תרגיל 1 - מבוא ללמידה חישובית

אייל גרינברג 207129792

Calculus and Probability

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

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

נתחיל מחישוב ה CDF של Y:

כדי שהמקס' יהיה $\le t$ כל X_i צריך להיות $\le t$.

ב"ת X_i

$$F_Y(t) = P_r(Y \le t) = P_r(\max\{X_1, ..., X_n\} \le t) \overset{\downarrow}{=} P_r(X_1 \le t, X_2 \le t, ..., X_n \le t) =$$

X_i שווי התפלגות   לפי הגדרת CDF.

$$= \prod_{i=1}^{n} P_r(X_i \le t) \overset{\downarrow}{=} \prod_{i=1}^{n} P_r(X_1 \le t) \overset{\downarrow}{=} \prod_{i=1}^{n} F_{X_1}(t) = \left[F_{X_1}(t)\right]^n = \begin{cases} 1 & 1 < t \\ t^n & 0 \le t \le 1 \\ 0 & t < 0 \end{cases}$$

לפי CDF של $X_1 \sim U[0,1]$

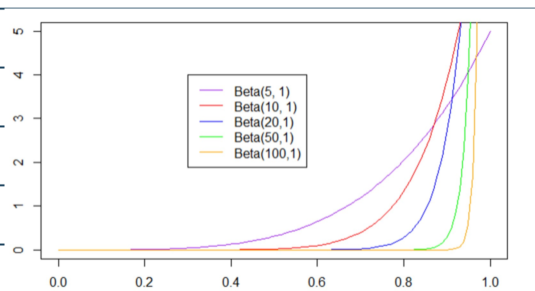
מכיוון שהתפלגות המקסימום של מ"מ $U[0,1]$ הינה התפלגות רציפה, ניתן לקבל את הצפיפות של Y ע"י גזירת ה CDF:

$$f_Y(t) = \left[F_Y(t)\right]' = n \cdot t^{n-1} \cdot \underline{I_{(0,1)}}$$

אינדיקטור המציין האם $t \in [0,1]$, שכן מתקיים:

$$f_{X_i}(t) = \begin{cases} 1 & , t \in [0,1] \\ 0 & , else \end{cases}$$

נשים לב ש: $Y \sim Beta(n,1)$ (עפ"י ה PDF).



$\Leftarrow f_{Beta(n,1)}(y)$

להתפלגות Beta 2 פרמטרים: α, β במקרה שלנו $\alpha = n, \beta = 1$.

אם $X \sim Beta(\alpha, \beta)$ מתקיים: $E[X] = \frac{\alpha}{\alpha+\beta}$, $Var[X] = \frac{\alpha \cdot \beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

בגלל $X \sim Beta(\alpha,\beta)$ מתקיים: $E[X] = \frac{\alpha}{\alpha+\beta}$, $Var[X] = \frac{\alpha \cdot \beta}{(\alpha+\beta)^2 \cdot (\alpha+\beta+1)}$

נציב $\alpha=n, \beta=1$ עבור $Y \sim Beta(n,1)$ ונקבל:

$$E[Y] = \frac{n}{n+1} \Rightarrow \lim_{n\to\infty} E[Y] = \lim_{n\to\infty} \frac{n}{n+1} = 1$$

$$Var[Y] = \frac{n}{(n+1)^2 \cdot (n+2)} \Rightarrow \lim_{n\to\infty} Var[Y] = \lim_{n\to\infty} \frac{n}{(n+1)^2 \cdot (n+2)} = 0$$

2. **(extra credit - 5 pts)** Let X be a continuous random variable. The *median* of X is defined to be the number m which satisfies $\Pr[X \le m] = \frac{1}{2}$ (Verify that you understand why there must exist such m. You can assume for simplicity that m is unique). Show that

$$m = \text{argmin}_{a\in\mathbb{R}} \mathbb{E}[|X-a|]$$

נסמן $h(a)=E[|X-a|]$ ונביא למינימום את $h(a)$.

