פתרון שיעורי בית 1 –רוני קוסיצקי 205817893

שאלה 1

(→

נוכיח שאם המטריצה A היא סימטרית אז היא ניתנת לכתיבה כ $A = XX^T$ ולכן A היא PSD.
יהי $\mathrm{A} = XX^T$ עבור מטריצה כלשהיא X. יהיה v ווקטור כלשהוא. נוכל לכתוב את המשוואה הבאה:

$$v^T Av = v^T XX^T v = (v^T X) \cdot (\mathrm{X}^T v)$$

המכפלה של $v^T X$ היא ווקטור, נסמנו $u = (u_1 \dots, u_m)$. בסה"כ אנחנו נקבל:

$$v^T Av = uu^T = u_1^2 + \dots + u_n^2 \geq 0$$

על פי ההגדרה המטריצה הינה PSD.

(←

<u>טענה</u> – אם A היא PSD אז כל ע"ע של A הוא אי שלילי.

<u>הוכחה</u> – יהיה λ ע"ע ולכן מקיים את המשוואה $Av = \lambda v$. נכפיל את 2 האגפים בv^T משמאל ונקבל

$$v^T Av = v^T \lambda v = \lambda v^T v = (v_1^2 + \dots + v_n^2) \geq 0$$

המטריצה היא PSD והסכום של ערכי הווקטור הוא אי שלישי לכל ערכי הווקטורים. לכן הע"ע חייב להיות אי שלילי.

מהרמז אנחנו יכולים לרשום את המטריצה $A = QDQ^T$ לכסינה, כלומר ש Q היא הוקטורים העצמיים וD היא אלכסונים עם הע"ע על האלכסון. מהטענה שהוכחנו נוכל להסיק שניתן לכתוב את D בצורה הבאה:

$$D = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix}$$

מכיוון שהיא אלכסונית וכל ערכי האלכסון אי שליליים, נוכל להגדיר $D = C^2$ כך ש

$$C = \begin{pmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_n} \end{pmatrix}$$

המטריצה החדשה אלכסונית גם כן וגם סימטרית. לכן, מקיימת $D = CC^T$ ובסך הכל נוכל להגדיר

$$A = QDQ^T = QCC^T Q^T = XX^T \ where \ X = QC$$

שאלה 2

נפתח את שני האגפים במקביל עד שנגיע לשוויון באגפים:

$$\left(\frac{\partial x^T A x}{\partial x}\right)_i = \frac{\partial x^T A x}{\partial x_i} = \frac{\partial\left((x_1, \dots, x_n) A (x_1, \dots, x_n)^T\right)}{\partial x_i}$$

$$= \frac{\partial\left((x_1, \dots, x_n)\begin{pmatrix} a_{11}x_1 + \cdots + a_{1i}x_1 + \cdots a_{1n}x_1 \\ \vdots \\ a_{n1}x_n + \cdots + a_{ni}x_n + \cdots a_{nn}x_n \end{pmatrix}\right)}{\partial x_i}$$

$$= \frac{\partial(a_{11}x_1^2 + \cdots + a_{1i}x_1x_i + \cdots a_{1n}x_1x_n + \cdots a_{ni}x_nx_i + \cdots a_{nn}x_n^2)}{\partial x_i}$$

$$= a_{1i}x_i + a_{2i}x_2 + \cdots + a_{i1}x_1 + a_{i2}x_2 + \cdots 2a_{ii}x_i + a_{ii+1}x_{i+1} \dots$$

$$= \sum_{j=1}^{n} a_{ji}x_j + \sum_{j=1}^{n} a_{ij}x_j$$

נפתח כעת את הצד השני ונראה שוויון:

$$\left((A + A^T)x\right)_i = \left(A\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + A^T\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}\right)_i$$

$$= \begin{pmatrix} a_{11}x_1 + \cdots + a_{1i}x_i + \cdots + a_{1n}x_n \\ a_{21}x_1 + \cdots + a_{2i}x_i + \cdots + a_{2n}x_n \\ \vdots \\ a_{n1}x_1 + \cdots + a_{ni}x_i + \cdots + a_{nn}x_n \end{pmatrix}_i + \begin{pmatrix} a_{11}x_1 + \cdots + a_{i1}x_i + \cdots + a_{n1}x_n \\ a_{12}x_1 + \cdots + a_{i2}x_i + \cdots + a_{n2}x_n \\ \vdots \\ a_{1i}x_1 + a_{2i}x_2 + \cdots a_{ii}x_i + \cdots + a_{ni}x_n \\ \vdots \\ a_{n1}x_1 + \cdots + a_{ni}x_i + \cdots + a_{nn}x_n \end{pmatrix}_i$$

