Intro to Machine Learning: Assignment #1

May 2, 2023

Dor Bourshan

1 Linear Algebra

1.1 Problem

תזכורת: מטריצה סימטרית A מעל $\mathbb{R}$ ניתנת לכתיבה הבאה $A = QDQ^T$ כאשר :

- D מטריצה אלכסונית, ובאלכסון נמצאים הערכים העצמיים של D
- המטריצה Q היא מטריצה אורתוגונלית כאשר עמודות המטריצה הם וקטורים עצמיים.

סימון: עבור מטריצה B עם ערכים אי שליליים נסמן $\sqrt{B}$ את המטריצה כך שהערך בתא (i,j) הוא $\left(\sqrt{b}\right)_{i,j}$.

נראה שעבור מטריצה סימטרית A מעל $\mathbb{R}$ התנאים הבאים שקולים :

- המטריצה חיובית.
- ניתן לכתוב $A = XX^T$.
- הערכים העצמיים של A אי שליליים.

נחל בשרשרת גרירות מעגלית :

- נניח הערכים העצמיים של A אי שליליים ונראה כי ניתן לכתוב $A = XX^T$.

  מכך שהמטריצה A ניתנת לכתיבה $A = QDQ^T$ כאשר המטריצה D אלכסונית וכן באלכסון של D מופיעים ערכים אי שליליים, נטען כי השוויון הבא תקף $D = \sqrt{D}\sqrt{D}$. לפיכך A ניתנת לכתיבה כ $A = Q\sqrt{D}\sqrt{D}Q^T$. מהיותה של D אלכסונית כך גם $\sqrt{D}$ ולכן מתקיים $\sqrt{D} = \left(\sqrt{D}\right)^T$.

  כידוע עבור מטריצות A, B מתקיים $(AB)^T = B^T A^T$ אם כך מתקיים :

  $$\sqrt{D}Q^T = \left(\sqrt{D}\right)^T Q^T = \left(Q\sqrt{D}\right)^T$$

  נסמן $X = Q\sqrt{D}$. נוכל לרשום

  $$A = XX^T$$

  כנדרש.

- נניח כי ניתן לכתוב $A = XX^T$ ונראה המטריצה A חיובית.

  יהיי $\mathbf{v} \in \mathbb{R}^n$, אנו מעוניינים להראות כי $\mathbf{v}^T A \mathbf{v} \geq 0$. ואכן

  $$\mathbf{v}^T A \mathbf{v} = \mathbf{v}^T XX^T \mathbf{v} \overset{(1)}{=} \left(X^T\mathbf{v}\right)^T \left(X^T\mathbf{v}\right) = \left\langle X^T\mathbf{v}, X^T\mathbf{v}\right\rangle \overset{(2)}{\geq} 0$$

  כאשר :

  (1) : תכונות transpose

  (2) : תכונת מרחב מכפלה פנימית.

- תהיי המטריצה A מטריצה חיובית ונראה כי הערכים העצמיים של A אי שליליים.

יהיי $\vec{0} \neq \mathbf{v} \in \mathbb{R}^n$ וקטור עצמי המתאים לערך עצמי λ. מהגדרת מטריצה חיובית לכל וקטור $\mathbf{x}$ מתקיים כי $\mathbf{x}^T A\mathbf{x} \geq 0$ ולכן מתקיים כי:

$$0 \leq \mathbf{v}^T A\mathbf{v} = \mathbf{v}^T \lambda\mathbf{v} = \lambda\mathbf{v}^T\mathbf{v} = \lambda \langle \mathbf{v}, \mathbf{v} \rangle$$

מכך ש $\mathbf{v} \neq \vec{0}$ מתקיים כי $\langle \mathbf{v}, \mathbf{v} \rangle > 0$ מהגדרת מכפלה פנימית ולכן

$$\lambda \langle \mathbf{v}, \mathbf{v} \rangle \geq 0 \implies \lambda \geq 0$$

כדי להשלים את התמונה הע"ע המתאים לו"ע $\vec{0}$ הוא $\lambda = 0 \geq 0$.

לסיכום הערכים העצמיים של A אי שליליים.

□

1.2 Problem

יהיו $\alpha, \beta \geq 0$ ו $A, B \in \mathbb{R}^{n\times n}$ מטריצות חיוביות.

נראה כי $\alpha A + \beta B$ מטריצה חיובית.

