מטלה 1 - ת.ז 208935916

Linear Algebra

1. (**15 pts**) A symmetric matrix A over $\mathbb{R}$ is called *positive semidefinite* (PSD) if for every vector $\mathbf{v}$, $\mathbf{v}^\mathsf{T} A\mathbf{v} \geq 0$.

   (a) Show that a symmetric matrix A is PSD if and only if it can be written as $A = XX^T$, if and only if all of its eigenvalues are non-negative.

   Hint: Recall that a real symmetric matrix A can be decomposed as $A = QDQ^T$, where Q is an orthogonal matrix whose columns are eigenvectors of A and D is a diagonal matrix with eigenvalues of A as its diagonal elements.

הוכחה

נוכיח: 1 ⇐ 3 ⇐ 2 ⇐ 1.

נוכיח 1 ⇐ 3:

נניח כי A היא PSD. אזי לכל וקטור v מתקיים $v^tAv \geq 0$. בפרט עבור x_0 וע"ע של A מתקיים

$0 \leq x_0^t A x_0 = x_0^t x_0 \lambda$, כאשר x_0 ע"ע הינו לעצמו. $\Rightarrow 0 \leq x_0^t \cdot x_0 \lambda \Rightarrow 0 \leq \|x_0\|^2 \cdot \lambda$

כפל בסקלר

$\Rightarrow 0 \leq \lambda$.

מכפלה פנימית של x הינו אי-שלילית.

3 ⇐ 2: יהיו $\lambda_1, \dots, \lambda_n$ ע"ע של A. נניח כי לכל $i \in [n]$, $\lambda_i \geq 0$. נוכיח כי A ניתנת להצגה כך - $A = XX^T$.

$$A = QDQ^T = Q \cdot \sqrt{D} \cdot \sqrt{D} \cdot Q^T = (Q\sqrt{D})(\sqrt{D}Q^T)$$

ניתן שכל מטריצה

עם ערכים אי שליליים

$$(\sqrt{D})_{i,j} = \begin{cases} \sqrt{\lambda_i} & i = j \\ 0 & \text{אחרת} \end{cases}$$

* מתקיים $\sqrt{D} \cdot \sqrt{D}^T$, אבן

ונתון $\lambda_i \geq 0$, ולכן מוגדר היטב.

אזי עבור $M = Q\sqrt{D}$, נקבל $A = MM^T$ כנדרש

סימן

2 ← 1: נניח A ניתנת להצגה באופן הבא: $A = XX^T$. נראה כי A היא PSD.

יהי v וקטור. אי שלילי -

$$v^t A v = v^t X X^T v = (v^t x)(x^T v) = \|v^t x\|^2 \geq 0$$

נורמה בריבוע

על הוקטור כולל

$\square$

הראנו $1 \to 3 \to 2 \to 1$ כנדרש.

(b) Show that for all $\alpha, \beta \geq 0$ and PSD matrices $A, B \in \mathbb{R}^{n\times n}$, the matrix $\alpha A + \beta B$ is also PSD. Does this mean that the set of all $n \times n$ PSD matrices over $\mathbb{R}$ is a vector space over $\mathbb{R}$?

הוכחה: יהיו $\alpha, \beta \geq 0$ וכן $A, B \in \mathbb{R}^{n\times n}$ PSD. נראה כי $\alpha A + \beta B$ היא PSD.

יהי וקטור v. נראה כי $v^t(\alpha A + \beta B)v \geq 0$.

$$v^t(\alpha A + \beta B)v = (v^t \alpha A + v^t \beta B)v = v^t \alpha A v + v^t \beta B v = \alpha \underbrace{(v^t A v)}_{\geq 0} + \beta \underbrace{(v^t B v)}_{\geq 0}$$

$$= \alpha \cdot \gamma_1 + \beta \cdot \gamma_2 \geq 0$$

כדי לראות כי קומבינציה מאוגל חיובי אין PSD לא מרחב וקטורי: סמן כאן

לדוגמה $\alpha, \beta < 0$, נקבל (בהתאם בהגדרה לעיל) - $\alpha, \beta, \gamma_1, \gamma_2 \geq 0$

$\alpha \cdot \gamma_1 + \beta \cdot \gamma_2$, כאשר: $\alpha, \beta < 0$, $\gamma_1, \gamma_2 \geq 0$

ובמקרה זה אנו נקבל כי נבחר וקטור שלילי ← בפרט לא מקיים $v^t(\alpha A + \beta B)v$ כנדרש.

