תרגיל בית 1 – מבוא ללמידה חישובית

מגיש: נתן בלוך

.Linear Algebra

שאלה 1.

נדרש להוכיח את הטענה הבאה:

$$A \text{ symmetric matrix } A \text{ is PSD } \boldsymbol{if\ and\ only\ if} \text{ it can be written as } A = X\,X^T$$

ואכן, **בכיוון הראשון**, $\Leftarrow$, נניח כי $A = X\,X^T$. אכן A הינה סימטרית, שכן מתקיים –

$$A^T = (X\,X^T)^T = (X^T)^T\,X^T = X\,X^T = A$$

כמו כן, נרצה להראות שלכל וקטור v מתקיים $v^TAv \geq 0$. יהי וקטור v. נסמן ראשית -

$$y := X^T \cdot v \in \mathbb{R}^{n\times 1}$$

ובעת, מתקיים –

$$v^TAv = v^TXX^Tv = (X^Tv)^T(X^Tv) = y^Ty \geq 0$$

כאשר, האי-שיוויון האחרון מתקיים לכל וקטור y, שכן בעצם זה נובע מכך שאם נסמן $y = (y_1, \dots, y_n)$, נקבל

$$y^Ty = \sum_{i=1}^{n} y_i \cdot y_i = \sum_{i=1}^{n} {y_i}^2 \geq 0$$

כנדרש בעצם, שכן הראנו כי לכל וקטור v מתקיים $v^TAv \geq 0$, ולכן A הינה מטריצה סימטרית מוגדרת חיובית.

בכיוון השני, $\Rightarrow$, נניח כי המטריצה A הינה מטריצה סימטרית המוגדרת חיובית. כיוון ש-A מוגדרת חיובית, ניתן להסיק כי כל הערכים העצמיים שלה הינם אי-שליליים, כלומר לכל ערך עצמי λ, מתקיים ש- $\lambda \geq 0$. ניתן לראות זאת בקלות שאם נסמן ב-u וקטור עצמי של A עם ערך עצמי λ, כלומר $Au = \lambda u$. מכאן נקבל כי –

$$u^TAu = u^T\lambda u = \lambda u^Tu$$

מצד שני, מההנחה כי A הינה מוגדרת חיובית, לכל וקטור v מתקיים - $v^TAv \geq 0$.
על כן, נסיק כי בפרט עבור הוקטור u מתקיים ש- $\lambda u^Tu \geq 0$, וכיוון ש- $u^Tu \geq 0$, מנימוקים קודמים, נסיק כי מתקיים גם כן $\lambda \geq 0$. על כן, הנ"ל מוכיח כי כל הערכים העצמיים של A הינם אי-שליליים.
כעת, על ידי שימוש ברמז, וכן שימוש בפירוק($decomposition$) של מטריצות סימטריות מוגדרות חיובית, ניתן לכתוב את המטריצה A באופן הבא - $A = Q^TDQ$, כאשר Q מטריצה אורתוגונלית, וכן D מטריצה אלכסונית, כך שבאלכסון מופיעים כל הערכים העצמיים של A. לכן, ממה שהוכחנו עד כה, מתקיים ש- $D_{ii} \geq 0$ לכל i. על כן, ניתן להגדיר את המטריצה האלכסונית B באופן הבא - $B_{ii} = \sqrt{D_{ii}}$, וכן לכל $i \neq j$, מתקיים $B_{ij} = 0$ וכך נקבל שמתקיים - $D = B^2$. כמו כן, B אלכסונית, ולכן סימטרית, ומתקיים $B = B^T$.
כעת נסתכל השיוויון –

$$A = Q^TDQ \underbrace{=}_{D=B^2} Q^TB^2Q \underbrace{=}_{B=B^T} Q^TB^TBQ = (BQ)^TBQ$$

על כן, בחירה של המטריצה X, באופן הבא - $X = (BQ)^T$, תשיג את הנדרש, שכן יתקיים –

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

כנדרש בכיוון זה של הטענה.
בסה"כ, הוכחנו את שני כיווני הטענה כנדרש.

שאלה 2.

יהיו A, B מטריצות ממשיות סימטריות המוגדרות חיוביות, וכן יהי סקלר $0 \leq \theta \leq 1$. נדרש להראות כי המטריצה $\theta A + (1-\theta)B$ הינה מטריצה ממשית סימטרית המוגדרת חיובית. ואכן, מההנחות מתקיים כי –

$$A = A^T \,, B \;=\; B^T$$

ולכן אם נסמן את המטריצה $C = \theta A + (1-\theta)B$, נקבל כי מתקיים –

$$C^T = (\theta A + (1-\theta)B)^T = (\theta A)^T + \left((1-\theta)B\right)^T = \theta A^T + (1-\theta)B^T = \theta A + (1-\theta)B = \; C$$

כאשר הנ"ל מוכיח את הסימטריות. כמו כן, נדרש להראות כי המטריצה הנ"ל הינה מוגדרת חיובית.

מהנחה, מתקיים **לכל** וקטור v -

$$v^T A v \geq 0 \quad , v^T B v \geq \; 0$$

כמו כן, מכיוון ש-$0 \leq \theta \leq 1$, מתקיים כי $0 \leq \theta, 1 - \theta$ ולכן ניתן לקבל מכאן –

$$v^T C v \;=\; v^T(\theta A + (1-\theta)B)v \;=\; \underbrace{\theta}_{\geq 0} \cdot \underbrace{v^T A v}_{\substack{\geq 0 \\ \forall v}} \;+\; \underbrace{(1-\theta)}_{\geq 0} \cdot \underbrace{v^T B v}_{\substack{\geq 0 \\ \forall v}} \;\; \geq \;\; \mathbf{0}$$

ומכאן, ניתן להסיק כי המטריצה $\boldsymbol{\theta A + (1-\theta)B}$ הינה מוגדרת חיובית וכן סימטרית, כנדרש.

.Calculus and Probability

שאלה 1.

