותר) יש יתרון מסקנה ויכולים לעשות.

SVM - formulation

(5) Quadratic Program (QP) Canonical form $\arg\min_{v\in\mathbb{R}} \left(\frac{1}{2}v^T Q v - a^T v\right)$ s.t. $Av \le d$

$Q \in \mathbb{R}^{n\times n}$, $A \in \mathbb{R}^{m\times n}$, $a \in \mathbb{R}^n$, $d \in \mathbb{R}^m$

נרצה לנסות שה SVM hard margin בצורה הקנונית של בעיית QP

ע"מ שנוכל לפתור שפתרון עליהן תוכל לעזור לנו בשלב הבא.

$$\arg\min_{(w,b)} \|w\|^2 \quad s.t. \quad \forall i,\ y_i(\langle w, x_i\rangle - b) \ge 1$$

נרצה לבנות מטריצה Q, a, v לפי הנוסחה הקנונית לגבי w

ו b לכלול ב- v המטריצה A ו d שיביאו ל $Av \le d$

בלומר נשאר: $\arg\min \|w\|^2 \to$ $\arg\min\left(\frac{1}{2}v^T Q v - a^T v\right)$ s.t. $\forall i\ y_i(\langle w, x_i\rangle - b) \ge 1 \to Av \le d$

(1) $\arg\min \|w\|^2 = \langle w, w\rangle = w^T w = w^T I w$

ולכן מצד $w^T I w = \frac{1}{2}v^T Q v - a^T v$ יהיה: $a=0,\ Q=I,\ \arg\min v = w$

כל שנשאר לנו לבדוק ב $\frac{1}{2}$ אחרונה בנוסף: $\arg\min \frac{1}{2}\|w\|^2 = \frac{1}{2}w^T I w$

(2) $\forall i\ y_i(\langle w, x_i\rangle - b) \ge 1 \iff \forall i\ y_i\langle w, x_i\rangle - y_i b \ge 1$ s.t.

כי יש לנו $w, x_i \in \mathbb{R}^n$, y_i, b ו $\langle w, x_i\rangle, b \in \mathbb{R}$ אזי

$$= y_i w^T \vec{x}_i - y_i b \ge 1 = \begin{bmatrix} y_i x_i & y_i \end{bmatrix} \begin{bmatrix} w \\ b \end{bmatrix} \ge \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} =$$

$$\begin{bmatrix} (y_1 x_1) & y_1 \\ \vdots & \vdots \\ (y_m x_m) & y_m \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ \vdots \\ w_m \\ b \end{bmatrix} \ge \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \iff$$

$$= \underbrace{\begin{bmatrix} (y_1 x_1) & y_1 \\ \vdots & \vdots \\ (y_m x_m) & y_m \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} w_1 \\ \vdots \\ w_m \\ b \end{bmatrix}}_{v} \le \underbrace{\begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \end{bmatrix}}_{d}$$

נשים לב שגם לגבי $\frac{1}{2}w^T I w$ אנחנו נשים ב $\|w\|^2$ $\frac{1}{2}w^T I w$

הוא חלק, ולכן אנחנו לא נצטרך לדאוג לשים בלוקים, אבל חלק.

לגבי נקודת הפתרון w בעצמו לא ישנה שלפני $\|w\|^2$ יהיה מינימום כך

ג"ל הוא ב $\|w\|^2$ יגיע לאותה נקודה שבה הבעיה המקורית מתנהגת

$\frac{1}{2}w^T I w$ s.t. $-y \cdot x^T(w) \le -\vec{1}$: $(b/w = w)$

(b/w)

⑥ In the soft-SVM we defined the problem:

(I) $\min_{w,\xi_i} \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m}\xi_i$ s.t. $\forall i\ y_i\langle w, x_i\rangle \geq 1-\xi_i$ and $\xi_i \geq 0$

כדי להראות שהבעיה הזו שקולה לבעיה הבאה קבה:

