100

מבוא ללמידה חישובית, תרגיל 3

אורי ברגר 200791275 uriberger@mail.tau.ac.il
אופיר גרין 305625048 ofirgreen@mail.tau.ac.il

שאלות תיאורטיות

שאלה 1

א) נעבור על כל הדוגמאות. אם ניתקל בדוגמה (x,y) עבור $y=1$, נחזיר את h_x. אם נסיים את כל הדוגמאות ולא ניתקל בדוגמה עם סימון 1, נחזיר את $h_$.
נוכיח שאלגוריתם זה מביא למינימום את הסיכון האמפירי:
יהי $EX(c,D)$ אוסף נתונים, ונסמן את הפונקציה שהאלגוריתם החזיר ב- h'. נראה שלכל $(x,y)\in EX(c,D)$, מתקיים ש- $h'(x)=c(x)$. נחלק למקרים:
$\boldsymbol{h'\neq h_}$: אז בהכרח קיים $z\in X$ כך ש- $h'=h_z$, לכן $h'(z)=1$. מכיוון שהחזרנו את h', בהכרח $(z,1)\in EX(c,D)$. הפונקציה היחידה שנותנת ל- z את הערך 1, לכן בוודאות $c=h'$. בפרט, לכל $(x,y)\in EX(c,D)$, מתקיים ש- $h'(x)=c(x)$.
$\boldsymbol{h'=h_}$: אז כל הדוגמאות קיבלו סימון של -1. לכן בהכרח, לכל $(x,y)\in EX(c,D)$, מתקיים ש- $c(x)=-1$. מהגדרת $h_$, לכל x כנ"ל $h_(x)=-1$. לכן לכל $(x,y)\in EX(c,D)$, יתקיים ש- $h_(x)=c(x)$.

וכעת:

$$\widehat{error}(h')=\frac{1}{m}\sum_{i=1}^{m} I\left(h'(x_i)\neq c_t(x_i)\right)=0$$

ומכיוון שלכל פונקציה $h\in H$, $\widehat{error}(h)\geq 0$, לכל פונקציה $h\in H$ מתקיים $\widehat{error}(h)\geq\widehat{error}(h')$, ואכן האלגוריתם מביא למינימום את הסיכון האמפירי.

ב) תהי c פונקציית קונספט, תהי D התפלגות על X, ויהיו $\varepsilon,\delta>0$. נבחר: $m:=\log_{1-\varepsilon}\delta$, ונבצע את האלגוריתם מסעיף א' עם m דגימות. נראה שבהסתברות לפחות $1-\delta$ אנו מקבלים טעות קטנה מ- ε. נסמן את הפונקציה שקיבלנו ב- h'.
נחלק למקרים:
$\boldsymbol{c=h_}$: ואז לכל $(x,y)\in EX(c,D)$ יתקיים $y=-1$, ולכן בוודאות נבחר $h'=h_$. במקרה זה בהסתברות 1 אנו מקבלים טעות 0.
$\boldsymbol{c\neq h_}$: אז קיים $z\in X$ כך ש- $c=h_z$. נשים לב שהמקרה היחיד בו ננבא פונקציה שהיא לא c, הוא אם נבחר $h'=h_$ (אנו בוחרים פונקציה h_x רק במקרה שראינו במדגם $(x,1)$, ו- c לא מסמנת ב- 1 אף $x\neq z$). לכן האיבר היחיד ב- X ש- h',c יכולות לא להסכים עליו הוא z. כעת, אם נסמן: $p:=D(z)$, נקבל:

$$error(h')=\Pr\left(h'(x)\neq c_t(x)\right)\leq p$$

נחלק למקרים:
1) $\boldsymbol{p<\varepsilon}$: ואז בהסתברות 1, $error(h')\leq p<\varepsilon$, כרצוי.
2) $\boldsymbol{p\geq\varepsilon}$: ההסתברות שבדגימה אחת לא ייצא לנו z היא $1-p$. לכן ההסתברות שב- m דגימות בלתי תלויות לא ייצא לנו z היא

$$(1-p)^m=(1-p)^{\log_{1-\varepsilon}\delta}\leq(1-\varepsilon)^{\log_{1-\varepsilon}\delta}=\delta$$

כאשר אי השוויון הוא בגלל ש- $p\geq\varepsilon$. כלומר בהסתברות לפחות $1-\delta$ נקבל באחת הדגימות את z. במקרה זה בטוח $h'=h_z=c$ (לא יכול להיות שעצרנו את האלגוריתם לפני כי אנחנו עוצרים אותו רק כשאנחנו רואים 1, וכל שאר הדוגמאות מסומנות על ידי c ב- -1), והשגיאה שנקבל היא 0.