נגדיר את הנגזרת של סקלר y ביחס לוקטור x כוקטור עמודה המוגדר באופן הבא –

$$\left(\frac{\partial y}{\partial x}\right)_i = \frac{\partial y}{\partial x_i}$$

לכל $i = 1, \dots, n$. תהי A מטריצה מסדר $n \times n$. נדרש להוכיח כי מתקיים –

$$\frac{\partial x^T A x}{\partial x} = (A + A^T)x$$

ואכן, נסמן ראשית $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, וכן $A = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,n} \end{pmatrix}$. מכאן מתקיים ש-

$$x^T A x = (x_1, \dots, x_n) \cdot \begin{pmatrix} \sum_{j=1}^{n} a_{1,j} \cdot x_j \\ \vdots \\ \sum_{j=1}^{n} a_{n,j} \cdot x_j \end{pmatrix} = \sum_{i=1}^{n} x_i \cdot \left(\sum_{j=1}^{n} a_{i,j} \cdot x_j \right)$$

$$\Longrightarrow \quad x^T A x = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{i,j} x_i x_j$$

על מנת להוכיח את השיוויון הנדרש, נראה שמתקיים שיוויון עבור כל קורדיאנטה. יהי $k \in \{1, \dots, n\}$. נרצה להראות כי מתקיים –

$$\left(\frac{\partial x^T A x}{\partial x}\right)_k = \left((A + A^T)x\right)_k$$

ואכן מתקיים לפי הגדרה,

$$\left(\frac{\partial x^T A x}{\partial x}\right)_k = \frac{\partial x^T A x}{\partial x_k} = \frac{\partial\left(\sum_{i=1}^{n} \sum_{j=1}^{n} a_{i,j} x_i x_j\right)}{\partial x_k} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \partial\left(a_{i,j} x_i x_j\right)}{\partial x_k} =$$

$$= \underbrace{2 \cdot a_{k,k} x_k}_{\substack{case\ of \\ i=j=k}} + \underbrace{\sum_{k \neq j \geq 1}^{n} a_{k,j} x_j}_{\substack{case\ of \\ i=k, j \neq k}} + \underbrace{\sum_{k \neq i \geq 1}^{n} a_{i,k} x_i}_{\substack{case\ of \\ j=k, i \neq k}} = \sum_{j=1}^{n} a_{k,j} x_j + \sum_{i=1}^{n} a_{i,k} x_i =$$

$$= (Ax)_k + (A^T x)_k = \left((A + A^T)x\right)_k$$

כאשר, נבחין שעבור המקרים בהם $i \neq k, j \neq k$, בעת גזירה לפי המשתנה x_k, מדובר באיברים **קבועים** מהצורה $a_{i,j} x_i x_j$, ולכן נגזרתם לפי x_k הינה 0. **מכאן, ניתן להסיק כאמור כי מתקיים השיוויון הדרוש -**

$$\frac{\partial x^T A x}{\partial x} = (A + A^T)x$$

שכן הוכחנו כי השיוויון מתקיים לכל קואורדינטה, כנדרש.

שאלה 2.

נתון $\boldsymbol{p} = (p_1, \dots, p_n)$ התפלגות דיסקרטית, כאשר מתקיים $\sum_{i=1}^{n} p_i = 1$, וכן $p_i \geq 0$ לכל i. נתונה גם –

$$H(\boldsymbol{p}) = -\sum_{i=1}^{n} p_i \cdot log(p_i)$$

נראה באמצעות כופלי לגרנז' כי עבור ההתפלגות האחידה, מקבלים אנטרופיה מקסימלית. לשם כך, נגדיר את הפונקציה הבאה, לפי האילוץ הנתון –

$$g_1(\boldsymbol{p}) = \sum_{i=1}^{n} p_i - 1$$

וכעת נגדיר פונק' חדשה באופן הבא –

$$h(\boldsymbol{p}, \lambda_1) = H(\boldsymbol{p}) + \lambda_1 g_1(\boldsymbol{p}) = -\sum_{i=1}^{n} p_i \cdot log(p_i) + \lambda_1 \sum_{i=1}^{n} p_i - \lambda_1$$

כעת, נרצה לגזור לפי כל אחד מהמשתנים ולהראות כי המקסימום מתקבל עבור $p_1 = \cdots = p_n = \frac{1}{n}$.

נגזור ראשית לפי p_i, ונקבל –

$$\frac{\partial h(\boldsymbol{p}, \lambda_1)}{\partial p_i} = \frac{\partial\left(-\sum_{i=1}^{n} p_i \cdot log(p_i) + \lambda_1 \sum_{i=1}^{n} p_i - \lambda_1\right)}{\partial p_i} =$$

$$= -\left(p_i \cdot \frac{1}{p_i} + 1 \cdot log(p_i)\right) + \lambda_1 = \boldsymbol{-log(p_i) + \lambda_1 - 1}$$

כעת, נגזור לפי λ_1, ונקבל –

$$\frac{\partial h(\boldsymbol{p}, \lambda_1)}{\partial \lambda_1} = \frac{\partial\left(-\sum_{i=1}^{n} p_i \cdot log(p_i) + \lambda_1 \sum_{i=1}^{n} p_i - \lambda_1\right)}{\partial \lambda_1} =$$

$$= \sum_{i=1}^{n} p_i - 1$$

מכאן, נסיק כי מתקיים –

$$\nabla h(\boldsymbol{p}, \lambda_1) = \begin{pmatrix} \frac{\partial h(\boldsymbol{p}, \lambda_1)}{\partial p_1} \\ \vdots \\ \frac{\partial h(\boldsymbol{p}, \lambda_1)}{\partial p_n} \\ \frac{\partial h(\boldsymbol{p}, \lambda_1)}{\partial \lambda_1} \end{pmatrix}$$

ואם נרצה למצוא את המקסימום, נחפש את הערכים עבורם $\nabla h(\boldsymbol{p}, \lambda_1) = 0$, ונקבל כי מתקיים –

$$\nabla h(\boldsymbol{p},\lambda_1) = \begin{pmatrix} \frac{\partial h(\boldsymbol{p},\lambda_1)}{\partial p_1} \\ \vdots \\ \frac{\partial h(\boldsymbol{p},\lambda_1)}{\partial p_n} \\ \frac{\partial h(\boldsymbol{p},\lambda_1)}{\partial \lambda_1} \end{pmatrix} = \begin{pmatrix} -log(p_1) + \lambda_1 - 1 \\ \vdots \\ -log(p_n) + \lambda_1 - 1 \\ \sum_{i=1}^{n} p_i - 1 \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\Longrightarrow \ log(p_i) = \lambda_1 - 1 \quad \forall i \in \{1, \dots n\} \ \Longrightarrow \ log(p_1) \ = \dots = log(p_n) \ \Longrightarrow \ p_1 = p_2 = \cdots = p_n$$

ומכאן, נשתמש באילוץ האחרון, כדי לקבל את הרצוי –

$$\sum_{i=1}^{n} p_i - 1 = 0 \quad \Longrightarrow \quad \sum_{i=1}^{n} p_i = 1 \quad \Longrightarrow \quad n \cdot p_i \ = \ 1 \quad \Longrightarrow \quad \boldsymbol{p_i} \ = \ \frac{\boldsymbol{1}}{\boldsymbol{n}}$$

ומכאן קיבלנו כי $\boldsymbol{p} = (\boldsymbol{p_1}, \dots, \boldsymbol{p_n}) = \ (\frac{\boldsymbol{1}}{\boldsymbol{n}}, \dots, \frac{\boldsymbol{1}}{\boldsymbol{n}})$, כלומר $\boldsymbol{p}$ הינה ההתפלגות אחידה.