יהיי $\mathbf{v} \in \mathbb{R}^n$.

$$\mathbf{v}^T (\alpha A + \beta B)\,\mathbf{v} \overset{(1)}{=} \left(\alpha\mathbf{v}^T A + \beta\mathbf{v}^T B\right)\mathbf{v} \overset{(2)}{=} \alpha\mathbf{v}^T A\mathbf{v} + \beta\mathbf{v}^T B\mathbf{v} \overset{(3)}{\geq} 0$$

כאשר:

$(1+2)$: אסוציאטיבות ודיסטרטיביות של כפל מטריצות

(3): מהיותן של A, B מטריצות חיובית וכן $\alpha, \beta \geq 0$ מתקיים:

$$\mathbf{v}^T A\mathbf{v} \geq 0 \implies \alpha\mathbf{v}^T A\mathbf{v} \geq 0$$
$$\mathbf{v}^T B\mathbf{v} \geq 0 \implies \beta\mathbf{v}^T B\mathbf{v} \geq 0$$

□

נשים לב כי עבור $A = I$ המטריצה חיובית, כיוון שערכיה העצמיים הם חיובים. מנגד, המטריצה αI עבור $\alpha < 0$ אינה חיובית כיוון שערכיה העצמיים שלילים.

לפיכך אין שמירה על אקסיומת כפל בסקלר.

2 Calculus and Probability

יהיו $X_1, \dots, X_n$ משתנים מקרים בלתי תלויים רציפים מהתפלגות $U([0,1])$. נגדיר

$$Y = \max(X_1, \dots, X_n)$$

2.1 Problem

נמצא את ה PDF של Y, $\mathbb{E}[Y]$ וכן $\mathrm{Var}(Y)$.

- נמצא את PDF

כידוע עבור $X \sim U([0,1])$ מתקיים

$$F_X(x) = \begin{cases} x & 0 \le x \le 1 \\ 1 & x > 1 \\ 0 & x < 0 \end{cases}$$

לפיכך

$$\Pr(Y \le y) = \Pr(\max(X_1, \dots, X_n) \le y) = \Pr(X_1 \le y, \dots, X_n \le y) \overset{\text{i.d.d}}{=} \prod_{i=1}^{n} \Pr(X_i \le y) = (F_X(x))^n$$

אם כן

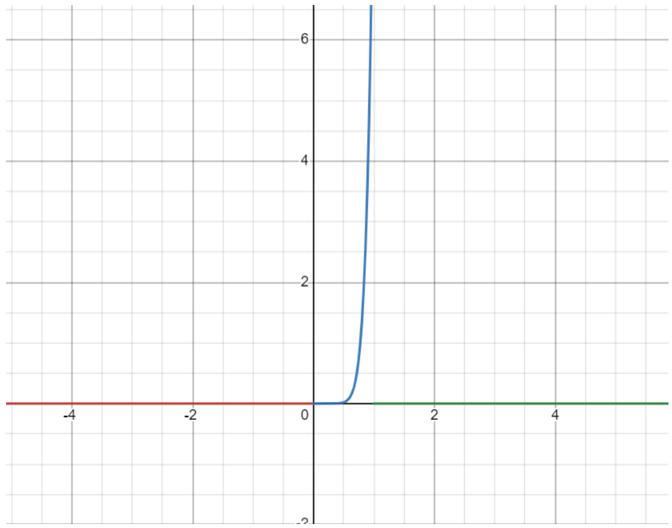
$$F_X^n(x) = \begin{cases} x^n & 0 \le x \le 1 \\ 1 & x > 1 \\ 0 & x < 0 \end{cases}$$

אנו יודעים כי

$$f_X(x) = \frac{\partial}{\partial x} F_X(x)$$

ולכן

$$\boxed{f_Y(y) = \begin{cases} ny^{n-1} & 0 \le y \le 1 \\ 0 & y > 1 \\ 0 & y < 0 \end{cases}}$$



:1 Plot PDF

- נמצא $\mathbb{E}\left[Y\right]$

$$\mathbb{E}\left[Y\right]=\int\limits_{-\infty}^{\infty}y\cdot f_{Y}\left(y\right)\mathrm{d}y=\int\limits_{-\infty}^{0}y\cdot f_{Y}\left(y\right)\mathrm{d}y+\int\limits_{0}^{1}y\cdot f_{Y}\left(y\right)\mathrm{d}y+\int\limits_{1}^{\infty}y\cdot f_{Y}\left(y\right)\mathrm{d}y$$

$$\boxed{\mathbb{E}\left[Y\right]=\frac{n}{n+1}}$$

מתקיים כי

$$\int\limits_{-\infty}^{0}y\cdot0\mathrm{d}y=\int\limits_{1}^{\infty}y\cdot0=0$$