$$h(a) = E[|X-a|] = \int_{-\infty}^{\infty} |X-a| \cdot f_X(x)\,dx = \underbrace{\int_{-\infty}^{a} (a-x) \cdot f_X(x)\,dx}_{*} + \underbrace{\int_{a}^{\infty} (x-a) \cdot f_X(x)\,dx}_{*}$$

נגזור את $h(a)$ ונשווה ל-0, נשתמש בכלל לייבניץ לגזירה תחת סימן האינטגרל:

ניסוח הכלל [עריכת קוד מקור | עריכה]

תהי $f(x,y)$ פונקציה רציפה במלבן $[a,b] \times [\alpha,\beta]$, וגזירה ברציפות לפי y ($\frac{\partial f}{\partial y}$ קיימת ורציפה). נניח בנוסף שהפונקציות $a(y), b(y)$ גזירות בקטע $[\alpha,\beta]$. אזי

$$\frac{d}{dy}\int_{a(y)}^{b(y)} f(x,y)\,dx = \int_{a(y)}^{b(y)} \frac{\partial f}{\partial y}(x,y)\,dx + f(b(y),y)b'(y) - f(a(y),y)a'(y).$$

$$\frac{d*}{da} = \int_{-\infty}^{a} f_X(x)\,dx + \cancelto{0}{\overbrace{(a-a)}^{0} \cdot f_X(a) \cdot 1} - \cancelto{0}{(a+\infty) \cdot f_X(-\infty) \cdot 0}$$

$$\frac{d*}{da} = \int_{a}^{\infty} -f_X(x)\,dx + \cancelto{0}{(\infty-a) \cdot f_X(\infty) \cdot 0} - \cancelto{0}{\overbrace{(a-a)}^{0} \cdot f_X(a) \cdot 1}$$

$$\Rightarrow \frac{dh}{da} = \frac{d*}{da} + \frac{d*}{da} = \int_{-\infty}^{a} f_X(x)\,dx - \int_{a}^{\infty} f_X(x)\,dx = 0 \Rightarrow \int_{-\infty}^{a} f_X(x)\,dx = \int_{a}^{\infty} f_X(x)\,dx$$

$$\Rightarrow Pr(X \le a) = Pr(X > a)$$

כיוון שהמאורעות $\{X>a\}, \{X \le a\}$ זרים ומשלימים מתקיים:

$$Pr(X \le a) + Pr(X > a) = 1$$

לכן נקבל $\Pr(X \le a) = \Pr(X > a) = \frac{1}{2} \Longleftarrow a$ הוא החציון והוא מביא למינימום את הביטוי. ■

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

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

נראה את הדרוש באופן מוסיף דומה למקרה הבינארי שהוצג בהרצאה.

נסמן בדומה להרצאה את תוחלת ההפסד שנרצה להביא למינימום:

$$\mathcal{L}(h) := \mathbb{E}\left[\ell(Y, h(X))\right] = \sum_{j=1}^{L} \Pr[X = x, Y = i] \cdot \ell(i, h(x))$$

כעת, בהינתן $X = \bar{x}$, נשתמש בנוסחת בייס שלפיה:

$$\Pr(Y = y | X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)} \Longleftrightarrow \Pr(X = x) \cdot \Pr(Y = y | X = x) = \Pr(X = x, Y = y)$$

ונקבל:

$$\mathcal{L}(h) = \sum_{i=1}^{L} \Pr(X = \bar{x}) \cdot \Pr(Y = i | X = \bar{x}) \cdot \ell(i, h(\bar{x}))$$

$$= \Pr(X = \bar{x}) \cdot \sum_{i=1}^{L} \Pr(Y = i | X = \bar{x}) \cdot \ell(i, h(\bar{x}))$$

עפ"י הגדרת המסווג h, $h(\bar{x}) \in \{1, \ldots, L\}$.