שאלה 3.

נתונים n משתנים מקריים חיוביים, $X_0, \dots, X_{n-1}$, שהינם בלתי תלויים, ומתפלגים זהה(כלומר $i.i.d$). כמו כן, נתונה פונקציית התפלגות רציפה f_X.

סעיף א.

נדרש להוכיח כי מתקיים השיוויון הבא –

$$\mathbb{P}\left(X_0 \geq max(X_1, X_2, \dots, X_n)\right) = \int_0^\infty \left(F_{X_0}(a)\right)^{n-1} f_{X_0}(a)da$$

ואכן, על מנת להוכיח זאת, נשתמש בנוסחת ההסתברות השלמה, וכן נסתכל על $max(X_1, X_2, \dots, X_n)$ כמשתנה מקרי. נקבל כי –

$$\mathbb{P}\left(X_0 \geq max(X_1, X_2, \dots, X_n)\right) = \int_0^\infty \mathbb{P}(max(X_1, X_2, \dots, X_n) \leq a) \cdot f_{X_0}(a)da$$

כמו כן, מתקיים כי –

$$\mathbb{P}(max(X_1, X_2, \dots, X_n) \leq a) = \mathbb{P}(X_1 \leq a, X_2 \leq a, \dots, X_{n-1} \leq a) =$$
$$= \mathbb{P}(X_1 \leq a) \cdot \mathbb{P}(X_2 \leq a) \cdots \mathbb{P}(X_{n-1} \leq a) = \left(\mathbb{P}(X_0 \leq a)\right)^{n-1} = \left(F_{X_0}(a)\right)^{n-1}$$

כך שמתקיים בסה"כ השיוויון הדרוש –

$$\mathbb{P}\left(X_0 \geq max(X_1, X_2, \dots, X_n)\right) = \int_0^\infty \left(F_{X_0}(a)\right)^{n-1} f_{X_0}(a)da$$

סעיף ב.

ניזכר בכך שמתקיים –

$$F_{X_0}(a) = \int_{-\infty}^a f_{X_0}(t)dt \underbrace{=}_{X_i \geq 0} \int_0^a f_{X_0}(t)dt$$

כך שלמעשה, F_{X_0} הינה הפונק' הקדומה של f_{X_0}. נסתכל על הפונקציה $\left(F_{X_0}(a)\right)^n$ ונגזור אותה כך שנקבל –

$$\left(\left(F_{X_0}(a)\right)^n\right)' = n \cdot \left(F_{X_0}(a)\right)^{n-1} \cdot f_{X_0}(a)$$

כעת, נשתמש בכך על מנת לפתור את האינטגרל הנתון, ונקבל –

$$\mathbb{P}\left(X_0 \geq max(X_1, X_2, \dots, X_n)\right) = \int_0^\infty \left(F_{X_0}(a)\right)^{n-1} f_{X_0}(a)da = \frac{1}{n} \cdot \int_0^\infty n\left(F_{X_0}(a)\right)^{n-1} f_{X_0}(a)da$$

$$= \frac{1}{n} \cdot \int_0^\infty \left(\left(F_{X_0}(a)\right)^n\right)' da = \frac{1}{n} \cdot \left(\underbrace{\lim_{x\to\infty} F_{X_0}(x)}_{\substack{=1 \\ as\ a\ CDF}} - \underbrace{F_{X_0}(0)}_{\substack{=0 \\ as\ X_i \geq 0}}\right) = \frac{1}{n} \cdot (1 - 0) = \frac{1}{n}$$

כנדרש.

Decision Rules and Concentration Bounds.

שאלה 1.

יהיו X ו-Y משתנים מקריים, כך ש- $Y \in \{1,2...,L\}$. תהי ℓ_{0-1} פונקציית ההפסד 0-1 שהוגדרה בביתה.
נרצה להראות ש- $h = arg \min_{f:\chi\to\mathbb{R}} \mathbb{E}[\ell_{0-1}(f(X),Y)]$ נתונה על ידי –

$$h(x) = arg \max_{i\in\{1,2...,L\}} \mathbb{P}[Y = i|\ X = x]$$

פיתרון. נניח כי $X = x_0$ ונבצע את החישוב תחת התניה, על מנת להוכיח כי מתקיים –

$$h(x_0) = arg \max_{i\in\{1,2...,L\}} \mathbb{P}[Y = i|X = x_0]$$

נסתכל על פונק' ההפסד הצפוי כפי שהוגדרה בביתה –

$$L(h) = \mathbb{E}[\ell_{0-1}(h(X),Y)] = \sum_{\substack{x\in\chi \\ y\in\mathcal{Y}}} \mathbb{P}[X = x, Y = y] \cdot \ell_{0-1}(h(x),\ \ y)$$

וכאמור נרצה להביא למינימום את פונק' הזו. נרצה למצוא את הערך של $h(x_0)$ המביא למינימום את פונק' ההפסד הצפוי $L(h)$. ואכן, נמשיך בחישובים ונקבל –

$$L(h) = \underbrace{\sum_{\substack{x\in\ \chi\backslash\{x_0\} \\ y\in\mathcal{Y}}} \mathbb{P}[X = x, Y = y] \cdot \ell_{0-1}(h(x),\ \ y)}_{\substack{doesn't\ \ depend\ \ on\ \ x_0 \\ \Rightarrow\ doesn't\ \ depend\ \ on\ \ h(x_0)}} + \underbrace{\sum_{y\in\mathcal{Y}} \mathbb{P}[X = x_0, Y = y] \cdot \ell_{0-1}(h(x_0),\ \ y)}_{=:\ g_{x_0}(h)}$$

מכיוון שבעת הנחנו כי $X = x_0$, וכן אנו רוצים למצוא את הערך האופטימלי של $h(x_0)$ המביא למינימום את פונק' ההפסד הצפוי $L(h)$, ניתן להתייחס רק לגורמים התלויים ב-x_0 בפונק' ההפסד, ולכן נרצה להביא למינימום את הגורמים הללו. נסמן ב-$g_{x_0}(h)$, את הגורמים שתלויים ב-x_0 בפונק' ההפסד, וכעת נבחין בתוצאות הללו –

$$\boldsymbol{if\quad h(x_0) \notin\ \mathcal{Y}} = \{1,2,...,L\}\quad \Rightarrow\ \forall y \in \mathcal{Y}.\quad \ell_{0-1}(h(x_0),\ \ y) = 1$$

$$\Rightarrow g_{x_0}(h)\ = \sum_{y\in\mathcal{Y}} \mathbb{P}[X = x_0, Y = y] \cdot \underbrace{\ell_{0-1}(h(x_0),\ \ y)}_{=1} = \sum_{y\in\mathcal{Y}} \mathbb{P}[X = x_0, Y = y] = \mathbb{P}[X = x_0]$$