ולכן

$$\mathbb{E}\left[Y\right]=\int\limits_{0}^{1}y\cdot f_{Y}\left(y\right)\mathrm{d}y=\int\limits_{0}^{1}y\cdot ny^{n-1}\mathrm{d}y=\int\limits_{0}^{1}ny^{n}\mathrm{d}y=\left.\frac{n\cdot y^{n+1}}{n+1}\right|_{0}^{1}=\frac{n}{n+1}$$

$$\boxed{\mathbb{E}\left[Y\right]=\frac{n}{n+1}}$$

- נמצא $\mathrm{Var}\left(Y\right)$

כידוע מתקיים כי

$$\mathrm{Var}\left(Y\right)=\mathbb{E}\left[Y^{2}\right]-\mathbb{E}^{2}\left[Y\right]$$

אם כך נמצא $\mathbb{E}\left[Y^{2}\right]$ כאשר ההסברים עבור המעברים דומים מאוד לעיל ולכן נקצר

$$\mathbb{E}\left[Y^{2}\right]=\int\limits_{0}^{1}y^{2}\cdot f_{Y}\left(y\right)\mathrm{d}y=\int\limits_{0}^{1}y^{2}\cdot ny^{n-1}\mathrm{d}y=\int\limits_{0}^{1}ny^{n+1}\mathrm{d}y=\left.\frac{n\cdot y^{n+2}}{n+2}\right|_{0}^{1}=\frac{n}{n+2}$$

מתקיים אם כך

$$\mathrm{Var}\left(Y\right)=\mathbb{E}\left[Y^2\right]-\mathbb{E}^2\left[Y\right]=\frac{n}{n+2}-\left(\frac{n}{n+1}\right)^2$$

$$\boxed{\mathrm{Var}\left(Y\right)=\frac{n}{n+2}-\left(\frac{n}{n+1}\right)^2}$$

- אסימפטוטיקה

$$\mathbb{E}\left[Y\right]=\mathbb{E}\left[\max\left(X_1,\ldots,X_n\right)\right]=\frac{n}{n+1}\xrightarrow[n\to\infty]{}1$$

$$\mathrm{Var}\left(Y\right)=\frac{n}{n+2}-\left(\frac{n}{n+1}\right)^2\xrightarrow[n\to\infty]{}1-1=0$$

□

3 Optimal Classifiers and Decision Rules

3.1 Poblem

3.1.1 section

יהיו X, Y משתנים מקריים כך ש Y מקבל ערכים ב $\mathcal{Y} = \{1, \ldots, L\}$. נגדיר ℓ_{0-1} 0-1 loos function שהוגדרה בכיתה.

נראה כי

$$h = \arg\min_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}\left[\ell_{0-1}\left(Y, f\left(X\right)\right)\right]$$

נתונה ע"י

$$h\left(x\right) = \arg\max_{i\in\mathcal{Y}} \Pr\left[Y = i \mid X = x\right]$$

לפי ההגדרה

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

כאשר אנו משתמשים בהצגה זו רק לצורך הצגה נוחה וכן יכול להיות כי הקבוצה $\mathcal{X}$ אינה בת מנייה.

נקבע x. לפיכך אנו נדרשים למצוא f המביאה למינימום את הביטוי הבא :

$$\sum_{y\in\mathcal{Y}} \Pr\left[X = x, Y = y\right] \ell_{0-1}\left(y, f\left(x\right)\right) = \sum_{y\in\mathcal{Y}} \Pr\left[Y = y \mid X = x\right] \Pr\left[X = x\right] \cdot \ell_{0-1}\left(y, f\left(x\right)\right)$$

כעת נשים לב כי $\Pr\left[X = x\right]$ קבוע אי שלילי ולכן אנו נדרשים למעשה למנמם את הסכום הבא :

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

מתקיים כי $\Pr\left[Y = y \mid X = x\right] \geq 0$ ולכן אנו נרצה להתעלם מהערך y כך ש $\Pr\left[Y = i \mid X = x\right]$ גדול ביותר עבורו (כאשר השתמשנו בהגדרת/ מסקנה מהגדרת ℓ_{0-1}). אם כן :

$$\boxed{h\left(x\right) = \arg\max_{i\in\mathcal{Y}} \Pr\left[Y = i \mid X = x\right]}$$

3.1.2 section

יהיו X,Y משתנים מקריים כך ש Y מקבל ערכים ב $\mathcal{Y}=\{0,1\}$. נגדיר Δ פונקציית loss אסימטרית :

$$\Delta\left(y,\hat{y}\right)=\begin{cases}0 & y=\hat{y}\\ a & y=0,\hat{y}=1\\ b & y=1,\hat{y}=0\end{cases}$$

כאשר $a,b\in(0,1]$.

