:Linear Algebra

.1

תהי A מטריצה סימטרית שהיא PSD.
מכיוון ש-A סימטרית מתקיים כי קיימת מטריצה Q אורתוגונלית כך ש $A = QDQ^T$ כאשר D אלכסונית (ובאלכסון נמצאים הערכים העצמיים של A).
מתקיים כי הע"ע של A אי-שליליים:
יהי λ ע"ע של A ו-v ו"ע מתאים, מתקיים כי $v^T Av = v^T \lambda v = \lambda \|v\|_2^2 \geq 0 \Rightarrow \lambda \geq 0$
לכן נוכל להגדיר מטריצה חדשה D' באופן הבא: $D'_{ij} = \sqrt{D_{ij}}$.
מכיוון שD אלכסונית, מתקיים כי $D'D' = D$.

לכן

$$A = QDQ^T = QD'D'Q^T = \left(D'^T Q^T\right)^T D'Q^T = (D'Q^T)^T D'Q^T$$

וזה מה שרצינו להוכיח.

תהי A מטריצה כך שניתנת לפירוק באופן הבא: $A = X\,X^T$.
נראה כי היא סימטרית:

$$A^T = (XX^T)^T = XX^T = A$$

כעת נראה כי היא PSD:
יהי $v \in \mathbb{R}^n$ כלשהו. מתקיים כי

$$v^T Av = v^T XX^T v = (X^T v)^T X^T v = \sum_{i=1}^{n} X^T v_i^2 \geq 0$$

לכן A היא מטריצה סימטרית ו-PSD.

2. יהיו $\alpha, \beta \geq 0$ ממשיים ו- $A, B \in \mathbb{R}^{n\times n}$ מטריצות PSD. נראה כי $\alpha A + \beta B$ היא PSD.
ראשית, נראה כי היא סימטרית:

$$(\alpha A + \beta B)^T = \alpha A^T + \beta B^T = \alpha A + \beta B$$

כעת נבדוק כי היא PSD:
יהי $v \in \mathbb{R}^n$ כלשהו. מתקיים כי

$$v^T(\alpha A + \beta B)v = v^T(\alpha Av + \beta Bv) = \alpha v^T Av + \beta v^T Bv$$

ומכיוון ש - $\alpha, \beta \geq 0, v^T Av \geq 0, v^T Bv \geq 0$, נקבל כי

$$v^T(\alpha A + \beta B)v = \alpha v^T Av + \beta v^T Bv \geq 0$$

לכן $\alpha A + \beta B$ היא PSD.

לא מתקיים כי $M_{psd} \stackrel{\text{def}}{=} \{A \in \mathbb{R}^{n\times n} \mid A\ is\ PSD\}$ הוא מרחב וקטורי מעל $\mathbb{R}$.
נתבונן במטריצת היחידה I_n. מתקיים כי היא ניתנת לפירוק ע"י $I_n = I_n I_n^T$, לכן לפי הסעיף הראשון היא סימטרית והיא PSD, כלומר $I_n \in M_{psd}$.
לו M_{psd} היה מרחב וקטורי מעל $\mathbb{R}$ היה מתקיים כי $-I_n \in M_{psd}$, אבל לכל $v \in \mathbb{R}^n$ מתקיים כי $v^T(-I_n)v = -v^T v \leq 0$ ולכן I_n לא PSD.

:Calculus and Probability

1. בהינתן $X_1, \dots, X_n$ משתנים מקריים ב"ת המתפלגים $U[0,1]$ נמצא צפיפות , ונחשב את התוחלת והשונות של $Y = \max(X_1, \dots, X_n)$.

$$\begin{aligned}
F_Y(t) &= \mathbb{P}[Y \le t] = \mathbb{P}[\max(X_1, \dots, X_n) \le t] = \mathbb{P}[X_1 \le t, \dots, X_n \le t] \\
&= \prod_{i=1}^{n} \mathbb{P}[X_i \le t] = \mathbb{P}[U[0,1] \le t]^n = \left(\int_{-\infty}^{t} 1_{[0,1]} ds \right)^n \\
&= \begin{cases} 0 \ \ if\ t < 0\,, \\ t^n \ \ if\ t \in [0,1], \\ 1 \ \ if\ t > 1 \end{cases}
\end{aligned}$$