(II) $\min_{w} \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum \ell^{hinge}(y_i\langle w, x_i\rangle)$

כך ש: $\ell^{hinge}(a) = \max\{0, 1-a\}$

אפיון את השגיאה של כל דגם ואמר שהם שקולים באמצעות
פשוט הנסיון לשפר בעיה אחת לעומת $\frac{\lambda}{2}\|w\|^2$ ו
ולכן נוכל לשכתב כמה ביטויים ב תהיה כמו:

(I) מבוא: $\frac{1}{m}\sum_{i=1}^{m}\xi_i$ s.t. $\forall i\ y_i\langle w, x_i\rangle \geq 1-\xi_i$ and $\xi_i \geq 0$:

בהקשר הזה בדוגמה הנוכחית לדגם אחד שנדרש למבוא את
w המינימלי ושגיאת אז יש רק ξ_i אחד הקטנה ביותר מ w.
אז לאחר ש: $y_i\langle w, x_i\rangle \geq 1-\xi_i$ אם $\xi_i \geq 0$ כדי שמבין אחד
w אפשר יהיה separable hyperplane בין לדגם
מ מתוך הנתונים הנוכחי אפשר הרבה אחרים ולכן $\xi_i \geq 0$
אין שום סיבה לתת יותר מדי ξ_i.
כלומר, אם ההקשר הזה שיש נקודות מסוימות יותר מדי אפשר לבחור
רק, נוכל אולי הקרן יש אפשר מרחק בין כמות הרעש
שבה אפשרי כאשר הרעש שנאבד את הגלים מרעש אם אין
כמות הרעש שאפשר מרחק $\frac{1}{m}\sum\xi_i$ לתת כמה נקבה שהם תוצאה הכי
גדול.

(II) $\frac{1}{m}\sum \ell^{hinge}(y_i\langle w, x_i\rangle)$

נשתמש אם הקבוע נשים לו שנקבל:

$\frac{1}{m}\sum \ell^{hinge}(y_i\langle w, x_i\rangle) = \frac{1}{m}\sum \max\{0, 1-y_i\langle w, x_i\rangle\}$

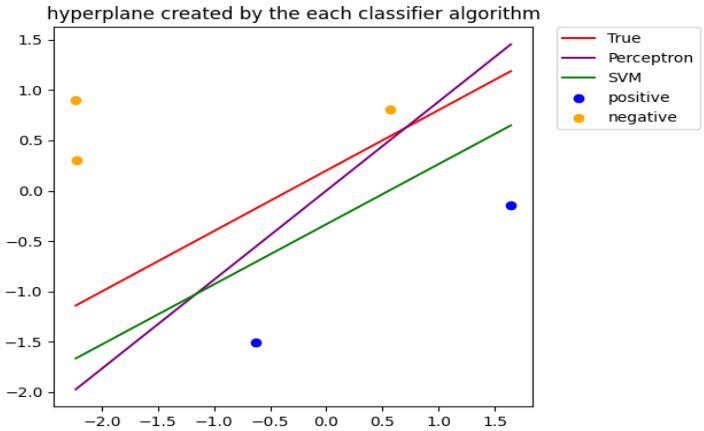
נשים לב שאם נקבל ב w את ξ_i שבגבוה הכי max נקבל ξ_i
לכן ξ_i אלף. כלומר כן, אם נסתכל ב הבעיה (I) ונסתכל
נקבל ש: $\xi_i \geq 1-y_i\langle w, x_i\rangle$ מה ξ_i $\Leftrightarrow$ $\xi_i \geq 0$ ו $1-\xi_i \leq y_i\langle w, x_i\rangle$