אי לא מרחב וקטורי. $\square$

Calculus and Probability

1. (**15 pts**) Let $X_1, ..., X_n$ be i.i.d $U([0,1])$ (uniform) continuous random variables. Let $Y = \max(X_1, ..., X_n)$.

   (a) What is the PDF of Y? Write the mathematical formula and plot the PDF as well. Compute $\mathbb{E}[Y]$ and $\mathrm{Var}[Y]$ - how do they behave as a function of n as n grows large?

בתרגיל:

$$P(Y \le x) = P(X_1 \le x, X_2 \le x, \dots, X_n \le x) = \prod_{i=1}^{n} P(X_i \le x) = F_X^n(x)$$

כאן

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

אז PDF תהיה: $y \in [0,1]$

(מתקבל ע"י נגזרת של F_X).

$$f_X(x) = \begin{cases} 0 & y < 0 \\ n \cdot y^{n-1} & y \in [0,1] \\ 0 & y > 1 \end{cases}$$

$$E[Y] = \int_0^1 x \cdot f_Y(x)\,dx = \int_0^1 x \cdot n \cdot x^{n-1}\,dx = n \cdot \int_0^1 x^n\,dx = n\left[\frac{x^{n+1}}{n+1}\right]\Bigg|_0^1$$

$$= \boxed{\frac{n}{n+1}}$$

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

$$E[Y^2] = \int_0^1 x^2 f_Y(x)\,dx = \int_0^1 x^2 n x^{n-1}\,dx = n\int_0^1 x^{n+1}\,dx = n\left[\frac{x^{n+2}}{n+2}\right]\Bigg|_0^1$$