$$= a_{i1}x_1 + \cdots + a_{in}x_n + a_{1i}x_1 + \cdots + a_{ni}x_n = \sum_{j=1}^{n} \boldsymbol{a_{ji}x_j} + \sum_{k=1}^{n} \boldsymbol{a_{ik}x_k}$$

הוקטורים בשני האגפים שווים כי כל איבר i שווה לאיבר i בצד השני.

שאלה 3

א – ניעזר בחוקי גזירה בסיסיים

$$R(\lambda) = \ln\left(E\left[e^{\lambda x}\right]\right)$$

$$R'(\lambda) = \frac{1}{E[e^{\lambda x}]} \cdot \frac{d}{dX}\left(E\left[e^{\lambda x}\right]\right) = \frac{1}{E[e^{\lambda x}]} \cdot \frac{d}{d\lambda} \sum_{x=-\infty}^{\infty} xe^{\lambda x} = \frac{1}{E[e^{\lambda x}]} \sum_{\lambda=-\infty}^{\infty} x^2 e^{\lambda x} = \frac{E\left[Xe^{\lambda x}\right]}{E\left[e^{\lambda x}\right]}$$

$$R''(\lambda) = \frac{d}{d\lambda}\left(\frac{E\left[Xe^{\lambda x}\right]}{E\left[e^{\lambda x}\right]}\right) = \frac{\frac{d}{d\lambda}\left(E\left[Xe^{\lambda x}\right] \cdot E\left[Xe^{\lambda x}\right] - \frac{d}{d\lambda}E\left[e^{\lambda x}\right]\right) \cdot E\left[Xe^{\lambda x}\right]}{E^2\left[e^{\lambda x}\right]}$$
$$= \frac{\frac{d}{d\lambda}\left(\sum_{x=-\infty}^{\infty} x \cdot xe^{\lambda x} \cdot E\left[Xe^{\lambda x}\right] - E\left[Xe^{\lambda x}\right]\right)^2}{E^2\left[e^{\lambda x}\right]} = \frac{E\left[X^2 e^{\lambda x}\right]E\left[e^{\lambda x}\right] - E^2\left[Xe^{\lambda x}\right]}{E^2\left[e^{\lambda x}\right]}$$
$$= \frac{E\left[X^2 e^{\lambda x}\right]}{E\left[e^{\lambda x}\right]} - \frac{E^2\left[Xe^{\lambda x}\right]}{E^2\left[e^{\lambda x}\right]}$$

ב – נשתמש ישירות בנוסחת השונות

$$Var(X) = E[X^2] - E[X]$$

$$E_g[X] = \int_{-\infty}^{\infty} x \frac{e^{\lambda x}}{E_f[e^{\lambda x}]} f(x)dx = \frac{1}{E_f[e^{\lambda x}]} \int_{-\infty}^{\infty} xe^{\lambda x} f(x)dx = \frac{E_f[Xe^{\lambda x}]}{E_f[e^{\lambda x}]}$$

$$E_g[X^2] = \int_{-\infty}^{\infty} x^2 \frac{e^{\lambda x}}{E_f[e^{\lambda x}]} f(x)dx = \frac{1}{E_f[e^{\lambda x}]} \int_{-\infty}^{\infty} x^2 e^{\lambda x} f(x)dx = \frac{E_f[X^2 e^{\lambda x}]}{E_f[e^{\lambda x}]}$$

ולכן נקבל שהשונות שווה בדיוק למה שמצאנו בסעיף א'.