ובנוסף,

$$\boldsymbol{if\quad h(x_0) = i \in \mathcal{Y}} \Rightarrow\ \forall y \in \mathcal{Y}\backslash\{i\}.\quad \ell_{0-1}(h(x_0),\ \ y) = 1\ \ and\ \ \ell_{0-1}(h(x_0),\ \ i) =\ 0$$

$$\Rightarrow g_{x_0}(h) = \sum_{y\in\mathcal{Y}} \mathbb{P}[X = x_0, Y = y] \cdot \ell_{0-1}(h(x),\ \ y) = \mathbb{P}[X = x_0] - \mathbb{P}[X = x_0, Y = i]$$

לכן, בעצם קיבלנו כי הגורם הנ"ל הינו –

$$g_{x_0}(h) \;=\; \begin{cases} \mathbb{P}[X = x_0] & if \;\; h(x_0) \notin \mathcal{Y} \\ \mathbb{P}[X = x_0] - \mathbb{P}[X = x_0, Y = i] & o.w. \;\; h(x_0) = i \in \mathcal{Y} \end{cases}$$

מכאן, ברור כי הגורם הנ"ל מגיע למינימום עבור $h(x_0) \in \mathcal{Y}$, שכן מתקיים –

$$\mathbb{P}[X = x_0] - \mathbb{P}[X = x_0, Y = i] \quad \le \quad \mathbb{P}[X = x_0]$$

ובפרט, $g_{x_0}(h)$ משיגה את המינימום שלה במקרה בו $\mathbb{P}[X = x_0, Y = i]$ מקסימלי, ועל כן, נרצה לבחור את ה-*predictor* האופטימלי עבור x_0 באופן הבא –

$$\begin{aligned} h(x_0) &= arg \max_{i \in \{1,2\ldots,L\}} \mathbb{P}[X = x_0, Y = i] \;= \\ &= arg \max_{i \in \{1,2\ldots,L\}} \mathbb{P}[Y = \; i \,|X = x_0] \cdot \underbrace{\mathbb{P}[X = x_0]}_{\substack{const \,\ge\, 0 \\ for\; all\; i}} = \\ &= arg \max_{i \in \{1,2\ldots,L\}} \mathbb{P}[Y = \; i \,|X = x_0] \end{aligned}$$

כנדרש.

שאלה 2.

נתון $X=(X_1,\dots,X_n)^T$ וקטור של משתנים מקריים. לפי ההגדרה, נרצה לחזות את $y\in\mathcal{Y}=\{0,1\}$ באופן הבא –

$$y=\begin{cases}1 & if\ \ \mathbb{P}[y=1|\boldsymbol{X}]\ >\ \mathbb{P}[y=0|\boldsymbol{X}]\\ 0 & otherwise\end{cases}$$

כאשר הסימון של $\mathbb{P}[y=y\mid\boldsymbol{X}]$ הינו סימון מקוצר ל- $\mathbb{P}[y=y|\boldsymbol{X}=\boldsymbol{x}]$.
נתונים גם – $\mathbb{P}[y=1]=p$, ולכן ניתן להסיק כי $\mathbb{P}[y=0]=1-p$. כמו כן, בהנתן $y=i$, פונק' הצפיפות של הוקטור X הינה $f_i(X)=f(X,\mu_i,\Sigma)$, כלומר למעשה $X_{|y=i}\sim N(\mu_i,\Sigma)$.

סעיף א.

נתונה נקודה $x\in\mathbb{R}^d$ ונדרש לבצע $labeling$ של y לפיה. נדרש למצוא תנאי פשוט יותר על $\boldsymbol{X}$ כך שתתקיים שקילות ביניהם. נשתמש בכלל בייס עבור התפלגות רציפה –

$$\boldsymbol{Continuous\ Bayes'\ rule}:\ \ f_X(x\mid Y=y)=\frac{\mathbb{P}[Y=y\mid X=x]\cdot f_X(x)}{\mathbb{P}[Y=y]}$$

ובעת, מתקיים –

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

$$\Leftrightarrow\quad\frac{f_X(x|Y=1)\cdot\mathbb{P}[y=1]}{f_X(x)}>\frac{f_X(x|Y=0)\cdot\mathbb{P}[y=0]}{f_X(x)}\quad\Leftrightarrow$$

$$\Leftrightarrow f_X(x|Y=1)\cdot\mathbb{P}[y=1]>f_X(x|Y=0)\cdot\mathbb{P}[y=0]\Leftrightarrow$$

$$\Leftrightarrow\quad\frac{f_X(x|Y=1)}{f_X(x|Y=0)}\ >\ \frac{\mathbb{P}[y=0]}{\mathbb{P}[y=1]}\quad\Leftrightarrow$$

$$\Leftrightarrow\quad\frac{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)}{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)}>\frac{1-p}{p}\quad\Leftrightarrow$$

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

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

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

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

$$\Leftrightarrow \quad {\color{red}x^T\Sigma^{-1}x} - x^T\Sigma^{-1}\mu_0 - {\mu_0}^T\Sigma^{-1}x + {\mu_0}^T\Sigma^{-1}\mu_0 \qquad (continues)$$

$$-\ ({\color{red}x^T\Sigma^{-1}x} - x^T\Sigma^{-1}\mu_1 - {\mu_1}^T\Sigma^{-1}x + {\mu_1}^T\Sigma^{-1}\mu_1) \quad > 2\cdot ln\left(\frac{1-p}{p}\right) \quad \Leftrightarrow$$

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

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

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

כעת, נבחין כי מתקיים $Cov(X,Y) = Cov(Y,X)$ שכן מדובר ביחס סימטרי, ולכן לכל $i,j \in \{1,\dots n\}$, מתקיים כי - $\Sigma_{i,j} = Cov(X_i,X_j) = Cov(X_j,X_i) = \Sigma_{j,i}$, ולכן בעצם המטריצה Σ הינה סימטרית. מכאן, ניתן להסיק כי המטריצה Σ^{-1} הינה סימטרית גם כן. לכן $\Sigma^{-1} = (\Sigma^{-1})^T$. כעת, מתקיים –

$$x^T\Sigma^{-1}(\mu_1-\mu_0) = x^T(\Sigma^{-1})^T(\mu_1-\mu_0) = (\Sigma^{-1}x)^T(\mu_1-\mu_0) = ((\mu_1-\mu_0)^T\Sigma^{-1}x)^T$$