לפי הגדרת פונקציית ההפסד, $\ell_{0-1}(i, h(\bar{x})) = \begin{cases} 0 & , h(\bar{x}) = i \\ 1 & , h(\bar{x}) \neq i \end{cases}$

כעת נקבל:

$$\mathcal{L}(h) = \Pr(X = \bar{x}) \cdot \sum_{\substack{j=1 \\ j \neq i}}^{L} \Pr(Y = j | X = \bar{x}) \quad \text{if } h(\bar{x}) = i$$

נשתמש במאורע המשלים ונקבל:

$$\mathcal{L}(h) = \Pr(X = \bar{x}) \cdot \left(1 - \Pr(Y = i | X = \bar{x})\right) \quad \text{if } h(\bar{x}) = i$$

כעת, קל לראות ש $\mathcal{L}(h)$ מקבל ערך מינימלי כאשר $h(\bar{x}) = \arg\max_{i\in\mathcal{Y}} \Pr(Y = i | X = \bar{x})$, כנדרש. ■

(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:

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

נפעל באופן דומה לסעיף הקודם, בהינתן $X=\bar{x}$

$$\mathcal{L}(h) := \mathbb{E}[\Delta(Y, h(X))] = Pr(X=\bar{x}, Y=0)\cdot\Delta(0, h(\bar{x})) + Pr(X=\bar{x}, Y=1)\cdot\Delta(1, h(\bar{x}))$$

נפעל את כלל ב״ס ונקבל:

$$\mathcal{L}(h) = Pr(X=\bar{x})\cdot\left[\overbrace{Pr(Y=0 \mid X=\bar{x})\cdot\Delta(0, h(\bar{x})) + Pr(Y=1 \mid X=\bar{x})\cdot\Delta(1, h(\bar{x}))}^{*}\right]$$

$h(\bar{x}) \in \{0,1\}$, נחלק למקרים:

• אם $h(\bar{x}) = 0$:

$$* = \cancel{Pr(Y=0 \mid X=\bar{x})\cdot\overbrace{\Delta(0,0)}^{0}} \xrightarrow{} 0 + Pr(Y=1 \mid X=\bar{x})\cdot\overbrace{\Delta(1,0)}^{b} = b\cdot Pr(Y=1 \mid X=\bar{x})$$

• אם $h(\bar{x}) = 1$:

$$* = a\cdot Pr(Y=0 \mid X=\bar{x})$$

אנו חפצים ש $*$ יהיה נמוך לכן נקבע את $h(\bar{x})$ להיות:

$$h(\bar{x}) = \begin{cases} 0 & , b\cdot Pr(Y=1 \mid X=\bar{x}) < a\cdot Pr(Y=0 \mid X=\bar{x}) \\ 1 & , \text{else} \end{cases}$$

2. (**25 pts**) Let $\mathbf{X} = (X_1, \dots, X_d)^\mathsf{T}$ be a vector of random variables. $\mathbf{X}$ is said to have a **multivariate normal (or Gaussian) distribution** with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and a $d \times d$ positive definite[1] covariance matrix Σ, if its probability density function is given by

$$f(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

where $\mathrm{E}[X_i] = \mu_i$ and $cov(X_i, X_j) = \Sigma_{ij}$ for all $i, j = 1, \dots, d$. We write this as $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

In this question, we generalize the decision rule we have seen in the recitation to more cases. Assume that the data is $\langle \mathbf{x}, y \rangle$ pairs, where $\mathbf{x} \in \mathbb{R}^d$ and $y \in \{0, 1\}$. Denote by $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ the probability density functions of $\mathbf{x}$ given each of the label values. It is known that f_0, f_1 are multivariate Gaussian:

$$f_0(\mathbf{x}) = f(\mathbf{x}; \boldsymbol{\mu}_0, \Sigma_0),$$
$$f_1(\mathbf{x}) = f(\mathbf{x}; \boldsymbol{\mu}_1, \Sigma_1).$$

Also, the probability to sample a positive sample (i.e. $y = 1$) is p. Assume throughout the question that Σ_0, Σ_1 are positive definite diagonal matrices.