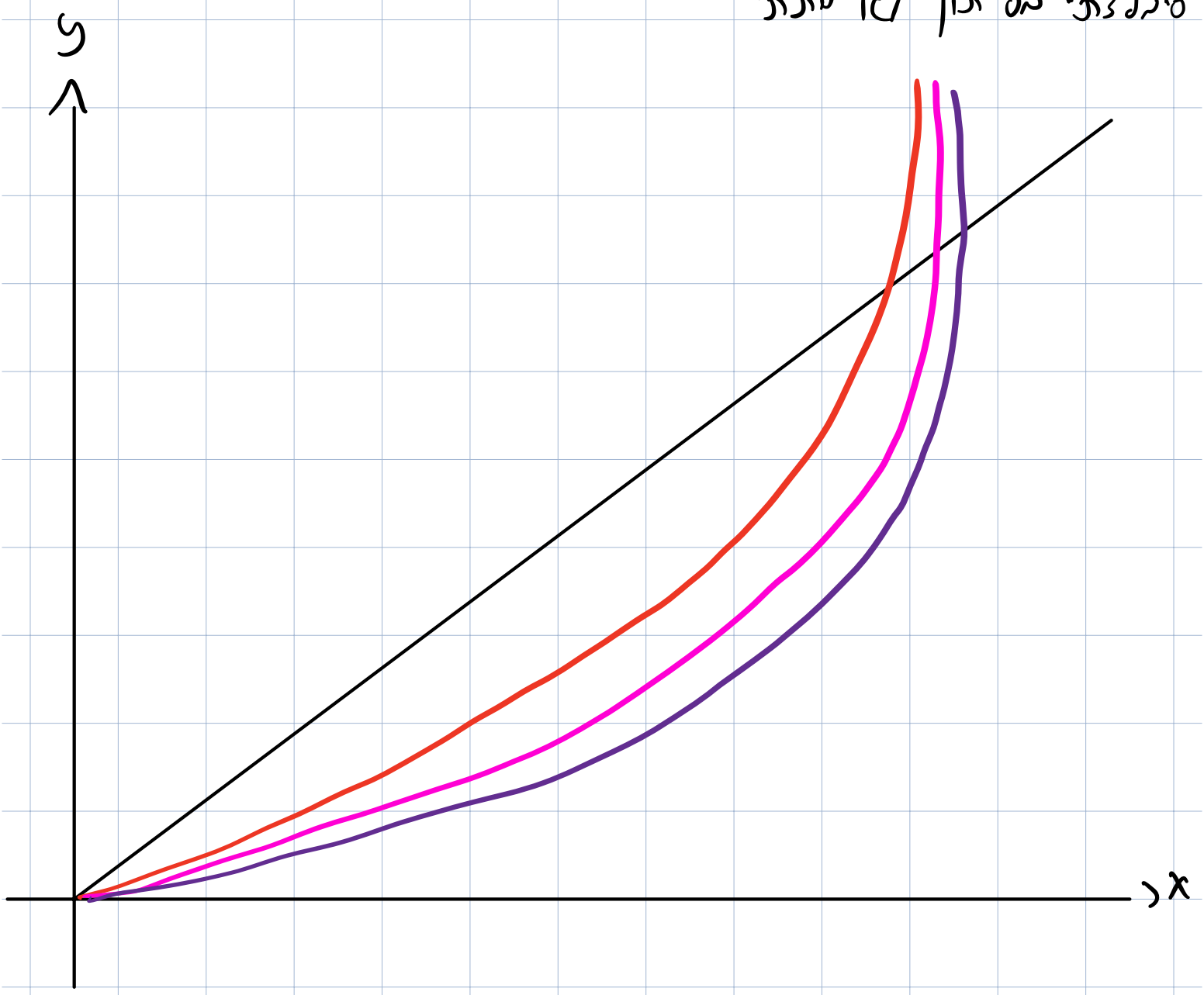
$$= \frac{n}{n+2}$$

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

$$= \frac{n^3+2n^2+n-n^3-2n^2}{(n+2)(n+1)^2} = \boxed{\frac{n}{(n+2)(n+1)^2}}$$

איור המבוקש.

סימלצית הפיתון שאיתה:



Optimal Classifiers and Decision Rules

1. (**15 pts**)

   (a) Let X and Y be random variables where Y can take values in $\mathcal{Y} = \{1, \ldots, L\}$. Let ℓ_{0-1} be the 0-1 loss function defined in class. Show that $h = \arg\min_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}\left[\ell_{0-1}(Y, f(X))\right]$ is given by

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

נרצה למקסם כמה הניתן את הביטוי

$P[X=x, Y=y]\, \ell_{0-1}(y, h(x))$

נפרק כפי שראינו בשיעור -

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

$\sum_{i\in\mathcal{Y}} P[X=x, Y=i] \cdot \ell_{0-1}(i, h(x)) = P[X=x] \left(\sum_{i\in\mathcal{Y}} P[Y=i|X=x] \cdot \ell_{0-1}(i, h(x))\right)$

$$= \begin{cases} \sum_{1\neq j\in\mathcal{Y}} P[Y=i|X=x] & h(x)=1 \\ \\ \sum_{L\neq j\in\mathcal{Y}} P[Y=i|X=x] & h(x)=L \end{cases}$$

$\leftarrow (*) \rightarrow$ * השמטת הקבוע

הכפלי $P[X=x]$

על מנת למקסם את $(1 - P[Y=h(x)|X=x])$ נרצה לבחור את

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

(b) Let X and Y be random variables where Y can take values in $\mathcal{Y} = \{0, 1\}$. Let Δ be the following asymmetric loss function:

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

where $a, b \in (0, 1]$ (note that this loss function generalizes the 0-1 loss defined in class). Compute the optimal decision rule h for the loss function Δ, i.e. the decision rule which satisfies:

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

פתרון בדומה לסעיף (א), יש שינוי בהתאם.

לכל $x \in X$ נסה להקטין את הביטוי

$$P[X=x, Y=y]\,\Delta(y, h(x))$$

מתקיים

$$P[X=x, Y=y]\,\Delta(y, h(x)) = P[X=x, Y=0]\,\Delta(0, h(x)) +$$

$$+ P[X=x, Y=1]\cdot\Delta(1, h(x)) =$$

$$= P[X=x]\cdot\left(P[Y=0 \mid X=x]\,\Delta(0, h(x)) + P[Y=1 \mid X=x]\,\Delta(1, h(x))\right)$$

$$= \begin{cases} a\cdot P[Y=0 \mid X=x] & h(x)=1 \\ b\cdot P[Y=1 \mid X=x] & h(x)=0 \end{cases}$$

$$h(x) = \arg\min_{y \in \{0,1\}} (a\cdot y)(b(1-y))\cdot P[Y=1-y \mid X=x] \iff$$

$$a\cdot P[Y=0 \mid X=x] \le b\cdot P[Y=1 \mid X=x]$$

$$\Rightarrow h(x) = \begin{cases} 1 & \\ 0 & b\cdot P[Y=1 \mid X=x] \le a\cdot P[Y=0 \mid X=x] \end{cases}$$