נמצא

$$h=\arg\min_{f:\mathcal{X}\to\mathcal{Y}}\mathbb{E}\left[\Delta\left(Y,f\left(X\right)\right)\right]$$

לפי ההגדרה

$$\mathbb{E}\left[\Delta\left(Y,f\left(X\right)\right)\right]=\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\Pr\left[X=x,Y=y\right]\Delta\left(y,f\left(x\right)\right)$$

כאשר אנו משתמשים בהצגה זו רק לצורך הצגה נוחה וכן יכול להיות כי הקבוצה $\mathcal{X}$ אינה בת מנייה.

נקבע x. לפיכך אנו נדרשים למצוא f המביאה למינימום את הביטוי הבא :

$$\sum_{y\in\mathcal{Y}}\Pr\left[X=x,Y=y\right]\Delta\left(y,f\left(x\right)\right)=\sum_{y\in\mathcal{Y}}\Pr\left[Y=y\mid X=x\right]\Pr\left[X=x\right]\cdot\Delta\left(y,f\left(x\right)\right)$$

כעת נשים לב כי $\Pr\left[X=x\right]$ קבוע אי שלילי ולכן אנו נדרשים למעשה למנמם את הסכום הבא :

$$\Pr\left[Y=0\mid X=x\right]\Delta\left(0,f\left(x\right)\right)+\Pr\left[Y=1\mid X=x\right]\Delta\left(1,f\left(x\right)\right)$$

האפשרויות שלנו נתונות ע"י

$$\Pr\left[Y=0\mid X=x\right]\Delta\left(0,1\right)+\Pr\left[Y=1\mid X=x\right]\Delta\left(1,1\right)\qquad(1)$$
$$\Pr\left[Y=0\mid X=x\right]\Delta\left(0,0\right)+\Pr\left[Y=1\mid X=x\right]\Delta\left(1,0\right)\qquad(2)$$

כלומר

$$\Pr\left[Y=0\mid X=x\right]\cdot a\qquad(1)$$
$$\Pr\left[Y=1\mid X=x\right]\cdot b\qquad(2)$$

אם כן כדי למנמם נגדיר :

$$\boxed{h\left(x\right)=\begin{cases}0 & \Pr\left[Y=0\mid X=x\right]\cdot a>\Pr\left[Y=1\mid X=x\right]\cdot b\\ 1 & otherwise\end{cases}}$$

3.2 Problem 2

יהיו X, Y משתנים מקריים כך ש Y מקבל ערכים ב $\mathcal{Y} = \{0,1\}$ ו X מקבל ערכים ב $\mathcal{X}$. נניח כי אנו מעוניינים למצוא פונקציית חיזוי $h : \mathcal{X} \to [0,1]$. נגדיר Δ_{log}

$$\Delta_{log}(y, \hat{y}) = -y \log(\hat{y}) - (1-y)\log(1-\hat{y})$$

נמצא

$$h = \arg\min_{f:\mathcal{X}\to[0,1]} \mathbb{E}\left[\Delta_{log}(Y, f(X))\right]$$

לפי ההגדרה

$$\mathbb{E}\left[\Delta(Y, f(X))\right] = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \Pr[X = x, Y = y]\, \Delta_{log}(y, f(x))$$

כאשר אנו משתמשים בהצגה זו רק לצורך הצגה נוחה וכן יכול להיות כי הקבוצה $\mathcal{X}$ אינה בת מנייה.

נקבע x. לפיכך אנו נדרשים למצוא f המביאה למינימום את הביטוי הבא :

$$\sum_{y\in\mathcal{Y}} \Pr[X = x, Y = y]\, \Delta_{log}(y, f(x)) = \sum_{y\in\mathcal{Y}} \Pr[Y = y \mid X = x] \Pr[X = x] \cdot \Delta_{log}(y, f(x))$$

כעת נשים לב כי $\Pr[X = x]$ קבוע אי שלילי ולכן אנו נדרשים למעשה למנמם את הסכום הבא :

$$\Pr[Y = 0 \mid X = x]\, \Delta_{log}(0, f(x)) + \Pr[Y = 1 \mid X = x]\, \Delta_{log}(1, f(x))$$

נסמן לשם הנוחות

$$\Pr[Y = 0 \mid X = x] = p_0 \qquad (1)$$
$$\Pr[Y = 1 \mid X = x] = p_1 \qquad (2)$$

כעת נכתוב מחדש את הסכום עם כתיבה מפורשת של Δ_{log}

$$-p_o \cdot \log(1 - f(x)) + p_1 - p_1 \cdot \log(f(x))$$

ברור כי $0 < f(x) < 1$ אחרת הביטוי אינו מוגדר.