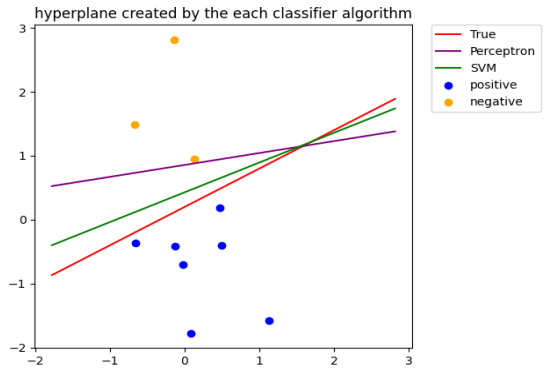
לכן הפתרון המקסימלי ב (הנתון) ℓ^{hinge} יהיה אותו המקסימום
ב (הנתון) ל הבעיה אחר ל הבעיה separable hyperplane עם מרחק לבעיה
ל הבעיה הקשה זו (I)

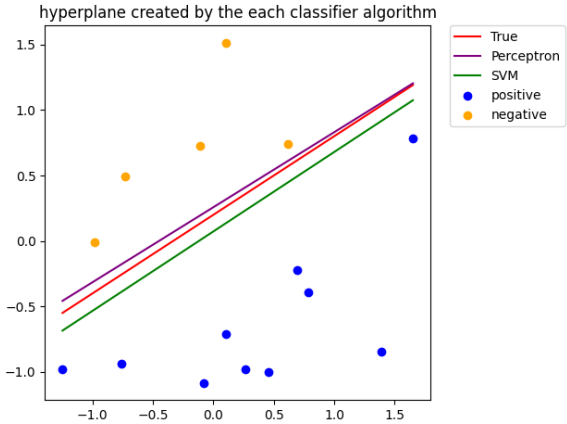
IML ex3 classifiers

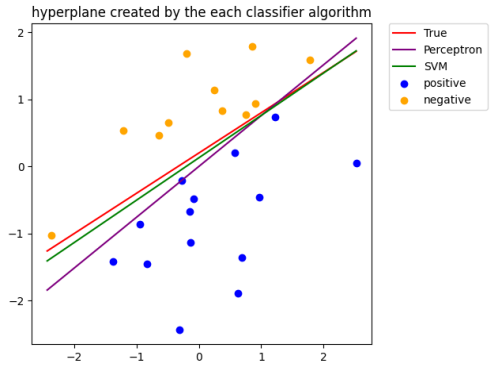
אביב אוחיון 313410458

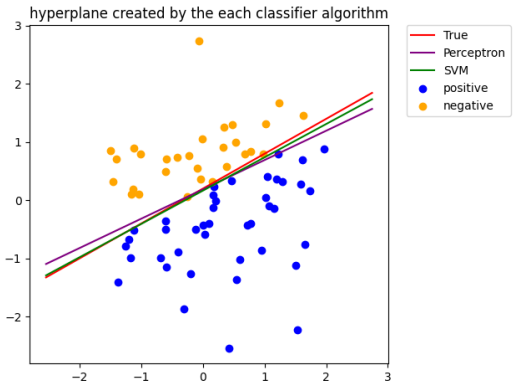
שאלה 9:



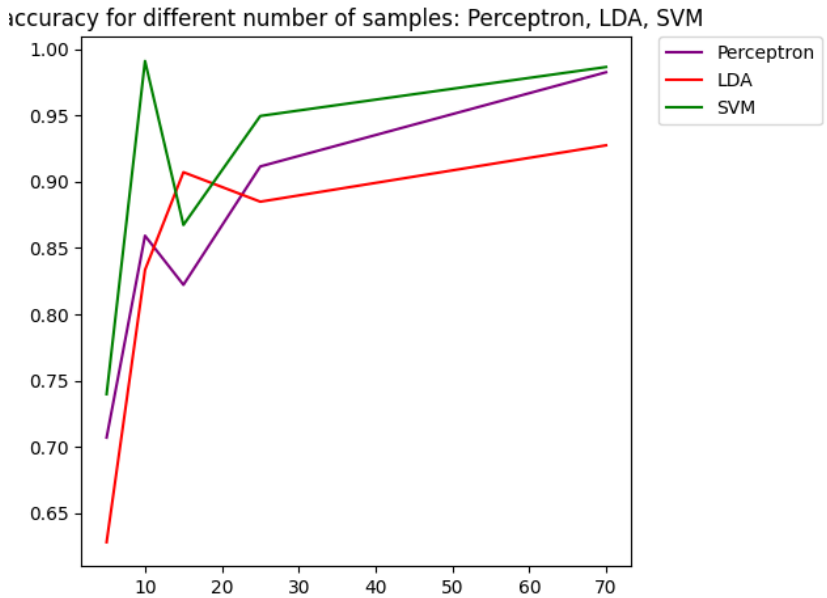






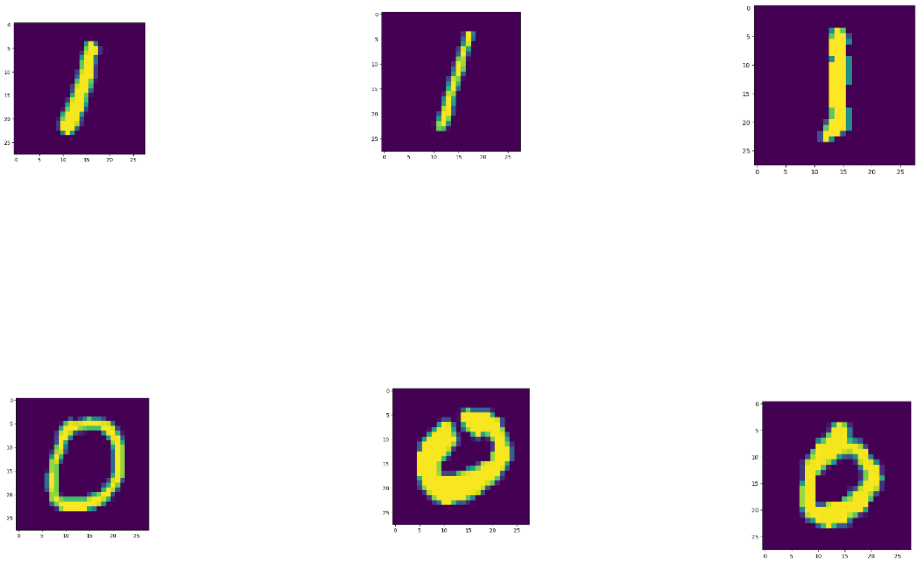


שאלה 11 + 10:



קלספייר ה SVM הוא הטוב ביותר לדעתי. הסיבה לכך היא שהמעקב הגדרתו הוא אינו מניח רליזאבליות על הדטא כלומר הוא "יודע" שאת הדטא לווא דווקא אפשר להפריד לינארית ופשוט מצמצם את טווח הטעות שאנחנו נגדיר לו בניגוד לperceptron שמניח ראלזיבליות (יוצר את המודל באופן איטרטיבי לפי ה training שלו ללא התחשבות בדגימות עתידיות שיגיעו) וLDA שמניח שעובד לפי משערכים הסתברותיים (כגון שונות משותפת לדאטא) ותלוי מאוד בסט הדגימות שהוא מקבל לאימון עקב כך בצורה רנדומית מההתפלגות שניתנה לו (במקרה זה נתנו לו את f שהיא דטרמניסטית ובפרט לא ההתפלגות הנורמאלית שהוא מצפה לה). עקב כך, ככל שכמות הדגימות עולה, ככה גם ההנחות על 2 האלגוריתם הללו הופכות ליותר קשות להבטחה ועל כן הניבוי שהן יניבו עלול להיות שגוי בעוד שSVM לא "בונה" על ההבטחות האלו ורק מנסה לצמצם את כמות השגיאה לפי הדגימות שניתנו לו לפי עקרון הלמידה של ה max margin. זו הסיבה שאנחנו רואים בנוסף לכך כי LDA מכמות מסויימת של דגימות נישאר יציב ונמוך יותר בעוד ש SVM מאותה כמות דגימות רק עולה ביציבות.