(b) Assume that $\Sigma_0 = \Sigma_1 = \Sigma$. We are given a point $\mathbf{x}$ and we need to label it with either $y = 0$ or $y = 1$. Suppose our decision rule is to decide $y = 1$ if and only if $\mathbb{P}[y = 1|\mathbf{X}] > \mathbb{P}[y = 0|\mathbf{X}]$. Find a simpler condition for $\mathbf{X}$ that is equivalent to this rule. Solve for general $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1 \in \mathbb{R}^d$ and diagonal $\Sigma \in \mathbb{R}^{d \times d}$.

כלל ב"ס במקרה הרציף:

$$\Pr[Y=y \mid X=x] = \frac{f_X(x \mid Y=y) \cdot \Pr(Y=y)}{f_X(x)}$$

כלל ההחלטה יהיה לקבוע $y=1$ אמ"מ:

$$\Pr[Y=1 \mid X] > \Pr[Y=0 \mid X] \iff$$

$$\frac{f_X(x \mid Y=1) \cdot \Pr(Y=1)}{\cancel{f_X(x)}} > \frac{f_X(x \mid Y=0) \cdot \Pr(Y=0)}{\cancel{f_X(x)}} \iff \frac{f_1(x)}{f_0(x)} > \frac{1-p}{p}$$

$$\iff \frac{f(x; \mu_1, \Sigma)}{f(x; \mu_0, \Sigma)} > \frac{1-p}{p} \iff \frac{\cancel{\frac{1}{(2\pi)^{n/2} \cdot |\Sigma|^{1/2}}} \cdot \exp\left[-\frac{1}{2}(x-\mu_1)^t \Sigma^{-1} \cdot (x-\mu_1)\right]}{\cancel{\frac{1}{(2\pi)^{n/2} \cdot |\Sigma|^{1/2}}} \cdot \exp\left[-\frac{1}{2}(x-\mu_0)^t \cdot \Sigma^{-1} \cdot (x-\mu_0)\right]} > \frac{1-p}{p}$$

$$\iff \exp\left[\frac{1}{2}\left((x-\mu_0)^t \Sigma^{-1}(x-\mu_0) - (x-\mu_1)^t \Sigma^{-1}(x-\mu_1)\right)\right] > \frac{1-p}{p}$$

$$\iff \underbrace{(x-\mu_0)^t \cdot \Sigma^{-1}(x-\mu_0)}_{*} - \underbrace{(x-\mu_1)^t \Sigma^{-1}(x-\mu_1)}_{*} > 2 \cdot \ln\left(\frac{1-p}{p}\right)$$

נשים לב שהביטוי משמאל מורכב מההפרש * - * כאשר * הוא ריבוע מרחק מהלנוביס בין x לבין μ_0 ו* הוא ריבוע מרחק מהלנוביס בין x לבין μ_1.

X לבין μ_0 ; * הוא ריבוע מרחק מהלנוביס בין X לבין μ_1.

לכן כלל ההחלטה השקול (והפשוט יותר) יהיה:

$$h(x)=\begin{cases}1, & \left(D_M(X,\mu_0)\right)^2-\left(D_M(X,\mu_1)\right)^2 > 2\ln\left(\frac{1-p}{p}\right)\\ 0, & \text{else}\end{cases}$$

(c) The decision boundary for this problem is defined as the set of points for which $\mathbb{P}[y = 1|\mathbf{X}] = \mathbb{P}[y = 0|\mathbf{X}]$. What is the shape of the decision a general $d > 1$ (think of $d = 1$ and $d = 2$ for intuition)?