כעת נגזור לפי $f(x)$ ונשווה ל 0 כדי למצוא מינימום, ונקבל

$$\frac{p_0}{1 - f(x)} - \frac{p_1}{f(x)} = 0 \implies f(x) = p_1$$

כאשר השתמשנו בעובדה כי $p_0 + p_1 = 1$.

נוודא כי אכן מינימום ע"י גזירה נוספת :

$$\left(\frac{p_0}{1-f(x)} - \frac{p_1}{f(x)}\right)' = \frac{p_0}{(1-f(x))^2} + \frac{p_1}{(f(x))^2}$$

כעת נציב $f(x) = p_1$ ואנו מניחים כי $p_1 \neq 0, 1$ ונקבל כי

$$\frac{p_0}{(1-p_1)^2} + \frac{p_1}{(p_1)^2} > 0$$

ולכן התקבל מינימום.

אם כך נגדיר

$$\boxed{h(x) = \Pr[Y = 1 \mid X = x] = p_1}$$

3.3 Problem

יהיו X, Y משתנים מקריים כך ש Y מקבל ערכים ב $\mathcal{Y} = \{0,1\}$ וכן $\mathcal{X} = \mathbb{R}$ בהתאמה . נניח כי מתקיים $X_{|Y=0} \sim \mathcal{N}(\mu, \sigma_0^2)$ וכן $X_{|Y=1} \sim \mathcal{N}(\mu, \sigma_1^2)$ כאשר $\sigma_0 \neq \sigma_1$. נניח כי $\Pr[Y=1] = p_1$.

נמצא את מסווג אופטימלי $h : \mathbb{R} \to \{0,1\}$ שמביא למינימום את $\mathbb{E}[\ell_{0-1}(Y, h(X))]$

לפי ההגדרה אם כן :

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

נקבע x

$$\begin{aligned}
&\sum_{y \in \mathcal{Y}} \Pr[X = x, Y = y] \, \ell_{0-1}(y, h(x)) = \\
=& f_{X_{|Y=0}}(x) \Pr[Y=0] \, \ell_{0-1}(0, h(x)) + f_{X_{|Y=1}}(x) \Pr[Y=1] \, \ell_{0-1}(1, h(x)) \\
=& f_{X_{|Y=0}}(x) \cdot (1-p_1) \cdot \ell_{0-1}(0, h(x)) + f_{X_{|Y=1}}(x) \cdot p_1 \cdot \ell_{0-1}(1, h(x))
\end{aligned}$$

כעת התנאי

$$h(x)=\begin{cases}0 & \frac{f_{X_{|Y=0}}(x)}{f_{X_{|Y=1}}(x)}>\frac{p_1}{(1-p_1)}\\ 1 & otherwise\end{cases}$$

נפרש את הביטוי : נשים לב מהיות $f_{X_{|Y=y}}$ פונקצית צפיפות של משתנה גאוסייני היא חיובית (ולכן נוכל לחלק בה) וההסתברות $p_1\in(0,1)$, אחרת לא יכול להיות כי $X_{|Y=y}$ משתנה גאוסייני.