* הצמצם קבוע כפל $P[X=x]$ $\square$

2. (**15 pts**) Let X and Y be random variables where X can take values in some set $\mathcal{X}$ and Y can take values in $\mathcal{Y} = \{0, 1\}$ (i.e. binary label space). Assume we wish to find a predictor $h : \mathcal{X} \to [0, 1]$ (note that the hypothesis can output any number between 0 and 1) which minimizes $\mathbb{E}[\Delta_{log}(Y, h(X))]$, where Δ_{log} is the following loss function known as the *log-loss*:

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

Find the predictor $h : \mathcal{X} \to [0, 1]$ which minimizes $\mathbb{E}[\Delta_{log}(Y, h(X))]$.

Note: This loss function may seem odd at first, but it is very important and we'll discuss it further in the future.

פתרון

מתקיים לכל $\hat{x} \in X$ נוכל להקטין את הביטוי

$$P[X=\hat{x} \mid Y=y] \cdot \Delta_{log}(y, h(\hat{x}))$$

נסמן $h(\hat{x}) = \hat{y}$. נקבל:

$$P[X=\hat{x}, Y=y] \cdot \Delta_{log}(y, \hat{y}) =$$

$$= P[X=x \mid Y=0] \cdot \Delta_{log}(0, \hat{y}) + P[X=x \mid Y=1] \Delta_{log}(1, \hat{y})$$

$$= P[X=x] \cdot \left(P[Y=0 \mid X=x] \cdot \Delta_{log}(0, \hat{y}) + P[Y=1 \mid X=x] \Delta_{log}(1, \hat{y}) \right)$$

$$= P[Y=0 \mid X=x] \left(-\log(1-\hat{y})\right) + P[Y=1 \mid X=x] \left(-\log(\hat{y})\right)$$

$\Downarrow$

צריך קבוע כפי שענה

אם מתקיים $P[Y=0 \mid X=x] = 0$ נבחר $\hat{y} = 1$

אם $P[Y=1 \mid X=x] = 0$ את נבחר $\hat{y} = 0$.

אחרת, נאור כדי למנוע מינוס - $0 = \frac{P[Y=0 \mid X=x]}{1-\hat{y}} - \frac{P[Y=1 \mid X=x]}{\hat{y}}$

$$\Rightarrow P[Y=0 \mid X=x] \cdot \hat{y} = P[Y=1 \mid X=x] (1-\hat{y})$$

$$\Rightarrow \frac{(1-\hat{y})}{\hat{y}} = \frac{P[Y=0 \mid X=x]}{P[Y=1 \mid X=x]} \Rightarrow \hat{y} = \left(\frac{P[Y=0 \mid X=x]}{P[Y=1 \mid X=x]} + 1 \right)^{-1}$$

$$= \left(\frac{P[Y=0 \mid X=x] + P[Y=1 \mid X=x]}{P[Y=1 \mid X=x]} \right)^{-1} = P[Y=1 \mid X=x]$$

$$= P[Y=0 \mid X=x] \qquad \square$$

3. (**10 pts**)

Let X and Y be random variables taking values in $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \{0, 1\}$ respectively, and assume that given $Y = 0$, X is distributed normally with mean μ and variance σ_0^2, i.e. $X \sim \mathcal{N}(\mu, \sigma_0^2)$, and similarly, given $Y = 1$, $X \sim \mathcal{N}(\mu, \sigma_1^2)$, where $\sigma_0 \neq \sigma_1$. Also assume $\Pr[Y = 1] = p_1$.

Find the optimal decision rule for this distribution and the zero-one loss, i.e. find $h : \mathbb{R} \to \{0, 1\}$ which minimizes $\mathbb{E}[\ell_{0-1}(Y, h(X))]$ where ℓ_{0-1} is the zero-one loss defined in class (write the decision rule only in terms of $x, \mu, \sigma_0, \sigma_1$ and p_1).