כאשר המעברים מוצדקים כי $X_1, \dots, X_n$ ב"ת ומתפלגים $U[0,1]$.
אם נגדיר $f_Y(t) = nt^{n-1}1_{[0,1]}$ נקבל כי $F_Y(t) = \int_{-\infty}^{t} f_Y(s)ds$. כלומר מצאנו צפיפות עבור Y.

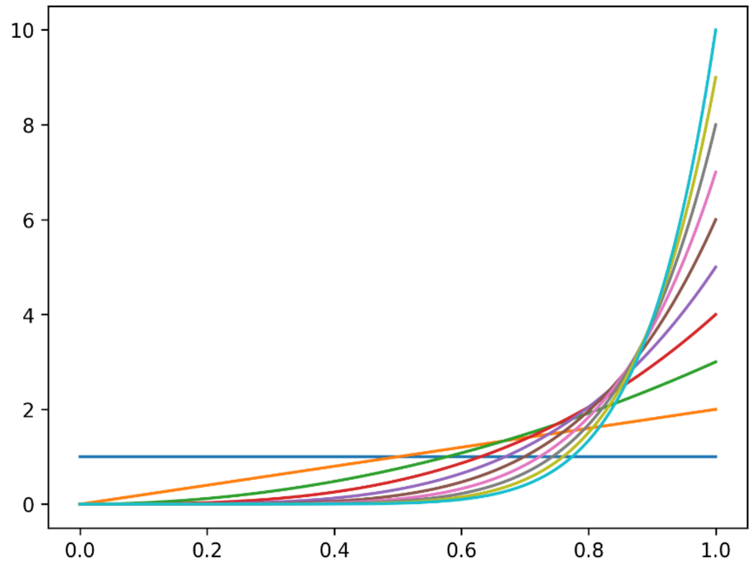
כעת נחשב את התוחלת והשונות:

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} t f_Y(t)dt = \int_0^1 tnt^{n-1}ds = \int_0^1 nt^n dt = \left[\frac{n}{n+1} t^{n+1} \right]_0^1 = 1 - \frac{1}{n+1}$$

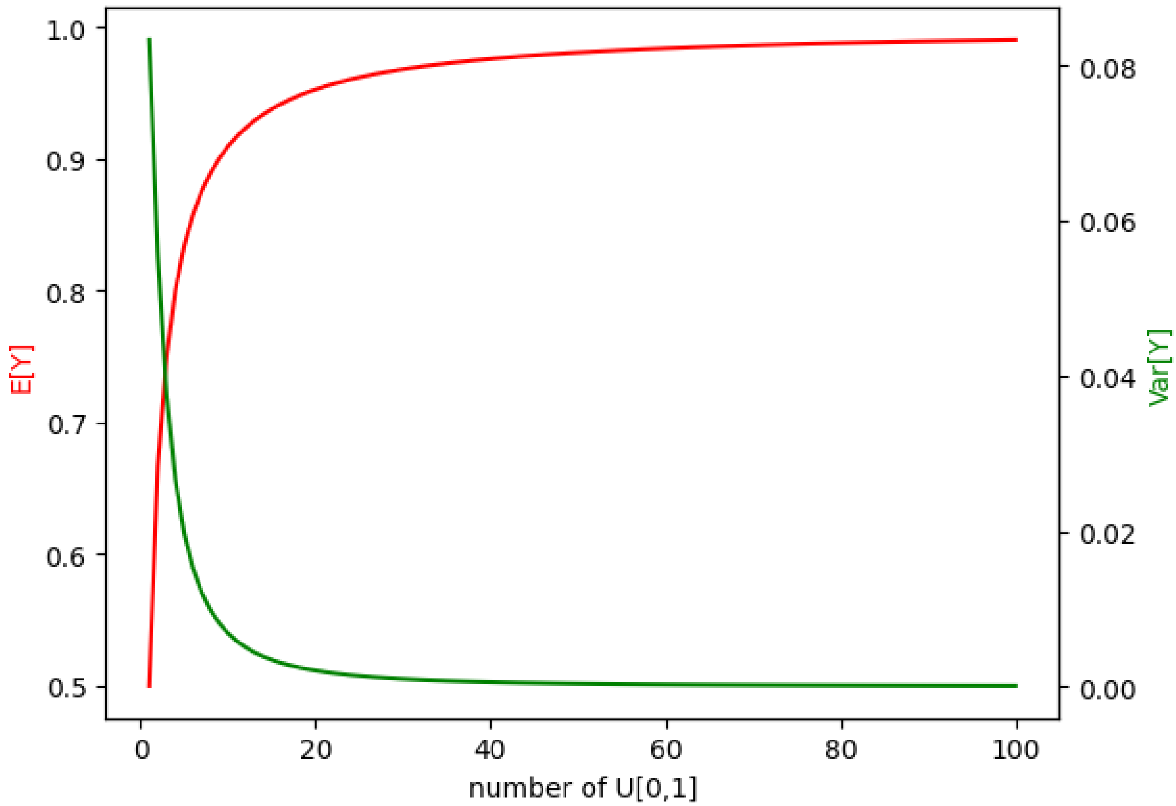
$$E[Y^2] = \int_{-\infty}^{\infty} t^2 f_Y(t)dt = \int_0^1 nt^{n+1}dt = \left[\frac{n}{n+2} t^{n+2} \right]_0^1 = 1 - \frac{2}{n+2}$$