נמשיך את השיוויונים מסעיף a הקודם:

$$\Pr(Y=1|X)=\Pr(Y=0|X) \iff \underbrace{(X-\mu_0)^t\Sigma^{-1}(X-\mu_0)-(X-\mu_1)^t\Sigma^{-1}(X-\mu_1)}_{*} = 2\ln\left(\frac{1-p}{p}\right)$$

נפשט את *:

$$* = \sum_{i=1}^{d}\frac{(X_i-\mu_{0_i})^2}{\Sigma_{ii}} - \sum_{i=1}^{d}\frac{(X_i-\mu_{1_i})^2}{\Sigma_{ii}} = \sum_{i=1}^{d}\frac{(X_i-\mu_{0_i})^2-(X_i-\mu_{1_i})^2}{\Sigma_{ii}}$$

$$= \sum_{i=1}^{d}\frac{(X_i^2-2X_i\mu_{0_i}+\mu_{0_i}^2)-(X_i^2-2X_i\mu_{1_i}+\mu_{1_i}^2)}{\Sigma_{ii}} = \sum_{i=1}^{d}\frac{\cancel{X_i^2}-2X_i\mu_{0_i}+\mu_{0_i}^2-\cancel{X_i^2}+2X_i\mu_{1_i}-\mu_{1_i}^2}{\Sigma_{ii}}$$

$$= \sum_{i=1}^{d}\frac{2X_i(\mu_{1_i}-\mu_{0_i})+\mu_{0_i}^2-\mu_{1_i}^2}{\Sigma_{ii}} = \sum_{i=1}^{d}\frac{2X_i(\mu_{1_i}-\mu_{0_i})}{\Sigma_{ii}} + \sum_{i=1}^{d}\frac{\mu_{0_i}^2-\mu_{1_i}^2}{\Sigma_{ii}}$$

$$= \sum_{i=1}^{d}\frac{2X_i(\mu_{1_i}-\mu_{0_i})}{\Sigma_{ii}} + \sum_{i=1}^{d}\frac{(\mu_{0_i}+\mu_{1_i})(\mu_{0_i}-\mu_{1_i})}{\Sigma_{ii}} = 2\cdot(\mu_1-\mu_0)^t\Sigma^{-1}\cdot X + (\mu_0+\mu_1)^t\Sigma^{-1}(\mu_0-\mu_1)$$

נחזור למשוואה המקורית:

$$\underbrace{2(\mu_1-\mu_0)^t\Sigma^{-1}}_{*}\cdot X = \underbrace{\frac{1-p}{p}-(\mu_0+\mu_1)^t\Sigma^{-1}(\mu_0-\mu_1)}_{*}$$

נשים לב שהביטוי * הוא קבוע, נסמנו c. הביטוי * הוא וקטור שורה כפול מטריצה ולכן הביטוי כולו הוא וקטור שורה, נסמנו $\bar{\alpha}$

לכן ניתן לכתוב את המשוואה הנ"ל בצורה: $\bar{\alpha}\cdot X = c$.

כלומר הוקטור $X_{1\times d}$ הוא הוקטור הפתרון למשוואה זו.

כאשר $d=1$ גבול ההחלטה הוא נקודה בודדת: $\alpha_1\cdot X_1 = c \iff X_1=\frac{c}{\alpha_1}$

כאשר $d=2$ גבול ההחלטה הוא קו ישר: $\alpha_1X_1+\alpha_2X_2=c$

כאשר $d>2$ גבול ההחלטה יהיה צורה d-מימדית.

(d) Assume now that $d = 1$, $\mu_0 = \mu_1 = \mu$ and $\sigma_0 \neq \sigma_1$. Find the decision rule whenever $f_0(x) = f(x; \mu, \sigma_0^2)$ and $f_1(x) = f(x; \mu, \sigma_1^2)$.