שאלה 12:



שאלה 14:

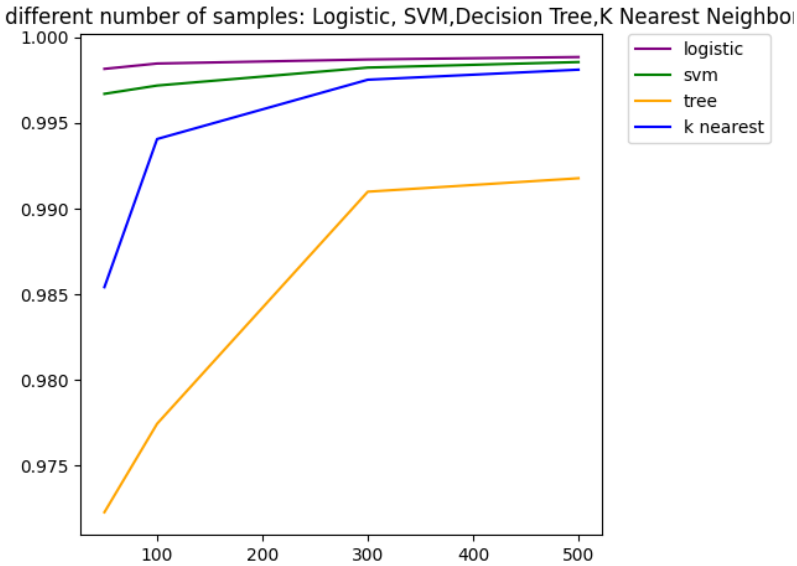
זמני הריצה של כל קלסיפייר (שמאל: קטן, ימין: גדול) הינם:

Logistic mean elapsed time: [0.06364116 0.0705077 0.09678286 0.12411702]

Svm mean elapsed time: [0.07811794 0.10009616 0.15179591 0.18879328]

tree_mean_elapsed_time: [0.01249707 0.01499523 0.02599104 0.04071514]

k nearest mean elapsed time: [0.15015065 0.15144804 0.1791475 0.21961463]



הדברים שהבחנתי בהם בהבדלים בין זמני הריצה של האלגוריתם הם:
ראשית, k nearest הינו הקלסיפר שלוקח הכי הרבה זמן, דבר זה אינו מפתיע עקב העובדה שהוא על כל דגימה, עובר על כל סט הדגימות על מנת להכריע מי השכנים הקרובים אליו ביותר. נשין לב שבנוסף עץ ההחלטה עובד על פיצור המרחב וצמצומו עד עומק קבוע ולפי כך הוא יוצר את הקלפסיפקיציה, עקב כך אומנם החישוב עצמו יהיה מהיר כי ההכרע במחינה חישובית היא די קלה (סך הכל להגיד אם משהו גדול ממשהו) לעומת השאר אבל כיוון שעומק העץ הוא זה שיכריע את הקלסיפקיה בהינתן עומק לא גדול במיוחד הכרעה של העץ צפוייה להיות לא תמיד הכי מדוייקת לעומת השאר. הSVM משתמש בכך שהוא פותר את בעייה שהיא אופטימוזציה ריבועית שיש לה מקסימום והיא חלקה, עקב ובשבל העובדה ש SVM הוא איטרטיבי על סט הדגימות (להבדיל מ k nearest לדוגמא) הוא גם עובד יחסית מהר ושילוב של 2 הסיבות מסביר למה הוא גם נותן את התוצאות הכי טובות.

הערה: בסעיף זה השתמשתי בפוקיציות הפסרייה של sklearn על מנת מנת לא לשנות את הלייבלים ל 1 0 בסעיף זה.