ראינו שבכל המקרים אנו מקבלים שגיאה של פחות מ- ε בהסתברות לפחות $1-\delta$, כרצוי.

<u>שאלה 2</u>

נציג אלגוריתם שלומד "pac" את המחלקה.
אינטואיטיבית: נעבור על כל הדוגמאות החיוביות ונצייר מעגל שהרדיוס שלו הוא המרחק בין ראשית הצירים לבין הדוגמא החיובית הרחוקה ביותר מראשית הצירים.
פורמאלית: נניח שהקונספט הנכון הוא c_r, שההתפלגות היא D ונניח שקיבלנו קבוצת דוגמאות $EX(c_r,D)$.
נסמן: $A=\{(x,y)\in\mathbb{R}^2 | ((x,y),1)\in EX(c_r,D)\}$, כלומר קבוצת כל הדוגמאות שסווגו חיובית. נבחר: $r'=\max_{(x,y)\in A}\{\sqrt{x^2+y^2}\}$, כלומר הנקודה החיובית הרחוקה ביותר מראשית הצירים. כעת נסמן:

$$h(x,y)=\begin{cases}1, & x^2+y^2\le (r')^2\\ 0, & else\end{cases}$$

נוכיח שהאלגוריתם לומד "pac" את המחלקה עבור ε,δ עם $m\ge\left(\frac{1}{\varepsilon}\right)\log\left(\frac{1}{\delta}\right)$:
נשים לב שהעיגול האמיתי לא יכול להיות קטן יותר (כי בחרנו את העיגול לפי הדוגמה החיובית הרחוקה ביותר, ואם העיגול האמיתי קטן יותר הוא היה מסווג אותה כשלילית). לכן $r\ge r'$. נסמן $B=\{(x,y)\in\mathbb{R}^2 | r'<x^2+y^2\le r\}$ (כלומר קבוצת הנקודות שיוצאות בעיגול האמיתי ולא בעיגול שלנו), ובנוסף:

$$p:=P_{(x,y)\sim D}(r'<x^2+y^2\le r)$$

כלומר, ההסתברות שנקודה יוצאת בעיגול האמיתי ולא בעיגול שלנו. נחלק למקרים:
$\boldsymbol{p<\varepsilon}$: ואז הטעות שלנו קטנה מ- ε בהסתברות 1, כרצוי.
$\boldsymbol{p\ge\varepsilon}$: ואז הטעות שלנו גדולה מ- ε. בשביל שנקבל את הרדיוס r' בסוף האלגוריתם, כל הדוגמאות שמגיעות צריכות להיות שליליות (כלומר מחוץ למעגל האמיתי) או חיוביות אבל בתוך המעגל של r', כלומר לא יקיימו $r'<x^2+y^2\le r$. מהגדרת p, ההסתברות של דוגמה אחת ליפול בשטח זה היא $1-p$. לכן, ההסתברות של m דוגמאות ליפול בשטח זה היא (מכיוון שהדוגמאות בלתי תלויות):

$$P\big((x_1,y_1)\notin B,\dots,(x_m,y_m)\notin B\big)=\prod_{i=1}^{m}P\big((x_i,y_i)\notin B\big)=\prod_{i=1}^{m}1-p=(1-p)^m=$$

$$(1-p)^{\left(\frac{1}{\varepsilon}\right)\log\left(\frac{1}{\delta}\right)}\underset{\text{כי } p\ge e}{\le}(1-\varepsilon)^{\left(\frac{1}{\varepsilon}\right)\log\left(\frac{1}{\delta}\right)}\underset{\text{מפיתוח טיילור}}{\le}(e^{-\varepsilon})^{\left(\frac{1}{\varepsilon}\right)\log\left(\frac{1}{\delta}\right)}=e^{-\log\left(\frac{1}{\delta}\right)}=e^{\log\left(\left(\frac{1}{\delta}\right)^{-1}\right)}=e^{\log\delta}=\delta$$