על סמך החישובים בסעיף ב׳ נחפש כל המקיים את השוויון:

$$\frac{f_1(x)}{f_0(x)}=\frac{1-p}{p} \iff \frac{f(x;\mu,\sigma_1^2)}{f(x;\mu,\sigma_0^2)}=\frac{1-p}{p} \iff \frac{\frac{\cancel{1}}{\sigma_1\cancel{\sqrt{2\pi}}}\cdot\exp\left(-\frac{(x-\mu)^2}{2\sigma_1^2}\right)}{\frac{\cancel{1}}{\sigma_0\cancel{\sqrt{2\pi}}}\cdot\exp\left(-\frac{(x-\mu)^2}{2\sigma_0^2}\right)}=\frac{1-p}{p}$$

$$\iff \frac{\sigma_0}{\sigma_1}\cdot\exp\left(\frac{(x-\mu)^2}{2\sigma_0^2}-\frac{(x-\mu)^2}{2\sigma_1^2}\right)=\frac{1-p}{p} \iff \exp\left(\frac{(x-\mu)^2}{2}\cdot\left(\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right)\right)=\frac{1-p}{p}\cdot\frac{\sigma_1}{\sigma_0}$$

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

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

Linear Algebra

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

   (a) Show that a symmetric matrix A is PSD if and only if it can be written as $A = XX^T$. Hint: 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.

כיוון 1: ניתן לכתוב את A כ- $X\cdot X^t$ $\Leftarrow$ A הוא PSD (כלומר מתקיים: $\forall v.\ v^t\cdot A\cdot v\geq 0$)

יהי וקטור v, אזי:

$$v^t A\cdot v=v^t\cdot X\cdot X^t\cdot v=(X^t\cdot v)^t\cdot X^t\cdot v=\|X^t\cdot v\|^2\geq 0 \quad \blacksquare$$

כיוון 2: לכל וקטור v מתקיים: $v^t A\cdot v\geq 0$ $\Leftarrow$ ניתן לכתוב את A כ- $X\cdot X^t$.

ראשית, נוכיח כי כל הע״ע של A הם אי שליליים:

$v^t\cdot A\cdot v\geq 0$ לכל וקטור v, ובפרט גם עבור $v_1,\dots,v_n$ הוקטורים העצמיים של A.

בה״כ נראה שעבור λ_1 הע״ע המתאים ל v_1, וניתן להראות באופן דומה גם עבור $\lambda_2,\dots,\lambda_n$.

מתקיים $v_1^t\cdot A\cdot v_1\geq 0$, אזי:

$$v_1^t\cdot\lambda v_1\geq 0 \iff \lambda\cdot v_1^t\cdot v_1\geq 0 \iff \lambda_1\cdot\|v_1\|^2\geq 0$$

מתקיים $v_1^t \cdot A \cdot v_1 \ge 0$, אזי:

$$v_1^t \cdot \lambda v_1 \ge 0 \iff \lambda \cdot v_1^t \cdot v_1 \ge 0 \iff \lambda \cdot \|v_1\|^2 \ge 0$$

ניתן לחלק את שני האגפים ב $\|v_1\|^2$ כיוון שהנחנו ש v_1 הוא ו"ע ולכן אינו וקטור האפס.

ונקבל $\lambda \ge 0$ כפי שרצינו.

כעת נשתמש ברמז, A היא מטריצה סימטרית ולכן ניתן להציג אותה כך: $A = Q \cdot D \cdot Q^t$

כאשר D היא מט' אלכסונית שעל אלכסונה הע"ע של A

מכיוון שכל הע"ע של A אי שליליים ניתן להציג את המט' D כך: $D = D^{1/2} \cdot D^{1/2}$ (D מט' אלכסונית).

סה"כ נקבל:

$$A = Q \cdot D \cdot Q^t = Q \cdot D^{1/2} \cdot D^{1/2} \cdot Q^t = Q \cdot D^{1/2} \cdot \left(Q \cdot (D^{1/2})^t\right)^t = Q \cdot D^{1/2} \cdot \left(Q \cdot D^{1/2}\right)^t$$

$(D^{1/2})^t = D^{1/2}$ כי $D^{1/2}$ מט' אלכסונית.