ג – מתקיים:

$$|X| \le B, Var_g(X) \le B^2$$

ולכן נוכל להסיק ש

$$R''(\lambda) \le B^2$$

$$R'(\lambda) \le \int B^2 d\lambda = \lambda B^2, \quad R(\lambda) \le \int \lambda B^2 = \frac{1}{2}\lambda^2 B^2$$

$$\ln\left(E\left[e^{\lambda x}\right]\right) \le \frac{1}{2}\lambda^2 B^2$$

$$E\left[e^{\lambda x}\right] \le e^{\frac{1}{2}\lambda^2 B^2}$$

שאלה 4

נפתח את הביטוי $\mathbb{E}[\ell_{0-1}(f(X),Y)]$ עפ"י הגדרה ונמצא את המינימום שלו:

$$\mathbb{E}[\ell_{0-1}(f(X),Y)] = \sum_{x,y} \mathbb{P}(X=x,Y=y)\cdot \ell_{0-1}(f(x),y)$$

כידוע, אנחנו יוצאים מנקודת הנחה שההתפלגות שלנו איננה ידועה, כלומר איננו יודעים מהו $\mathbb{P}(X=x,Y=y)$.
אבל, אנחנו נרצה שפונקציית ההתפלגות תיהיה כזו שעבור כל צמד $\{(x,y)|x\in instance\ space, y\in label\ space\}$ היא תמקסם את ההסתברות שפונקציית ההפסד תיהיה שווה ל0(כלומר אין הפסד). לכן, נרצה למקסם את ההסתברות שיתקיים $f(x)=y$. ולכן עבור כל ערך של x הפרדיקטור האופטימלי $h(x)$ יהיה בדיוק מה שאנחנו מחפשים, כלומר –

$$h(x) = \arg\max_{i\in\mathcal{Y}} \mathbb{P}[Y=i|X=x]$$

שאלה 5

סעיף ב -

הפרדיקטור שלנו יתייג $Y=1$ כאשר $\mathbb{P}(Y=1|X=x) > P(Y=0|X=x)$. נפעיך את חוק בייס ונקבל

$$\frac{f_{X|Y}(x|1)\cdot\mathbb{P}(Y=1)}{f_X(x)} > \frac{f_{X|Y}(x|0)\cdot\mathbb{P}(Y=0)}{f_X(x)}$$

$$f_1(x)\cdot p > f_0(x)\cdot(1-p)$$

$$f(x;\mu_1,\Sigma)\cdot p > f(x;\mu_0,\Sigma)\cdot(1-p)$$

$$\frac{p}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)} > \frac{1-p}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)}$$

$$pe^{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)} > (1-p)e^{-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)}$$

$$\frac{p}{1-p} > e^{-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)+\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}$$

$$\begin{aligned}2\ln\left(\frac{p}{1-p}\right) &> -(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) + (x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\\ &= -(x^T\Sigma^{-1}x + \mu_0^T\Sigma^{-1}\mu_0 - \mu_0^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_0) + x^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1 - \mu_1^T\Sigma^{-1}x\\ &\quad - x^T\Sigma^{-1}\mu_1 = \mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0 + (\mu_0^T-\mu_1^T)\Sigma^{-1}x + x^T\Sigma^{-1}(\mu_0-\mu_1)\end{aligned}$$

נתון לנו שΣ אלכסונית. לכן, נוכל לסמן $\mu=\mu_0-\mu_1$ ולקבל:

$$x^T\Sigma^{-1}\mu = (x_1,\dots,x_n)\Sigma^{-1}\begin{pmatrix}\mu_1\\ \vdots\\ \mu_n\end{pmatrix} = (x_1,\dots,x_n)\begin{pmatrix}\Sigma_{11}^{-1}\mu_1\\ \vdots\\ \Sigma_{nn}^{-1}\mu_n\end{pmatrix} = x_1\mu_1\Sigma_{11}^{-1} + \dots + x_n\mu_n\Sigma_{nn}^{-1}$$

$$\mu^T\Sigma^{-1}x = (\mu_1,\dots,\mu_n)\Sigma^{-1}\begin{pmatrix}x_1\\ \vdots\\ x_n\end{pmatrix} = (\mu_1,\dots,\mu_n)\begin{pmatrix}\Sigma_{11}^{-1}x_1\\ \vdots\\ \Sigma_{nn}^{-1}x_n\end{pmatrix} = x_1\mu_1\Sigma_{11}^{-1} + \dots + x_n\mu_n\Sigma_{nn}^{-1}$$

זה אומר ש $(\mu_0^T - \mu_1^T)\Sigma^{-1}x = x^T\Sigma^{-1}(\mu_0 - \mu_1)$ ולכן

$$2\ln\left(\frac{p}{1-p}\right) + \mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1 > 2(\mu_1 - \mu_0)^T\Sigma^{-1}x$$

$$\ln\left(\frac{p}{1-p}\right) + \frac{\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1}{2} > (\mu_0 - \mu_1)^T\Sigma^{-1}x$$

נגדיר משתנה v להיות $v = (\mu_0 - \mu_1)^T\Sigma^{-1}$ ובסה"כ נקבל

$$\ln\left(\frac{p}{1-p}\right) + \frac{\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1}{2} > v \cdot x$$

וזה התנאי הפשוט יותר אותו נדרוש על x.