<u>שאלה 3</u>

יהי $M\in\mathbb{N}$. נסתכל על וקטורים מעל $\mathbb{R}^M$. נבחר את הקו $w^*=\left(-\frac{1}{\sqrt{M}},-\frac{1}{\sqrt{M}},\dots,-\frac{1}{\sqrt{M}}\right)$ (נשים לב שהוא עם נורמה אחת), ונסתכל על הדוגמאות מהצורה:

$$x_i=(\underbrace{0,\dots,0}_{i-1\text{ פעמים}},1,0,\dots,0)$$

(נשים לב שהנורמה של כל דוגמה היא אכן אחת). ראשית נראה שהאלגוריתם יעשה M טעויות. נראה באינדוקציה שלכל $i\in\{1,\dots,M\}$, כשהאלגוריתם רואה את הדוגמא x_i, $w_i=(-1,\dots,-1,0,\dots,0)$ ($i-1$ פעמים -1).
<u>בסיס</u>: עבור $i=1$, אנו מתחילים את האלגוריתם עם $w_1=(0,\dots,0)$, לכן הטענה נכונה.

הנחה: נניח נכונות עבור i.
צעד: נוכיח עבור $i+1$.
מהנחת האינדוקציה, $w_i = (-1, \dots, -1, 0, \dots, 0)$. אז מגיע x_i, ומכיוון ש- $w_i^{(i)} = 0$ נקבל $w_i^T x_i = 0$, ולכן אנחנו מסווגים את x_i כחיובי. אבל $w^{*T} x_i = -\frac{1}{\sqrt{M}}$, לכן אנחנו טועים ומבצעים $w_{i+1} = w_i - x_i = (-1, \dots, -1, -1, 0, \dots, 0)$, עם i פעמים -1, כרצוי.

במהלך האינדוקציה הראנו שבכל שלב האלגוריתם טועה, לכן אחרי M שלבים יהיו לו M טעויות.

כעת נראה ש- $M = \frac{1}{\gamma^2}$. עבור כל $i \in \{1, \dots, M\}$, $\frac{|w^* \cdot x_i|}{||x_i||} = |w^* \cdot x_i| = \left|-\frac{1}{\sqrt{M}}\right| = \frac{1}{\sqrt{M}}$. לכן, לפי הגדרת השוליים, $\gamma = \frac{1}{\sqrt{M}}$. לכן:

$$\frac{1}{\gamma^2} = \frac{1}{\left(\frac{1}{\sqrt{M}}\right)^2} = M$$

שאלה 4

א) נגדיר את הבעיה בצורה הבאה:

$$\min_{w,b} \frac{1}{2} w^T w$$
$$s.t.$$
$$\forall n = 1, \dots, n: \begin{cases} w^T x_n + b - y_n \le \varepsilon \\ y_n - (w^T x_n + b) \le \varepsilon \end{cases}$$

ב) נעביר אגפים באילוצים ונקבל לכל $n = 1, \dots, N$:

$$y_n - (w^T x_n + b) + \varepsilon \ge 0$$
$$w^T x_n + b - y_n + \varepsilon \ge 0$$

ולכן פונקציית הלגרנז' תיראה כך:

$$L(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{n=1}^{N} \alpha_n (y_n - (w^T x_n + b) + \varepsilon) - \sum_{n=N+1}^{2N} \alpha_n \left(w^T x_{(n-N)} + b - y_{(n-N)} + \varepsilon\right)$$

ג) נגזור לפי w:

$$\nabla_w L = w - \sum_{n=1}^{N} -a_n x_n - \sum_{n=N+1}^{2N} a_n x_{n-N} = w + \sum_{n=1}^{N} a_n x_n - \sum_{n=N+1}^{2N} a_n x_{n-N} = 0$$

לכן נקבל:

$$w = \sum_{n=N+1}^{2N} a_n x_{n-N} - \sum_{n=1}^{N} a_n x_n = \sum_{n=1}^{N} x_n (\alpha_{n+N} - \alpha_n)$$

נגזור לפי b:

$$\frac{\partial}{\partial b} L = -\sum_{n=1}^{N} -\alpha_n - \sum_{n=N+1}^{2N} \alpha_n = \sum_{n=1}^{N} \alpha_n - \sum_{n=N+1}^{2N} \alpha_n = 0$$

לכן נקבל:

$$\sum_{n=1}^{N}\alpha_n - \sum_{n=N+1}^{2N}\alpha_n = \sum_{n=1}^{N}(\alpha_n - \alpha_{n+N}) = 0$$