אך מכיוון ש- $(\mu_1-\mu_0)^T\Sigma^{-1}x \in \mathbb{R}$, למעשה, מתקיים כי –

$$((\mu_1-\mu_0)^T\Sigma^{-1}x)^T = (\mu_1-\mu_0)^T\Sigma^{-1}x$$

ולכן בעצם קיבלנו כי מתקיים –

$$x^T\Sigma^{-1}(\mu_1-\mu_0) = (\mu_1-\mu_0)^T\Sigma^{-1}x$$

מכאן, נמשיך את פישוט התנאי מהתנאי האחרון אליו הגענו –

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

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

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

וזהו אם כך תנאי פשוט יותר על $\boldsymbol{X}=\boldsymbol{x}$ על מנת לבצע $labeling$ של y לפיה. **בסה"כ קיבלנו את התנאי –**

$$y = 1 \quad if\ and\ only\ if \qquad (\mu_1-\mu_0)^T\Sigma^{-1}x \ > ln\left(\frac{1-p}{p}\right) + \frac{{\mu_1}^T\Sigma^{-1}\mu_1}{2} - \frac{{\mu_0}^T\Sigma^{-1}\mu_0}{2}$$

כנדרש בסעיף זה.

סעיף ב.

גבול ההחלטה של הבעיה הזו מוגדר על ידי קבוצות הנקודות ב-$\mathbb{R}^d$ עבורן מתקיים –

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

נבין את צורת ההחלטה עבור $d > 1$ כללי, אך ראשית עבור אינטואיציה, נסתכל על $d = 1,2$. לפי הסעיף הקודם, קיבלנו את התנאי השקול –

$$\mathbb{P}[y = 1|\boldsymbol{X} = \boldsymbol{x}] = \mathbb{P}[y = 0|\boldsymbol{X} = \boldsymbol{x}] \iff$$
$$\iff (\mu_1 - \mu_0)^T \Sigma^{-1} x = ln\left(\frac{1-p}{p}\right) + \frac{{\mu_1}^T \Sigma^{-1} \mu_1}{2} - \frac{{\mu_0}^T \Sigma^{-1} \mu_0}{2}$$

עבור $\boldsymbol{d = 1}$, מתקיים – $\Sigma = Cov(X_1, X_1) = Var(X_1) \quad \Longrightarrow \Sigma^{-1} = \frac{1}{Var(X_1)}$. כמו כן, μ_1, μ_0 הם למעשה סקלרים, שכן מדובר בוקטורים באורך 1, ומכאן נקבל –

$$x = \frac{Var(X_1)}{\mu_1 - \mu_0} \cdot \left(ln\left(\frac{1-p}{p}\right) + \frac{{\mu_1}^2}{2 \cdot Var(X_1)} - \frac{{\mu_0}^2}{2 \cdot Var(X_1)} \right) =$$
$$= \frac{Var(X_1)}{\mu_1 - \mu_0} \cdot ln\left(\frac{1-p}{p}\right) + \frac{1}{2}\left(\frac{{\mu_1}^2 - {\mu_0}^2}{\mu_1 - \mu_0}\right) = \frac{Var(X_1)}{\mu_1 - \mu_0} \cdot ln\left(\frac{1-p}{p}\right) + \frac{1}{2}(\mu_1 + \mu_0)$$

ובמקרה זה x הינה נקודה על הישר, ובעצם מתקיים x הינה ממימד 0 בעוד מימד המרחב הינו $d = 1$.

עבור $\boldsymbol{d = 2}$, וכן ע"י סימון $\theta_1 := Var(X_1)$, $\theta_2 := Var(X_2)$, ו-$c = Cov(X_1, X_2)$, מתקיים –

$$\Sigma = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_2, X_1) & Var(X_2) \end{pmatrix} = \begin{pmatrix} \sigma_1 & c \\ c & \sigma_2 \end{pmatrix} \Longrightarrow \Sigma^{-1} = \begin{pmatrix} \frac{\sigma_2}{\sigma_1\sigma_2 - c^2} & -\frac{c}{\sigma_1\sigma_2 - c^2} \\ -\frac{c}{\sigma_1\sigma_2 - c^2} & -\frac{\sigma_1}{\sigma_1\sigma_2 - c^2} \end{pmatrix}$$

$$= a\ 2 \times 2\ matrix\ of\ constants$$

כאשר זה נובע מכך ש- $Var(X_1), Var(X_2), Cov(X_1, X_2)$ הינם קבועים ביחס ל- $X = x$, ותלויים רק בהתפלגות הרקע. על כן, אם נסתכל על השיוויון של גבול ההחלטה –

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

ונזכור כי $x \in \mathbb{R}^2$, כלומר $x = (x_1, x_2)$, נקבל בעצם **משוואה לינארית בשני הנעלמים $\boldsymbol{x_1, x_2}$ בלבד**, שכן כפי שהוסבר שאר הגורמים בשיוויון זה הינם קבועים.
כלומר, התנאי מגדיר קו ישר במישור דו-מימדי, וכן הישר ממימד 1, ונמצא במרחב ממימד 2 – $\mathbb{R}^2$.

באופן כללי, עבור $d \in \mathbb{N}$, ניתן להכליל ולהסיק כי התנאי $\mathbb{P}[y = 1|\boldsymbol{X} = \boldsymbol{x}] = \mathbb{P}[y = 0|\boldsymbol{X} = \boldsymbol{x}]$ יוצר משוואה אחת עם d משתנים, מה שמגדיר תת-מרחב מגודל $d - 1$ בתוך המרחב $\mathbb{R}^d$ שמימדו הינו d.
במילים אחרות, התנאי מגדיר מושג בשם "על-מישור"($hyperplane$), שהוא למעשה תת-מרחב בגודל $n - 1$ בתוך מרחב מגודל n.

שאלה 3.