סעיף ג –

עבור $d = 1$, משפחת הפרדיקטורים הכללית תיהיה קו ליניארי $x = c$. עבור $d = 2$ הגבול יהיה קו שיהיה תלוי ב2 משתנים – בx, y. עבור $d = 3$ הגבול יהיה מישור. עבור $d = 4$ הגבול יהיה צורה תלת מימדית. נמשיך בצורה הזאת עד גם במימדים הגבוהים יותר.

סעיף ד –

כפי שראינו בסעיף ב', נתשמש בכלל החלטה $\mathbb{P}(Y = 1|X = x) > P(Y = 0|X = x)$ ונפעיל שוב את כלל בייס:

$$\frac{f_{X|Y}(x|1) \cdot \mathbb{P}(Y = 1)}{f_X(x)} > \frac{f_{X|Y}(x|0) \cdot \mathbb{P}(Y = 0)}{f_X(x)}$$

$$f(x; \mu, \sigma_1^2) \cdot p > f(x; \mu, \sigma_0^2) \cdot (1 - p)$$

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

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

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

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

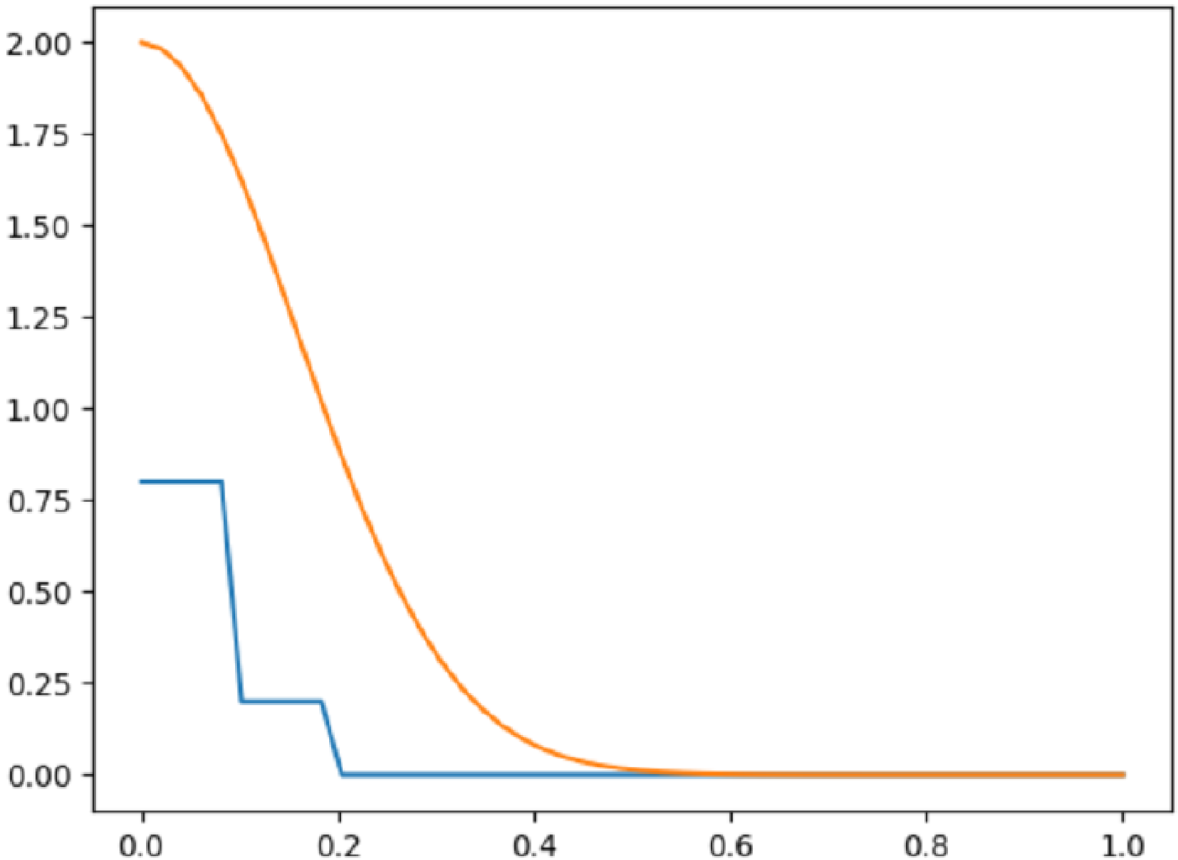
$$-\sqrt{\frac{2\ln\left(\frac{\sigma_0 p}{\sigma_1(1-p)}\right)}{\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)}} < x - \mu < \sqrt{\frac{2\ln\left(\frac{\sigma_0 p}{\sigma_1(1-p)}\right)}{\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)}}$$