$$\begin{aligned}
Var[Y] &= E[Y^2] - E[Y]^2 = 1 - \frac{2}{n+2} - \left(1 - \frac{1}{n+1}\right)^2 \\
&= 1 - \frac{2}{n+2} - \left(1 - \frac{2}{n+1} + \frac{1}{(n+1)^2}\right) = \frac{2}{n+1} - \frac{2}{n+2} - \frac{1}{(n+1)^2} \\
&= \frac{2(n+1)(n+2) - 2(n+1)^2 - n - 2}{(n+2)(n+1)^2} \\
&= \frac{2(n+1)(n+2-n-1) - n - 2}{(n+2)(n+1)^2} = \frac{2n+2-n-2}{(n+2)(n+1)^2} = \frac{n}{(n+2)(n+1)^2}
\end{aligned}$$

אפשר לראות כי ככל ש-n גדל, ה-PDF נהיה פולינום מדרגה גבוהה יותר.
שרטטתי גרף עבור n בין 1 ל-10, ואכן ניתן לראות זאת:



לגבי התוחלת ניתן לראות כי ככל ש-n גדל, כך היא שואפת ל-1, והשונות שואפת ל-0.
שרטטתי את התוחלת והשונות כפונקציה של n.



.2

יהי X משתנה מקרי רציף, כלומר F_X רציפה.
מכיוון ש-$\lim_{t\to\infty} F_X(t) = 1$ ו-$\lim_{t\to-\infty} F_X(t) = 0$ נובע ממשפט ערך הביניים כי קיים $m \in \mathbb{R}$ כך שמתקיים כי $\mathbb{P}[X \le m] = F_X(m) = \frac{1}{2}$.

בהנחה כי קיים m יחיד כזה, נוכיח כי $m = argmin_{a\in\mathbb{R}}\mathbb{E}[|X-a|]$.

יהי $a \in \mathbb{R}$.
נגדיר $Y = X - a$. מתקיים כי Y מ"מ (כי הרכבה של פונקציות מדידות היא מדידה). נחשב:

$$F_Y(t) = \mathbb{P}[Y \le t] = \mathbb{P}[X - a \le t] = \mathbb{P}[X \le a + t] = F_X(a + t)$$

וזו רציפה כהרכבה של פונקציות רציפות. לכן Y מ"מ רציף.