נתונים $X_1, \dots, X_n$ משתנים מקרים בלתי תלויים המתפלגים באופן זהה ואחיד בקטע $[-3,5]$. נגדיר את סכומם

$$S = X_1 + \cdots + X_n$$

נדרש למצוא באמצעות אי-שיוויון הופדינג $N \in \mathbb{N}$, כך שלכל $n \geq N$, מתקיים –

$$\mathbb{P}[S > n^2 + 0.2n] < 0.1$$

פיתרון. ראשית נבדוק את התנאים הדרושים כדי להשתמש בחסם הופדינג. אכן נתון כי $X_1, \dots, X_n$ בלתי תלויים. כמו כן, כיוון שהם מתפלגים אחיד על הקטע $[-3,5]$, ניתן להסיק כי –

$$\mathbb{P}[-3 \leq X_i \leq 5] = 1 \quad for\ every\ i$$

כאשר במקרה זה נשתמש ב- $a = -3, b = 5$.

כמו כן, מתקיים כי –

$$\mathbb{E}[X_i] \underbrace{=}_{\sim U(-3,5)} \frac{a+b}{2} = \frac{-3+5}{2} = 1$$

וכן, אם נסתכל על ממוצע המשתנים המקריים, ונסמנו $\bar{\boldsymbol{X}}$, נזכור כי מתקיים - $\bar{\boldsymbol{X}} = \frac{S}{n}$. ומכאן –

$$\mathbb{E}[\bar{\boldsymbol{X}}] = \mathbb{E}[\frac{X_1 + \cdots + X_n}{n}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n} 1 = \frac{1}{n} \cdot (n \cdot 1) = \boldsymbol{1}$$

ובעת, נחשב –

$$\mathbb{P}[S > n^2 + 0.2n] = \mathbb{P}\left[\frac{S}{n} > n + 0.2\right] = \mathbb{P}[\bar{\boldsymbol{X}} > n + 0.2] =$$

$$= \mathbb{P}[\bar{\boldsymbol{X}} - 1 > n - 0.8] = \mathbb{P}[\bar{\boldsymbol{X}} - \mathbb{E}[\bar{\boldsymbol{X}}] > n - 0.8] \leq$$

$$\leq \mathbb{P}[\bar{\boldsymbol{X}} - \mathbb{E}[\bar{\boldsymbol{X}}] \geq n - 0.8] \leq \mathbb{P}[|\bar{\boldsymbol{X}} - \mathbb{E}[\bar{\boldsymbol{X}}]| \geq n - 0.8] \leq 2e^{-\frac{2n(n-0.8)^2}{(b-a)^2}}$$

וכאמור, המטרה הינה למצוא $N \in \mathbb{N}$, כך שלכל $n \geq N$, מתקיים –

$$\mathbb{P}[S > n^2 + 0.2n] \leq 2e^{-\frac{2n(n-0.8)^2}{(b-a)^2}} < 0.1$$

על כן, נפתור את אי-השיוויון הימני על מנת להשיג את החסם הדרוש. מתקיים –

$$2e^{-\frac{2n(n-0.8)^2}{(b-a)^2}} < 0.1 \Leftrightarrow e^{-\frac{2n(n-0.8)^2}{(b-a)^2}} < 0.05 \Leftrightarrow -\frac{2n(n-0.8)^2}{\left(5-(-3)\right)^2} < ln(0.05) \Leftrightarrow$$

$$\Leftrightarrow n(n-0.8)^2 \geq -ln(0.05) \cdot 32 \Leftrightarrow n(n-0.8)^2 \geq ln(20) \cdot 32$$

ואכן, על ידי פתרון אי-שיוויון זה(שהינו פולינום ב-n), נקבל כי האי-שיוויון מתקיים עבור –

$$n \geq 5.12495$$

וכיוון ש- $\boldsymbol{n} \in \mathbb{N}$, נסיק כי למעשה לכל $\boldsymbol{n} \geq \boldsymbol{6}$, מתקיים האי-שיוויון –

$$\mathbb{P}[S > n^2 + 0.2n] < 0.1$$

כנדרש.

שאלה 4.

סעיף א.

נרצה לחשב את $\mathbb{E}[R_i]$. מתקיים כי - $R_i = \sum_{j=1}^{n} R_{ij}$, ולכן נקבל –

$$\mathbb{E}[R_i] = \mathbb{E}\left[\sum_{j=1}^{n} R_{ij}\right] = \sum_{j=1}^{n} \mathbb{E}[R_{ij}]$$

ועל כן, נרצה לחשב לשם כך את $\mathbb{E}[R_{ij}]$. מהגדרת המשתנה המקרי R_{ij}, מתקיים –

$$R_{ij} = \begin{cases} L_j & if\ server\ i\ was\ assigned\ to\ job\ j \\ 0 & otherwise \end{cases}$$

לפי ההנחה, לכל עבודה j ייבחר סרבר באופן אקראי, ולכן ניתן להסיק כי כל עבודה מתפלגת אחיד עם פרמטרים של $U(1,m)$, כאשר מדובר כאן בהתפלגות אחידה דיסקרטית. על כן –

$$\mathbb{P}[server\ i\ was\ assigned\ to\ job\ j] = \frac{1}{m}$$

מכאן, מהגדרת התוחלת, נחשב את $\mathbb{E}[R_{ij}]$ באופן הבא –

$$\mathbb{E}[\boldsymbol{R_{ij}}] = L_j \cdot \mathbb{P}[server\ i\ was\ assigned\ to\ job\ j] +$$
$$+\mathbf{0} \cdot (1 - \mathbb{P}[server\ i\ was\ assigned\ to\ job\ j]) = \boldsymbol{L_j} \cdot \frac{\mathbf{1}}{\boldsymbol{m}}$$

ובעת, ניתן לחשב את $\mathbb{E}[R_i]$ –

$$\mathbb{E}[\boldsymbol{R_i}] = \sum_{j=1}^{n} \mathbb{E}[R_{ij}] = \sum_{j=1}^{n} L_j \cdot \frac{1}{m} = \frac{1}{m} \cdot \sum_{j=1}^{n} L_j = \frac{\boldsymbol{L}}{\boldsymbol{m}}$$

סעיף ב.

נשתמש ב-$Chernoff\ Multiplicative\ Bound$ על מנת לחסום את הביטוי –

$$\mathbb{P}[R_i \geq (1+\delta)\mathbb{E}[R_i]]$$

ראשית, נבחין כי אכן כל התנאים הדרושים מתקיימים. מתקיים כי –

$$R_i = R_{i1} + R_{i2} + \cdots + R_{in}$$

$$R_{i1},\ R_{i2}, \dots,\ R_{in}\ are\ independent\ as\ the\ assignment\ of\ jobs\ to\ servers\ is\ random$$

מהנתונים מתקיים כי - $0 \leq L_j \leq 1$, $R_{ij} \in \{0, L_j\}$ לכל $0 \leq j \leq n$.