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

קיבלנו 2 ספים שנחזה ש$y = 1$ אם x נמצא בתוך התחום הנ"ל.

שאלה 6

הקוד מצורף לתרגיל. הגרף שלנו יהיה:



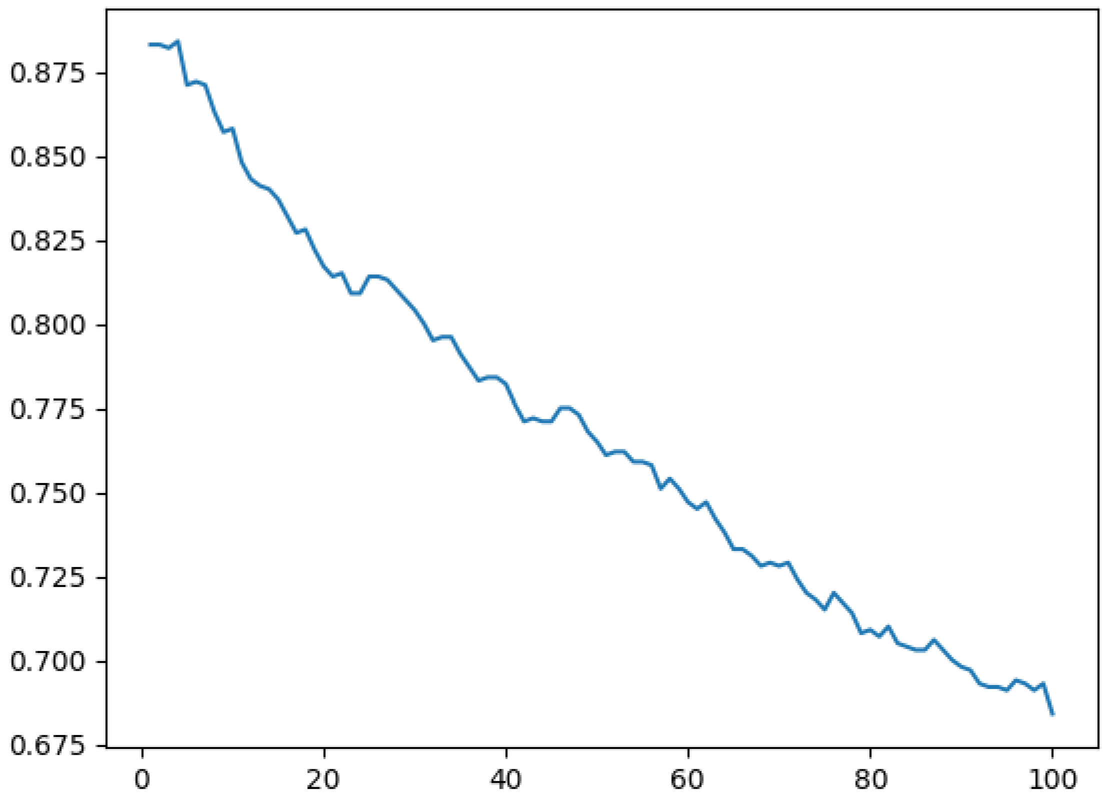
חלק תכנותי

סעיף ב -

עבור הערכים הנתונים אנו מקבלים אחוז דיוק של כ85 אחוז. במידה והיינו מנסים לנחש ספרה אחת מבין עשר הספרות אזי הסיכוי לתשובה נכונה יהיה רק 10 אחוז מכיוון שכל הספרות מפולגות אחיד. לכן, התוצאה שקיבלנו טובה משמעותית.