נחשב את התוחלת של $|Y|$:

$$\mathbb{E}[|Y|] = \mathbb{E}[Y^+ + Y^-] = \mathbb{E}[Y^+] + E[Y^-] = \int_0^\infty (1 - F_Y(t)\, dt + \int_0^\infty F_Y(-t)dt$$

$$= \int_0^\infty (1 - F_Y(t)\, dt + \int_{-\infty}^0 F_Y(t)\, dt$$

$$= \int_0^\infty (1 - F_X(t+a)\, dt + \int_{-\infty}^0 F_X(t+a)\, dt$$

ע"י החלפת משתנים $u = t + a$ נקבל

$$M(a) \stackrel{\text{def}}{=} \mathbb{E}[|X-a|] = \mathbb{E}[|Y|] = \int_a^\infty (1 - F_X(u)\, du + \int_{-\infty}^a F_X(u)\, dt$$

מכיוון שF_X רציפה נקבל מהמשפט היסודי של החשבון האינפיניטסימלי כי

$$M'(a) = -\big(1 - F_X(a)\big) + F_X(a) = 2F_X(a) - 1$$

אם נשווה את הנגזרת ל-0 נקבל כי קיימת נקודה קריטית כאשר $\mathbb{P}[X \le a] = F_X(a) = \frac{1}{2}$.
מתקיים כי זוהי נקודת מינימום כי:
- עבור $t \le a$ מתקיים כי $M'(t) \le 0$
- עבור $t \ge a$ מתקיים כי $M'(t) \ge 0$

(משום ש -F_X מונוטונית לא יורדת כפונקציית התפלגות צוברת).

לכן מתקיים כי החציון מתקבל ב

$$m = argmin_{a\in\mathbb{R}}\mathbb{E}[|Y|] = argmin_{a\in\mathbb{R}}\mathbb{E}[|X-a|]$$

: Optimal Classifiers and Decision Rule

.1

.a

תהי $f: \mathcal{X} \to \mathcal{Y}$ היפותזה כלשהי.

נחשב את תוחלת השגיאה עבור ℓ_{0-1}:

$$\mathbb{E}\left[\ell_{0-1}\left(Y, f(X)\right)\right] = \sum_{(x,y)\in(\mathcal{X}\times\mathcal{Y})} \mathbb{P}[X = x, Y = y]\ell_{0-1}\left(y, f(x)\right)$$

יהי $x \in \mathcal{X}$, מתקיים כי הגורמים היחידים שהוא מופיע בהם הם:

$$\left\{\mathbb{P}[X = x, Y = y]\ell_{0-1}\left(y, f(x)\right)\right\}_{y\in\mathcal{Y}}$$

נתבונן בשגיאה הנצברת שלהם:

$$\sum_{y\in\mathcal{Y}} \mathbb{P}[X = x, Y = y]\ell_{0-1}\left(y, f(x)\right) = \sum_{i=1}^{L} \mathbb{P}[X = x, Y = i]\ell_{0-1}\left(i, f(x)\right)$$

$$= \mathbb{P}[X = x] \sum_{i=1}^{L} \mathbb{P}[Y = i|X = x]\ell_{0-1}\left(i, f(x)\right)$$

אנו רוצים למזער את תוחלת השגיאה ככל הניתן ע"י קביעת $h(x)$.

נשים לב כי אם $h(x) = j, 1 \le j \le L$:

$$\mathbb{P}[X = x] \sum_{i=1}^{L} \mathbb{P}[Y = i|X = x]\ell_{0-1}\left(i, f(x)\right)$$

$$= \mathbb{P}[X = x] \sum_{i=1}^{L} \mathbb{P}[Y = i|X = x]\ell_{0-1}(i, j)$$

$$= \mathbb{P}[X = x] \sum_{i\in\mathcal{Y}/\{j\}} \mathbb{P}[Y = i|X = x]$$

לכן אם נבחר כי $h(x) \in argmax_{i\in\mathcal{Y}}\ \mathbb{P}[Y = i \mid X = x]$ נקבל כי הביטוי שלמעלה מזערי:

יהי $j \in \mathrm{argmax}_{\mathrm{i\in Y}}\, \mathbb{P}[\mathrm{Y} = \mathrm{i}|\mathrm{X} = \mathrm{x}]$, $i \in \mathcal{Y}/\{j\}$. מתקיים כי

$$\mathbb{P}[Y = i|X = x] \le \mathbb{P}[Y = j|X = x]$$

לכן

$$\mathbb{P}[X=x]\sum_{i\in\mathcal{Y}/\{i\}}\mathbb{P}[Y=i|X=x]$$
$$=\mathbb{P}[X=x]\left(\sum_{i\in\mathcal{Y}/\{i,j\}}\mathbb{P}[Y=i|X=x]+\mathbb{P}[Y=j|X=x]\right)$$
$$\geq\mathbb{P}[X=x]\left(\sum_{i\in\mathcal{Y}/\{i,j\}}\mathbb{P}[Y=i|X=x]+\mathbb{P}[Y=i|X=x]\right)$$
$$=\mathbb{P}[X=x]\sum_{i\in\mathcal{Y}/\{j\}}\mathbb{P}[Y=i|X=x]$$

כלומר הראנו כי עבור $x\in\mathcal{X}$ קביעת $\mathbb{P}[Y=i\,|\,X=x]$ $h(x)\in argmax_{i\in\mathcal{Y}}$ ממזערת את תוחלת השגיאה.

.b

תהי $f:\mathcal{X}\to\mathcal{Y}$ היפותזה כלשהי.

נחשב את תוחלת השגיאה עבור Δ:

$$\mathbb{E}\left[\ell_{0-1}\big(Y,f(X)\big)\right]=\sum_{(x,y)\in(\mathcal{X}\times\mathcal{Y})}\mathbb{P}[X=x,Y=y]\Delta\big(y,f(x)\big)$$

יהי $x\in\mathcal{X}$, מתקיים כי הגורמים היחידים שהוא מופיע בהם הם:

$$\left\{\mathbb{P}[X=x,Y=y]\Delta\big(y,f(x)\big)\right\}_{y\in\{0,1\}}$$

נתבונן בסכומים:

$$\sum_{y\in\mathcal{Y}}\mathbb{P}[X=x,Y=y]\ell_{0-1}\big(y,f(x)\big)=$$
$$\mathbb{P}[X=x](\mathbb{P}[Y=0|X=x]\Delta\big(0,f(x)\big)+\ \mathbb{P}[Y=1|X=x]\Delta\big(1,f(x)\big)$$

אנו רוצים למזער את תוחלת השגיאה ככל הניתן ע"י קביעת $h(x)$.

אם $h(x)=1$, נקבל כי השגיאה שתתרם היא

$$\mathbb{P}[Y=0|X=x]\Delta\big(0,f(x)\big)=\mathbb{P}[Y=0|X=x]\mathrm{a}$$

אם $h(x)=0$ נקבל כי השגיאה שתתרם היא

$$\mathbb{P}[Y=1|X=x]\Delta\big(1,f(x)\big)=\mathbb{P}[Y=1|X=x]\mathrm{b}$$

לכן אם $\{\mathbb{P}[Y=0|X=x]\mathrm{a},\mathbb{P}[Y=1|X=x]\mathrm{b}\}$ $h(x)\in argmax_{y\in\{0,1\}}$ נקבל כי תוחלת השגיאה תהיה מזערית.

2.

b.

נחשב כלל בחירה:

כפי שראינו בתרגול, מתקיים כי

$$\mathbb{P}[Y=1|\boldsymbol{X}=x] > \mathbb{P}[Y=0|\boldsymbol{X}=x] \Leftrightarrow$$

$$\mathbb{P}[Y=1]f_{\boldsymbol{X}|y}(x|1) > \mathbb{P}[Y=0]f_{\boldsymbol{X}|y}(x|0)$$

מהנתון $\boldsymbol{X}|Y=1\sim\mathcal{N}(\boldsymbol{\mu_1},\boldsymbol{\Sigma}),\boldsymbol{X}|Y=0\sim\mathcal{N}(\boldsymbol{\mu_0},\boldsymbol{\Sigma})$**.**