אם נסמן $X = Q \cdot D^{1/2}$ נקבל את הדרוש $A = X \cdot X^t$. $\blacksquare$

(b) Show that for all $\alpha, \beta \ge 0$ 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 $\mathbb{R}$?

A,B הן PSD, אזי לכל וקטור v מתקיים: $v^t \cdot B \cdot v \ge 0$, $v^t \cdot A \cdot v \ge 0$

כפל בסקלר אי שלילי עדיין ישמור על אי שליליות: $\beta \cdot v^t \cdot B \cdot v \ge 0$, $\alpha \cdot v^t \cdot A v^t \ge 0$

נראה שלכל וקטור v מתקיים $v^t \cdot (\alpha A + \beta B) \cdot v \ge 0$:

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

כנדרש.

בנוסף, קבוצת המטריצות מעל הממשיים שהן PSD היא לא מרחב וקטורי מעל $\mathbb{R}$

כיוון שקבוצה זו לא סגורה לכפל בסקלר:

תהי A מט' PSD ויהי סקלר $\alpha < 0$, אזי לא מתקיים לכל וקטור v: $\alpha \cdot v^t A v \ge 0$

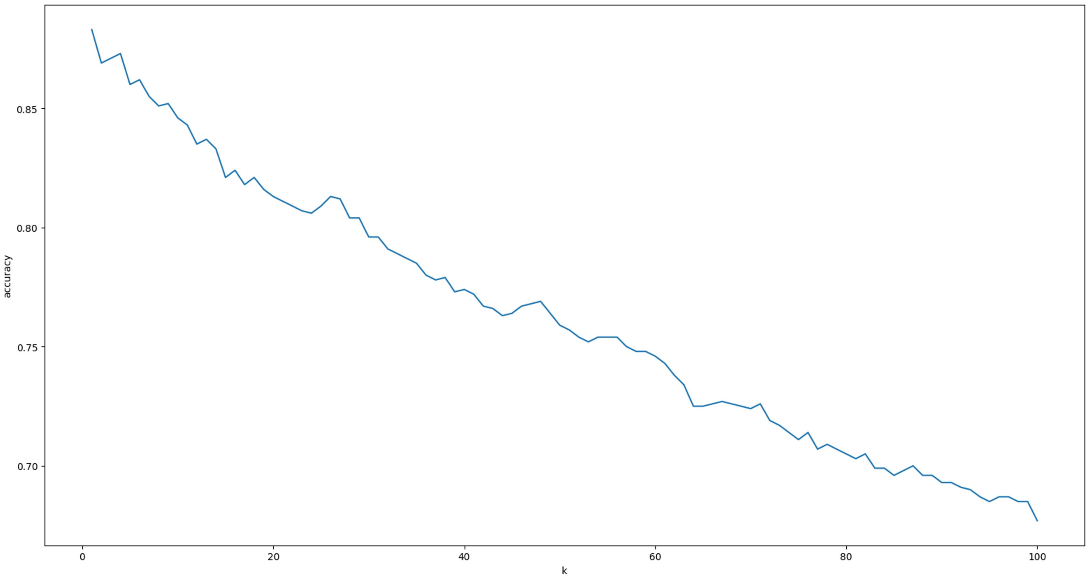
ולכן קבוצה זו אינה מרחב וקטורי.

Programming Assignment

(b) Run the algorithm using the first $n = 1000$ training images, on each of the test images, using $k = 10$. What is the accuracy of the prediction (i.e. the percentage of correct classifications)? What would you expect from a completely random predictor?

מבין 1000 התמונות שנבדקו לחיזוי ניתן לייבל נכון ל 846, כלומר אחוז דיוק של 84.6%.
אילו היינו משתמשים בפרדיקטור רנדומי שהיה נותן לכל תמונה לייבל אקראי מבין כל הלייבלים האפשריים (הספרות 0,1,...,9), בהנחה של הלייבלים מתפלגים אחיד בדיוק בין 0 ל-9 כל לייבל שהפרדיקטור היה נותן היה מדויק בסיכוי $\frac{1}{10}$ ולכן נצפה שאחוז הדיוק של פרדיקטור רנדומי יהיה 10%.