$$\begin{aligned}
f_{X_{|Y=0}}(x)\cdot(1-p_1)>f_{X_{|Y=1}}(x)\cdot p_1 &\iff \frac{f_{X_{|Y=0}}(x)}{f_{X_{|Y=1}}(x)}>\frac{p_1}{(1-p_1)}\\
&\iff \frac{\frac{1}{\sigma_0\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma_0^2}\right)}{\frac{1}{\sigma_1\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma_1^2}\right)}>\frac{p_1}{(1-p_1)}\\
&\iff \frac{\sigma_1}{\sigma_0}\exp\left(\frac{(x-\mu)^2}{2\sigma_1^2}-\frac{(x-\mu)^2}{2\sigma_0^2}\right)>\frac{p_1}{(1-p_1)}\\
&\iff \left(\frac{(x-\mu)^2}{2\sigma_1^2}-\frac{(x-\mu)^2}{2\sigma_0^2}\right)>\ln\left(\frac{p_1}{(1-p_1)}\cdot\frac{\sigma_0}{\sigma_1}\right)\\
&\iff \left(\frac{x^2-2x\mu+\mu^2}{2\sigma_1^2}-\frac{x^2-2x\mu+\mu^2}{2\sigma_0^2}\right)>\ln\left(\frac{p_1}{(1-p_1)}\cdot\frac{\sigma_0}{\sigma_1}\right)\\
&\iff \left(\frac{x^2-2x\mu+\mu^2}{2\sigma_1^2}-\frac{x^2-2x\mu+\mu^2}{2\sigma_0^2}\right)>\ln\left(\frac{p_1}{(1-p_1)}\cdot\frac{\sigma_0}{\sigma_1}\right)\\
&\iff \left(\frac{\sigma_0^2x^2-\sigma_0^22x\mu+\sigma_0^2\mu^2-\sigma_1^2x^2+\sigma_1^22x-\sigma_1^2\mu^2}{\sigma_1^2\sigma_0^2}\right)>2\ln\left(\frac{p_1}{(1-p_1)}\cdot\frac{\sigma_0}{\sigma_1}\right)\\
&\iff \left(\frac{x^2\left(\sigma_0^2-\sigma_1^2\right)+2x\mu\left(\sigma_1^2-\sigma_0^2\right)+\mu^2\left(\sigma_0^2-\sigma_1^2\right)}{\sigma_1^2\sigma_0^2}\right)>2\ln\left(\frac{p_1}{(1-p_1)}\cdot\frac{\sigma_0}{\sigma_1}\right)\\
&\iff x^2\left(\sigma_0^2-\sigma_1^2\right)+2x\mu\left(\sigma_1^2-\sigma_0^2\right)+\mu^2\left(\sigma_0^2-\sigma_1^2\right)>\sigma_1^2\sigma_0^22\ln\left(\frac{p_1}{(1-p_1)}\cdot\frac{\sigma_0}{\sigma_1}\right)
\end{aligned}$$

נסמן $\sigma_1^2\sigma_0^22\ln\left(\frac{p_1}{(1-p_1)}\cdot\frac{\sigma_0}{\sigma_1}\right)=C$ וכן $\left(\sigma_0^2-\sigma_1^2\right)=A$:

נפריד כעת לשני מקרים (כיוון ש $A\neq 0$ לפי הנתון) :

- $A>0$:

$$\begin{aligned}
&\iff x^2A-A2x\mu+\mu^2A>C\\
&\iff x^2-2\mu x+\mu^2>\frac{C}{A}\\
&\iff (x-\mu)^2>\frac{C}{A}
\end{aligned}$$

אם $C<0$: מתקיים לכל x

אם $C\geq 0$:

$$\iff (x-\mu)^2>\frac{C}{A}\iff |x-\mu|>\sqrt{\frac{C}{A}}\iff x>\mu+\sqrt{\frac{C}{A}}\text{or }x<-\mu+\sqrt{\frac{C}{A}}$$

- $A < 0$

$$\iff x^2A - A2x\mu + \mu^2A > C$$
$$\iff x^2 - 2\mu x + \mu^2 < \frac{C}{A}$$
$$\iff (x-\mu)^2 < \frac{C}{A}$$

אם $C \leq 0$: :

$$\iff (x-\mu)^2 < \frac{C}{A} \iff |x-\mu| < \sqrt{\frac{C}{A}} \iff -\mu + \sqrt{\frac{C}{A}} < x < \mu + \sqrt{\frac{C}{A}}$$

אם $C > 0$: אין x המקיים את הנ"ל .

נסכם

$$h(x) = \begin{cases} 0 & A > 0, C < 0 \\ 0 & A > 0, C \geq 0, \left(x > \mu + \sqrt{\frac{C}{A}}\right) \bigvee \left(x < -\mu + \sqrt{\frac{C}{A}}\right) \\ 0 & A < 0, C \leq 0, -\mu + \sqrt{\frac{C}{A}} < x < \mu + \sqrt{\frac{C}{A}} \\ 1 & \text{otherwise} \end{cases}$$
$$C = \sigma_1^2\sigma_0^2 2\ln\left(\frac{p_1}{(1-p_1)} \cdot \frac{\sigma_0}{\sigma_1}\right)$$
$$A = \left(\sigma_0^2 - \sigma_1^2\right)$$