נציב ונקבל כי

$$\begin{aligned}
&\mathbb{P}[Y=1|\boldsymbol{X}=x] > \mathbb{P}[Y=0|\boldsymbol{X}=x]\\
&\Leftrightarrow \frac{e^{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}p > \frac{e^{-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)}}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}(1-p)\\
&\Leftrightarrow \frac{(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)-(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}{2} > \log\left(\frac{1-p}{p}\right)\\
&\Leftrightarrow \mu_0^T\Sigma^{-1}\mu_0 - 2\mu_0^T\Sigma^{-1}x + 2\mu_1^T\Sigma^{-1}x - \mu_1^T\Sigma^{-1}\mu_1 > 2\log\left(\frac{1-p}{p}\right)\\
&\Leftrightarrow 2(\mu_1-\mu_0)^T\Sigma^{-1}x > 2\log\left(\frac{1-p}{p}\right) + \mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0\\
&\Leftrightarrow 2\sum_{i=1}^{d}\left(\mu_{1_i}-\mu_{0_i}\right)\sigma_i^{-2}x_i > 2\log\left(\frac{1-p}{p}\right) + \sum_{i=0}^{d}\sigma_i^{-2}\left(\mu_{1_0}^{2}-\mu_{0_i}^{2}\right)
\end{aligned}$$

ואפשר לראות כי במקרה ש-$d=1$ נקבל את מה שראינו בתרגול.

c. עבור $d=1$ נקבל כי ה-*Decision boundary* הוא נקודה במישור חד מימדי
עבור $d=2$ נקבל כי ה-*Decision boundary* הוא קו ישר במישור דו מימדי
עבור d כלשהו נקבל מישור מישור $d-1$ מימדי.

d.

בלי הגבלת הכלליות נניח כי $\sigma_1 > \sigma_0$

$$\mathbb{P}[Y=1|\boldsymbol{X}=x] > \mathbb{P}[Y=0|\boldsymbol{X}=x]$$

$$\Leftrightarrow \mathbb{P}[Y=1]f_{\boldsymbol{X}|y}(x|1) > \mathbb{P}[Y=0]f_{\boldsymbol{X}|y}(x|0)$$

$$\Leftrightarrow \frac{e^{-\frac{(x-\mu)^2}{2\sigma_1^2}}}{(2\pi)^{\frac{1}{2}}\sigma_1}p > \frac{e^{-\frac{(x-\mu)^2}{2\sigma_0^2}}}{(2\pi)^{\frac{1}{2}}\sigma_0}(1-p)$$

$$\Leftrightarrow \frac{(x-\mu)^2}{2\sigma_0^2} - \frac{(x-\mu)^2}{2\sigma_1^2} > \log\left(\frac{(1-p)\sigma_1}{p\sigma_0}\right)$$

$$\Leftrightarrow (x-\mu)^2\left(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}\right) > \log\left(\frac{(1-p)\sigma_1}{p\sigma_0}\right)$$

$$\Leftrightarrow (x-\mu)^2 > \log\left(\frac{(1-p)\sigma_1}{p\sigma_0}\right)\frac{2\sigma_1^2\sigma_0^2}{\sigma_1^2 - \sigma_0^2}$$

$$\Leftrightarrow (x-\mu)^2 > \log\left(\frac{(1-p)\sigma_1}{p\sigma_0}\right)\frac{2\sigma_1^2\sigma_0^2}{(\sigma_1+\sigma_0)(\sigma_1-\sigma_0)}$$

$$\Leftrightarrow x > \sqrt{\log\left(\frac{(1-p)\sigma_1}{p\sigma_0}\right)\frac{2\sigma_1^2\sigma_0^2}{(\sigma_1+\sigma_0)(\sigma_1-\sigma_0)}} + \mu \; \boldsymbol{or}$$