פתרון

ראינו כי הוא decision הכי מקטן אותה 0-1 loss במקרה הבא:

$$h(x) = \begin{cases} 1 & P[Y=1 \mid X=x] > P[Y=0 \mid X=x] \\ 0 & \text{אחרת} \end{cases}$$

$$P[Y=1 \mid X=x] > P[Y=0 \mid X=x]$$ נמצא עבור אילו ערכי x מתקיים

$$\frac{f_X(x \mid Y=1)\, P[Y=1]}{f_X(x)} > \frac{f_X(x \mid Y=0)\, P[Y=0]}{f_X(x)}$$ לפי שקילות

$$\frac{f_X(x \mid Y=1)\, P[Y=1]}{f_X(x)} > \frac{f_X(x \mid Y=0)\, P[Y=0]}{f_X(x)}$$ $\Leftarrow$ ארבעת אי־השוויונות:

$$\frac{f_X(x \mid Y=1)}{f_X(x \mid Y=0)} > \frac{P[Y=0]}{P[Y=1]} \qquad \Leftrightarrow$$

$$\Leftrightarrow$$

$$\frac{\frac{1}{\sqrt{2\pi\sigma_1^2}}\cdot e^{\frac{(x-\mu)^2}{2\sigma_1^2}}}{\frac{1}{\sqrt{2\pi\sigma_0^2}}\cdot e^{\frac{(x-\mu)^2}{2\sigma_0^2}}} > \frac{1-P_1}{P_1} \iff$$

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

$$> \frac{1-P_1}{P_1} \iff \frac{(x-\mu)^2}{2\sigma_0^2}-\frac{(x-\mu)^2}{2\sigma_1^2} > \log\left(\frac{1-P_1}{P_1}\cdot\frac{\sigma_0}{\sigma_1}\right)$$

$$\iff (x-\mu)^2\left(\frac{1}{2\sigma_0^2}-\frac{1}{2\sigma_1^2}\right) > \log\left(\left(\frac{1-P_1}{P_1}\right)\cdot\frac{\sigma_0}{\sigma_1}\right)$$

נחלק לשתי מקרים:

1) $\sigma_0 < \sigma_1 \Rightarrow \frac{1}{2\sigma_0^2}-\frac{1}{2\sigma_1^2} > 0 \Rightarrow (x-\mu)^2\left(\frac{1}{2\sigma_0^2}-\frac{1}{2\sigma_1^2}\right) >$

$$> \log\left(\left(\frac{1-P_1}{P_1}\right)\frac{\sigma_0}{\sigma_1}\right) \iff (x-\mu)^2 > \log\left(\left(\frac{1-P_1}{P_1}\right)\cdot\frac{\sigma_0}{\sigma_1}\right)\left(\frac{1}{2\sigma_0^2}-\frac{1}{2\sigma_1^2}\right)$$

$(x-\mu)^2 > \log\left(\left(\frac{1-P_1}{P_1}\right)\cdot\frac{\sigma_0}{\sigma_1}\right)\left(\frac{1}{2\sigma_0^2}-\frac{1}{2\sigma_1^2}\right)$ אם $h(x)=1$

אחרת $h(x)=0$.

ב) $\sigma_0 > \sigma_1 \Rightarrow \frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2} < 0 \Rightarrow (x-\mu)^2\left(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}\right) >$

$> \log\left(\left(\frac{1-p_1}{p_1}\right)\frac{\sigma_0}{\sigma_1}\right) \Leftrightarrow (x-\mu)^2 < \log\left(\left(\frac{1-p_1}{p_1}\right)\cdot\frac{\sigma_0}{\sigma_1}\right)\left(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}\right)$

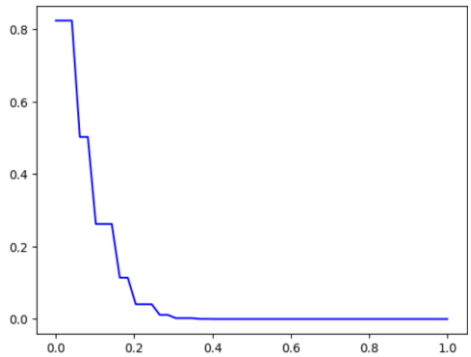
לכן $h(x)=1$ אם $(x-\mu)^2 < \log\left(\left(\frac{1-p_1}{p_1}\right)\frac{\sigma_0}{\sigma_1}\right)\left(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}\right)$