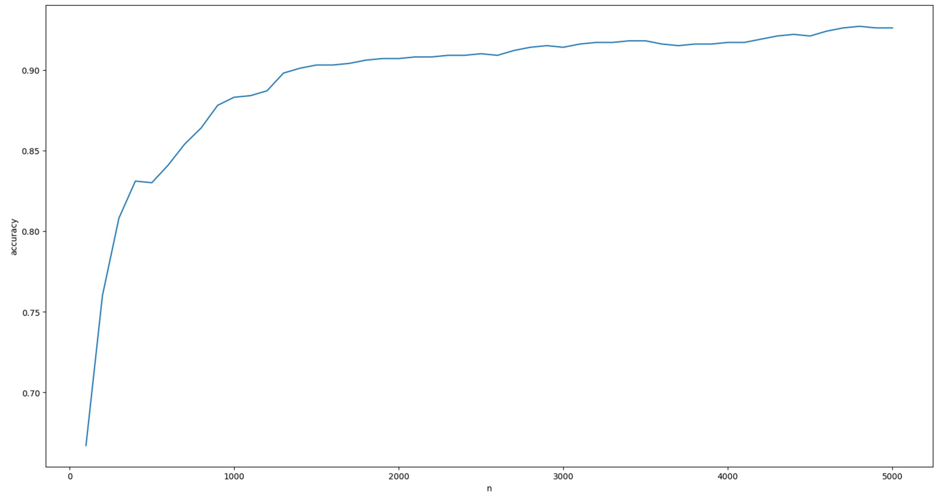
(c) Plot the prediction accuracy as a function of k, for $k = 1, \dots, 100$ and $n = 1000$. Discuss the results. What is the best k?



ניתן לראות שקיים קשר לינארי יורד בין מס' השכנים k לבין אחוז הדיוק, כלומר ככל שמספר השכנים גדל יותר אחוז הדיוק יורד. כמו כן ה-k שעבורו אחוז הדיוק הוא הגבוה ביותר הינו $k=1$.

לדעתי, הסיבה לכך היא ההגדרה שלנו לדיוק, אנו מגדירים תיוג כתיוג מדויק אם"ם התיוג שהוחזר עבור תמונה ע"י האלגוריתם זהה לתיוג האמיתי שלה. כלומר תמונה שתמסת קרובה לתמונה הנבדקת אבל שהלייבל שלה שונה מלייבל התמונה הנבדקת תיחשב כטעות בחישוב הדיוק.
ככל שנגדיל את מס' השכנים של התמונה הנבדקת נגדיל את הסיכוי לכך שהלייבל הנפוץ בקרב התמונות השכנות יהיה שונה מלייבל התמונה הנבדקת ונקבל חיזוי שגוי עבור אותה תמונה. לעומת זאת אם ניקח מס' קטן של שכנים, המרחקים שלהם מהתמונה הנבדקת יהיו יחסית קטנים יותר (אפשר לחשוב על ממוצע המרחקים למשל) וכך יש יותר סיכוי שהלייבל השכיח יהיה הלייבל של התמונה הנבדקת, ואחוז הדיוק יהיה גדול יותר.

(d) Using $k = 1$, run the algorithm on an increasing number of training images. Plot the prediction accuracy as a function of $n = 100, 200, \ldots, 5000$. Discuss the results.



ניתן לראות שיש קשר עולה בין גודל קבוצת האימון לאחוז הדיוק.
במודל זה לוקחים שכן אחד ($k=1$) כלומר את התמונה שהכי קרובה לתמונה הנבדקת. ככל שמגדילים את קבוצת האימון, יש יותר סיכוי שתימצא תמונה שמאוד קרובה לתמונה הנבדקת והסיכוי שהלייבל שלה יהיה זהה ללייבל של התמונה הנבדקת ולכן הסיווג יהיה נכון.