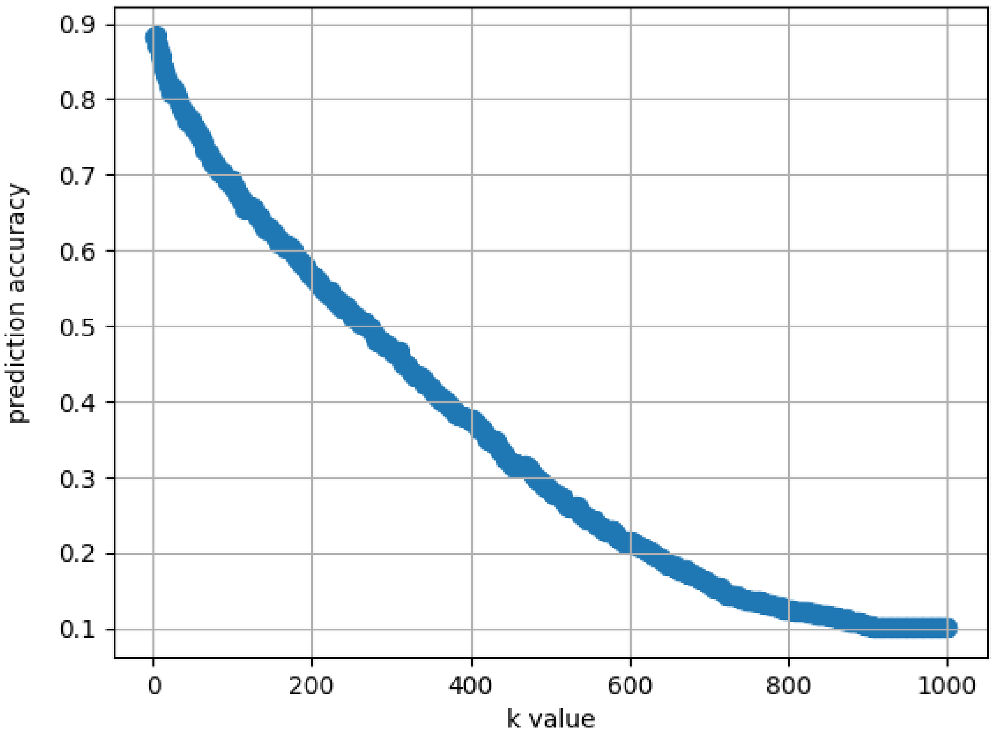
$$x < -\sqrt{\log\left(\frac{(1-p)\sigma_1}{p\sigma_0}\right)\frac{2\sigma_1^2\sigma_0^2}{(\sigma_1+\sigma_0)(\sigma_1-\sigma_0)}} + \mu$$

:Programming Assignment

b. אחוז הדיוק של האלגוריתם אותו מימשתי עבור 1000 דגימות כפי שהוגדר בתרגיל עמד על 85.8% (עבור $k = 10$).
עבור אלגוריתם שמקצה באופן רנדומלי הייתי מצפה לאחוז דיוק של 10%, מכיוון שגודל קבוצת התיוגים(Y) הוא 10.

c. ניתן לראות כי ככל ש-k גדל, כך הדיוק של $\mathrm{k\text{-}nn}$ יורד.
ערכי ה-k עבורם אחוז הדיוק היה הטוב ביותר הם $k = 1, 2, 3, 4$ כאשר אחוז הדיוק עמד על כ 88%.

כאשר ערך k גדול, האלגוריתם חשוף הרבה יותר ל"רעש" שיש בדגימות, לכן האפקטיביות שלו יורדת.



d. ניתן לראות כי אחוז הדיוק של האלגוריתם עולה משמעותית ככל שכמות דגימות האימון עולה.
לדעתי, ניתן לייחס את השיפור הזה לשני גורמים עיקריים:

1. ערך ה-k: בסעיף הקודם ראינו כי עבור ערכי k קטנים יעילות האלגוריתם טובה יותר בדרך כלל
2. איכות המדגם: ככל שכמות דגימות האימון עלתה, כך הן ייצגו את המציאות בצורה טובה יותר