על כן, כל התנאים הדרושים מתקיימים ולכן ניתן להשתמש באי-שיוויון זה עם $\delta = 0.1$ ולקבל –

$$\mathbb{P}[R_i \geq (1+\delta)\mathbb{E}[R_i]] \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\frac{L}{m}} = (\mathbf{0.995170})^{\frac{\boldsymbol{L}}{\boldsymbol{m}}}$$

סעיף ג.

נניח כי $\delta = 0.1$. נרצה למצוא חסם להסתברות כי לפחות אחד מהסרברים הינו עמוס ב- 10% יותר מאשר העומס הצפוי. כלומר, נרצה לחסום את ההסתברות הבאה –

$$\mathbb{P}\big[R_1 \geq (1+\delta)\mathbb{E}[R_1] \;\; or \;\; \ldots. \;\; or \;\; R_m \geq (1+\delta)\mathbb{E}[R_m]\big]$$

ניזכר ראשית בחסם האיחוד, ולפיו מתקיים –

$$\mathbb{P}[A_1 \cup \ldots \cup A_k] \leq \sum_{i=1}^{k} \mathbb{P}[A_i]$$

על כן, לפי חסם האיחוד, ולפי סעיף ב' מתקיים –

$$\mathbb{P}\big[R_1 \geq (1+\delta)\mathbb{E}[R_1] \;\; or \;\; \ldots. \;\; or \;\; R_m \geq (1+\delta)\mathbb{E}[R_m]\big] \;\; \leq$$

$$\leq \sum_{i=1}^{m} \mathbb{P}\big[R_i \geq (1+\delta)\mathbb{E}[R_i]\big] \underbrace{\leq}_{(b)} \sum_{i=1}^{m} (0.995170)^{\frac{L}{m}} = \boldsymbol{m} \cdot (\mathbf{0.995170})^{\boldsymbol{\frac{L}{m}}}$$

Programming Assignment.

שאלה 1.

סעיף א. הקוד –

```python
import numpy as np
from matplotlib import pyplot as plt
from collections import Counter
from scipy.spatial import distance
from sklearn.datasets import fetch_openml
mnist = fetch_openml('mnist_784')
data = mnist['data']
labels = mnist['target']

idx = np.random.RandomState(0).choice(70000, 11000)
train = data[idx[:10000], :].astype(int)
train_labels = labels[idx[:10000]]
test = data[idx[10000:], :].astype(int)
test_labels = labels[idx[10000:]]

# Pre-processing: Calculating distances between all test images and all train images. Done for efficiency reasons.
dists = distance.cdist(test, train)
# test_list variable is for inner use, for time-complexity issues.
test_list = list()
for i in range(0, len(test)):
    test_list.append(list(test[i]))



# parameters: train_imgs - train images, train_labels - train labels, image_query - query image, k - integer parameter
def kNN(train_imgs, train_labels, image_query, k):
    i = test_list.index(list(image_query))                                           # finds index of current image query
    label_counter = Counter()                                                        # useful data structure to count labels
    dists_i = [(dists[i][j], train_labels[j]) for j in range(len(train_imgs))]   # distances of query img from data imgs
    dists_i.sort(key=lambda tup: tup[0])                                             # sorting the distances
    for j in range(k):                                                       # choosing k nearest(first k in sorted list)
        label_counter[int(dists_i[j][1])] += 1                               # counting labels of k nearest neighbors
    return (label_counter.most_common(1))[0][0]                              # returns most common label among k nearest
```

סעיף ב.

נריץ עם $n = 1000$ התמונות הראשונות ב-$training\ images$, עם $k = 10$, על כל תמונה ב- $test\ images$.

הקוד –

```python
# Section b
print("Section (b) runs: ")
n = 1000
k = 10
correct = 0
train_b = train[:n]
train_labels_b = train_labels[:n]
for i in range(0, n):
    prediction = kNN(train_b, train_labels_b, test[i], k)
    if prediction == int(test_labels[i]):
        correct += 1

print("\tCorrect =", correct, "/", n, "=", correct / n)
print("\tPercentage is", 100 * correct / n, "%.")
print("Section (b) done.")
```

הפלט –

```
Section (b) runs:
    Correct = 858 / 1000 = 0.858
    Percentage is 85.8 %.
Section (b) done.
```

אחוז הדיוק הוא כאמור 85.8%, שכן קיבלנו כי האלגוריתם דייק ב-858 מתוך 1000 התמונות ב-$test\ images$. על כן, הדיוק הינו – $accuracy = 858 / 1000 = 0.858$.

אילו ה-$predictor$ היה מנבא תוצאות באופן אקראי, ניתן להסיק כי אחוז הדיוק היה באזור ה-10%, מכיוון שיש 10 תגיות (שכן $Label\ Space = \{0,1,\ldots 9\}$) שונות **וההתפלגות למעשה במצב זה הייתה אחידה כתוצאה מהאקראיות של ה-$predictor$, ולכן ניתן להסיק שאחוז הדיוק היה בתוחלת היה 10%.**
כלומר במקרה זה -

$$\mathbb{E}[prediction\ \ accuracy] = 0.1$$

סעיף ג.

בסעיף זה נחשב את ה-$prediction\ accuracy$ כתלות ב-k, הפרמטר המייצג את מספר השכנים לפיו ה-$predictor$ מחליט על ה-$label$. נשרטט את הגרף המתקבל עבור $k\ =\ 1,2,3,\dots,100$, וכן $n = 1000$.