אחרת $h(x)=0$. $\square$

חלק תכנותי:

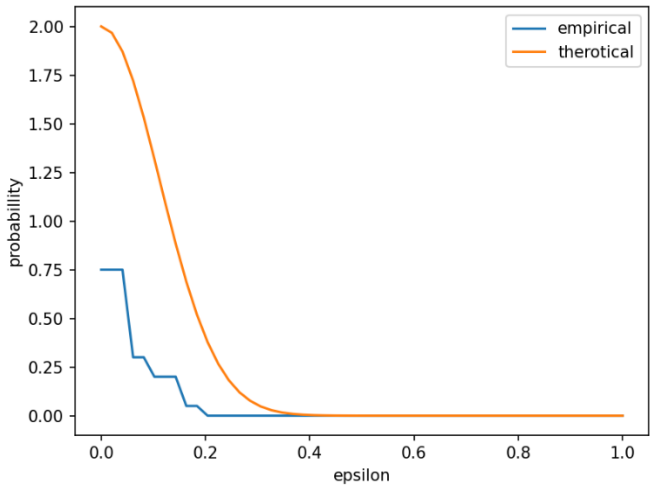
שאלה 1

סעיפים א+ב. מצורפת גרסה רציפה.



שאלה 1

סעיף ג'

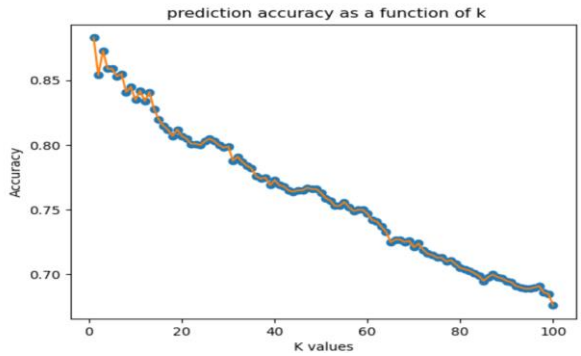


שאלה 2

עבור הפרמטרים איתם התבקשנו להריץ : k=10, n=1000. הדיוק שהתקבל הינו 0.846.

המשמעות הינה – 846 מתמונות המבחן קיבלו פרדיקציה נכונה. גודל המדגם של הlables הינו 10 (הספרות 9-0) -> נצפה לדיוק של 0.1 ע״י predictor שיפעל באופן רדנומלי.

סעיף ג׳



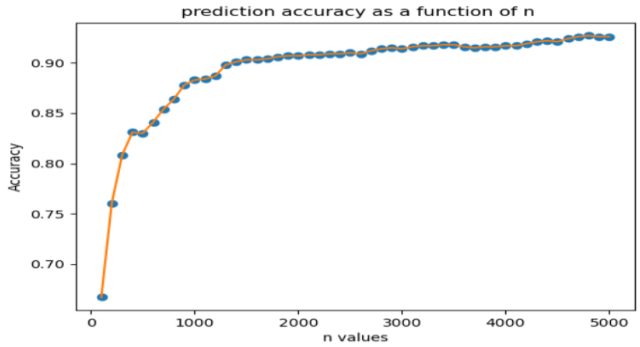
קל לראות כי ככל שערך K גדל -> רמת הדיוק של הפרדיקציה פוחתת.

נשים לב כי עבור k=1 אנו מקבלים את הדיוק הטוב ביותר.

הסבר אפשרי לכך : ככל שערך K גדל, תוצאת ההערכה (הפרדיקציה) תהיה מושפעת ע״י תמונות (ובפרט labels שלהן בהתאמה) אשר פחות קרובות לתמונה המקורית.

למשל עבור המקרה בו k הוא מספר כל התמונות, נקבל כי אין משמעות לקירוב הנל, שכן כל התמונות יהיו ״קרובות״ במידה שווה לתמונה שלנו.

סעיף ד׳



לעומת הסעיף הקודם – כאן רמת הדיוק של הפרדיקציה גדלה ככל שערך n גדל.

הסבר מתבקש – ככל שמדגם המבחן שלנו גדול יותר (בדוגמא הנל יותר תמונות להתאמן עליהן) נקבל רמת דיוק גבוהה יותר.