Programing Assignment

1. Visualizing the Hoeffding bound

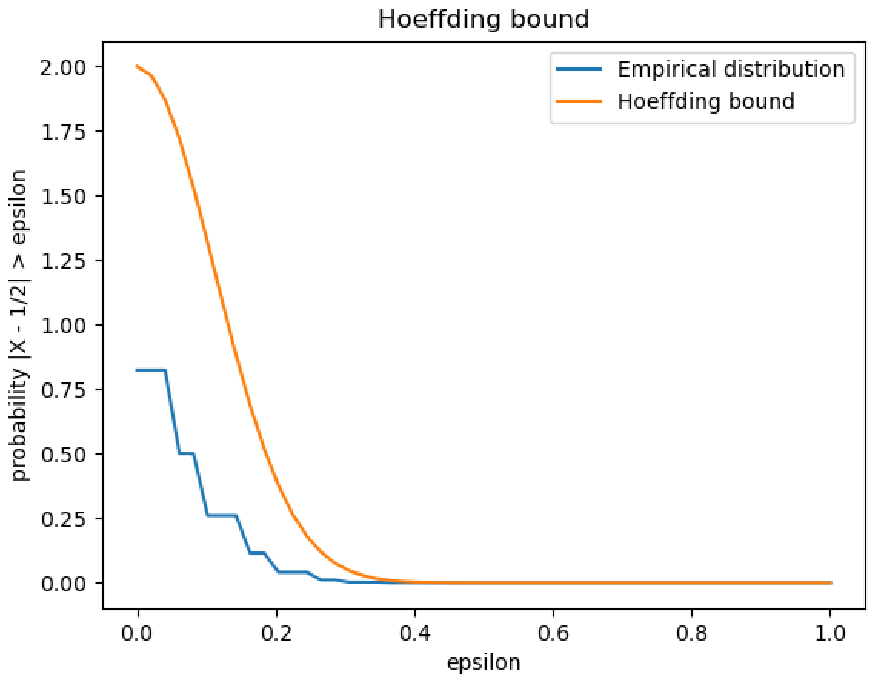


Figure 2: Visualizing the Hoeffding bound

2. Nearest Neighbor

2.1

Submitted to the programing section of the assignment.

2.2

אחוז הדיוק שהתקבל : 84.6%

ממסווג רנדומי לחלוטין היה מצופה שאחוז הדיוק יהיה (בתוחלת) 10% שכן יש עשרה מספרים שיכולים להתקבל בהסתברות שווה.