הקוד –

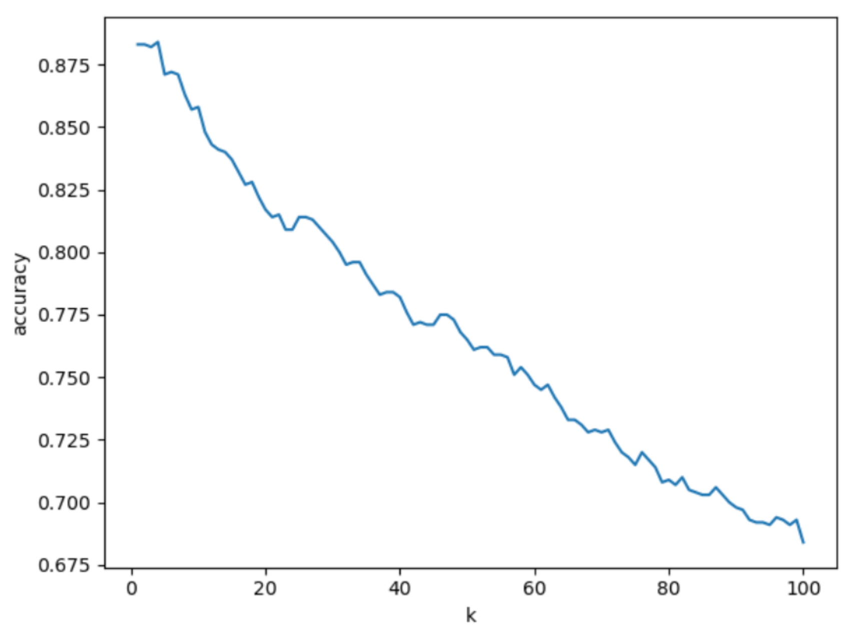
main.py

```python
# Section c
print("Section (c) runs: ")
n = 1000
train_c = train[:n]
train_labels_c = train_labels[:n]
x_labels = [i for i in range(1, 101)]                # x_labels[1,2,..., 100] to be used as k
y_labels = []
for k in x_labels:
    predictions = np.zeros(len(test))
    for i in range(len(test)):
        predictions[i] = kNN(train_c, train_labels_c, test[i], k)
    correct = np.sum(predictions == np.array(test_labels[:n], dtype=int))
    y_labels.append(correct / n)

plt.plot(x_labels, y_labels)
plt.xlabel("k")
plt.ylabel("accuracy")
plt.show()

maximumC = max(y_labels)
index = y_labels.index(maximumC) + 1        # + 1 as i'th index represents k=i+1
print("The best k is", index, "with a value of", maximumC)
print("Section (c) done.")
```

הגרף המתקבל –



הפלט המתקבל –

```
Section (c) runs:
The best k is 4 with a value of 0.884
Section (c) done.
```

נבחין כי ה-k הטוב ביותר מתקבל עבור $k\ =\ 4$, ובו מתקבל דיוק של $accuracy\ \ =\ \ 0.884$. **המגמה העיקרית שנראית בשרטוט היא שעבור *k* הולך וגדל, מתקבל דיוק הולך ופוחת של האלגוריתם שמימשנו.** ניתן להסביר מגמה זו על ידי כך שכאשר מסתכלים על מספר הולך וגדל של שכנים קרובים ביותר(כאשר *k* גדל), מסתכלים על שכנים שנמצאים במרחק הולך וגדל מהנקודה הנוכחית, ולכן ה-*label* שנקבע לפיהם נקבע גם על פי נקודות רחוקות יותר (באופן שבו **המשקל של כל נקודה שווה**), שלא בהכרח מנבאות באופן מיטבי את הנק' הנתונה.

במילים אחרות, כיוון שלכל נקודה באלגוריתם משקל שווה, ככל שנסתכל על מספר הולך וגדל של שכנים קרובים ביותר, **בעצם נקבל החלטה לסיווג ה-$label$ לפי מספר עולה של נקודות שנמצאות במרחקים הולכים וגדלים מהנקודה הנוכחית, ומכך שלכל נקודה ב-k הקרובות ביותר משקל שווה, יתקבל *label* לפי נקודות (שהן תמונות) רחוקות יותר, ולכן התוצאות(הסיווגים ל-$labels$) המתקבלות טובות פחות ככל ש-k גדל.**

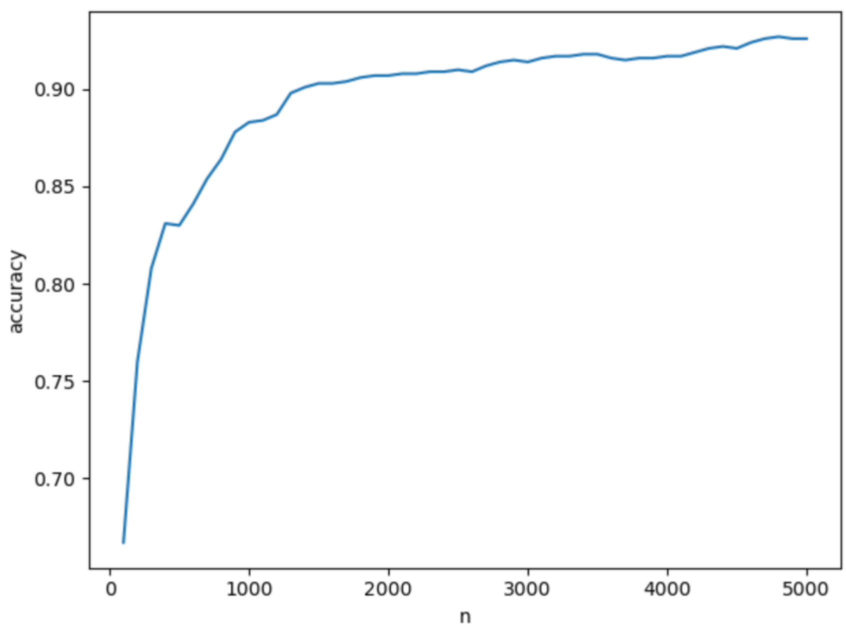
סעיף ד.

בסעיף זה נחשב את ה-$prediction\ accuracy$ כתלות ב-n, כאשר נגדיר גם $k = 1$. נשרטט את הגרף המתקבל עבור $n = 100,200, \ldots ,5000$. הקוד –

```python
# Section d
print("Section (d) runs: ")
k = 1
x_labels = [n for n in range(100, 5001, 100)]   # x_labels=[100,..., 5000] to be used as first n elems in train images
y_labels = []
for n in x_labels:
    train_d = train[:n]
    train_labels_d = train_labels[:n]
    predictions = np.zeros(len(test))
    for i in range(len(test)):
        predictions[i] = kNN(train_d, train_labels_d, test[i], k)
    correct = np.sum(predictions == np.array(test_labels, dtype=int))
    y_labels.append(correct / 1000)


plt.plot(x_labels, y_labels)
plt.xlabel("n")
plt.ylabel("accuracy")
plt.show()
print("Section d done.")
```

הגרף המתקבל –



ניתן לראות את המגמה לפיה כאשר n, שמייצג את מספר תמונות האימון$(training\ images)$ שלפיהם מבצעים את ה-$prediction$, גדל, גם אחוז הדיוק(ה-$accuracy$) בחיזוי ה-$label$ גדל גם הוא. ניזכור כי במקרה זה $k = 1$, כלומר מחפשים את הנקודה(התמונה) הקרובה ביותר לתמונה שרוצים לחזות, ועל כן, ככל שנסתכל על מספר רב יותר של תמונות, סביר להניח כי נמצא תמונה יותר קרובה ויותר מייצגת, שכן לפי אלגוריתם ה- $NN - k$, מרחק גיאומטרי קטן יותר אמור להניב חיזוי $label$ טוב יותר, ועל כן, ככל ש-n גדל, כך גדל גם אחוז הדיוק של האלגוריתם.