סעיף ג –

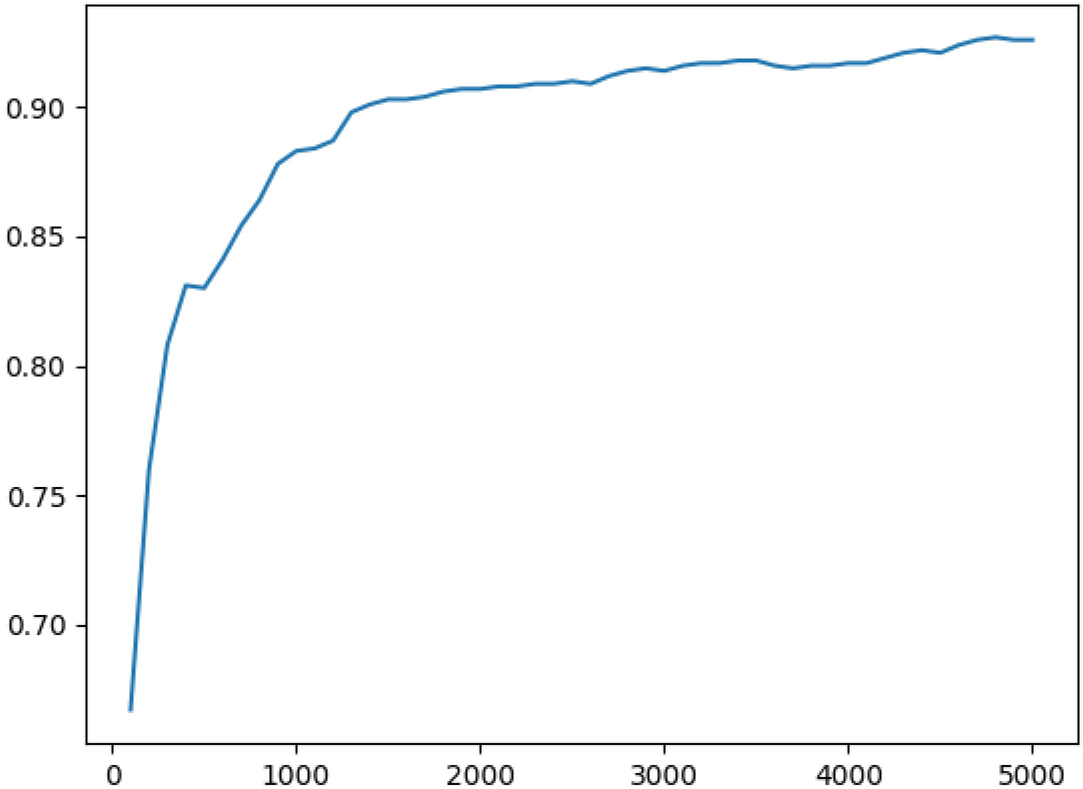
הגרף שייתקבל יהיה:



ככל שK יורד אזי הדיוק יורד. לכן, נקבל שהדיוק הטוב ביותר יהיה עבור K המינימלי, כלומר $k = 1$. ככל שנכניס יותר דגימות לבדיקת התווית של תמונה מסויימת, אנחנו נכניס גם בצורה רנדומלית תמונות שמרחקן משמעותית גדול יותר.
באופן ישיר אנחנו תמונות עם ערכים שונים לחלוטין ובכך נגדיר תווית שאיננה נכונה. לכן, ככל שנכניס יותר ערכים לדגימה בפועל נוריד את הדיוק שלנו ולכן הדיוק האופטימלי ייתקבל רק עבור התמונה שמרחקה האוקלידי לתמונה הרצויה היא הקרובה ביותר.

סעיף ד –

הגרף שייתקבל יהיה:



קל לראות שכל שסט האימון שלנו יותר גדול אז אנחנו מקבלים אלגוריתם חיזוי מדוייק יותר. הסיבה לכך היא שכאשר $k = 1$ אנחנו מתייחסים רק לתמונה הקרובה ביותר אלינו. כאשר אנחנו מוסיפים יותר תמונות למדגם שלנו ההסתברות לקבל תמונה קרובה אלינו גדל, ואז באופן ישיר גם ההסתברות לכך תגדל גם כן.