2.3

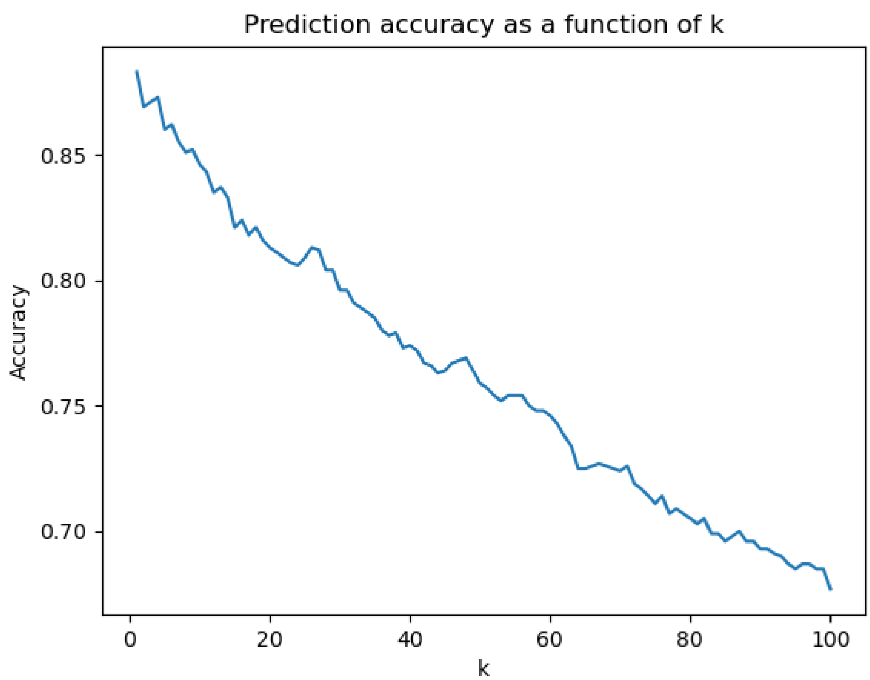


Figure 3: Prediction accuracy as a function of k

ניתן לראות כי ככל ש k גדל כך אחוז הדיוק קטן. ניתן להסביר זאת מפירוש ריצת האלגוריתם באופן הבא : ככל ש k גדל, "*pool*" התמונות שעומד לרשות המסווג גדול יותר לכן נצפה לראות יותר שגיאות. למשל ניתן הסבר שיכול לשכנע : תמונות של הספרה 1 ושל הספרה 7 יכולות להיות דומות, בהינתן תמונה של הספרה 7 שהאלגוריתם מקבל, הוא יחשב מרחקים מהתמונות הקרובות ביותר לפי המרחק האוקלידי. אם כן יש סיכוי לא מבוטל כי תמונות שמתוייגות כ 1, קרובות לתמונה שהאלגוריתם קיבל ולכן נמצאות ב "*pool*" התמונות שעומד לרשות המסווג. אם התמונות התוייגות כ 1 ב "*pool*" הן השכיחות ביותר, האלגוריתם ישגה ויתייג את התמונה כתמונה של הספרה 1.

כבאופן מובהק ניתן לראות כי עבור $k=1$ מתקבל הדיוק הטוב ביותר.

2.4

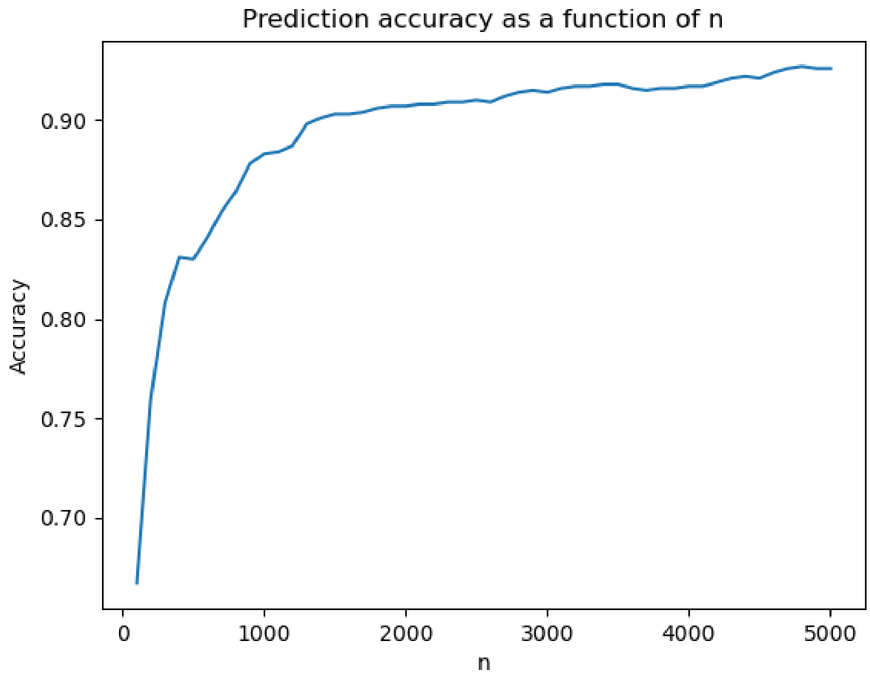


Figure 4: Prediction accuracy as a function of n

ניתן לראות כי ככל ש n גדל כך אחוז הדיוק עולה. ניתן להסביר שעבור n קטן, תמונה בעלת $Label = L_1$ הניתנת לאלגוריתם, תהיה קרובה לתמונה מתוייגת בעלת $Label = L_2$ כאשר $L_1 \neq L_2$ בסיכוי גבוה יותר (כיוון שמספר התמונות להשוות את המרחק קטן יחסית, והאלגוריתם קובע עבור התמונה היחידה, עד כדי מרחקים שווים, הקרובה ביותר אליו) . ככל שיש יותר מידע כך סביר יותר שתמונה בעלת $Label = L_1$ הניתנת לאלגוריתם, תהיה קרובה ביותר לתמונה מתוייגת בעלת אותו L_1.

ניתן לראות שהחל מ $n = 2000$ לערך, תוספת x תמונות מתוייגות אינה באופן מובהק כמו תוספת ל $100 < n < 1000$ ובאופן כללי ניתן לראות שהגרף מתנהג דומה לגרף $\log$. אפשר להסביר זאת שעבור מספיק תמונות מתוייגות (וכיוון שמספר התיוגים $\{0, 1, \ldots, 9\}$ קטן יחסית והתמונות אינן מורכבות, יחסית) האלגוריתם יצליח לקבוע ברמת מובהקות טובה את התיוג המתאים.