ד) בשביל למצוא את הבעיה הדואלית נציב את הערכים שמצאנו עבור w, b בפונקציית הלגרנז'. נקבל:

$$L(\alpha) = \frac{1}{2}w^Tw - \sum_{n=1}^{N}\alpha_n(y_n - (w^Tx_n + b) + \varepsilon) - \sum_{n=N+1}^{2N}\alpha_n\left(w^Tx_{(n-N)} + b - y_{(n-N)} + \varepsilon\right) =$$

$$\frac{1}{2}w^Tw - \varepsilon\left(\sum_{n=1}^{N}\alpha_n + \sum_{n=N+1}^{2N}\alpha_n\right) - \sum_{n=1}^{N}\alpha_n\left(y_n - (w^Tx_n + b)\right) + \sum_{n=N+1}^{2N}\alpha_n\left(y_{(n-N)} - \left(w^Tx_{(n-N)} + b\right)\right) =$$

$$\frac{1}{2}w^Tw - \varepsilon\left(\sum_{n=1}^{N}\alpha_n + \sum_{n=N+1}^{2N}\alpha_n\right) + \sum_{n=1}^{N}\left(y_n - (w^Tx_n + b)\right)(\alpha_{n+N} - \alpha_n) =$$

$$\frac{1}{2}w^Tw - \varepsilon\left(\sum_{n=1}^{N}\alpha_n + \sum_{n=N+1}^{2N}\alpha_n\right) + \sum_{n=1}^{N}(y_n - w^Tx_n)(\alpha_{n+N} - \alpha_n) - b\sum_{n=1}^{N}(\alpha_n - \alpha_{n+N}) =$$

שווים לפי הגזירה של b

שווה לאפס לפי הגזירה של b

$$\frac{1}{2}w^Tw - 2\varepsilon\sum_{n=1}^{N}\alpha_n + \sum_{n=1}^{N}(y_n - w^Tx_n)(\alpha_{n+N} - \alpha_n) =$$

$$\frac{1}{2}w^Tw - 2\varepsilon\sum_{n=1}^{N}\alpha_n + \sum_{n=1}^{N}y_n(\alpha_{n+N} - \alpha_n) - w^T\sum_{n=1}^{N}x_n(\alpha_{n+N} - \alpha_n) =$$

שווה ל- w לפי הגזירה של w

$$\frac{1}{2}w^Tw - 2\varepsilon\sum_{n=1}^{N}\alpha_n + \sum_{n=1}^{N}y_n(\alpha_{n+N} - \alpha_n) - w^Tw =$$

$$-\frac{1}{2}w^Tw - 2\varepsilon\sum_{n=1}^{N}\alpha_n + \sum_{n=1}^{N}y_n(\alpha_{n+N} - \alpha_n) =$$

נציב את הערך שמצאנו של w

$$-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(a_i - a_{i+N})(a_j - a_{j+N})x_i^Tx_j + 2\varepsilon\sum_{n=1}^{N}\alpha_n + \sum_{n=1}^{N}y_n(\alpha_{n+N} - \alpha_n)$$

כעת אם נגדיר $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_N)$, $\alpha' = (\alpha_{N+1}, \dots, \alpha_{2N})$, נקבל את התוכנית:

$$\min_{\alpha}\frac{1}{2}(\alpha - \alpha')^TM(\alpha - \alpha') - (2\varepsilon^T)\alpha - (y^T)(\alpha' - \alpha)$$

$$s.t.\ (1^T)\cdot\alpha + (-1^T)\cdot\alpha' = 0$$

$$\alpha \geq 0$$

$$M = \begin{pmatrix} x_1^Tx_1 & \cdots & x_1^Tx_N \\ \vdots & \ddots & \vdots \\ x_N^Tx_1 & \cdots & x_N^Tx_N \end{pmatrix}$$

<u>שאלה 5</u>

א) הפונקציה אינה בהכרח proper kernel. נראה דוגמה נגדית:
נסמן $K_1(x,z)=0$ לכל $x,z\in\mathbb{R}^d$, $K_2(x,z)=1$ לכל $x,z\in\mathbb{R}^d$. נראה שהם proper kernels:
1) סימטריות:

$$K_1(x,z)=0=K_1(z,x)$$
$$K_2(x,z)=1=K_2(z,x)$$

2) חיוביות: לכל N, לכל $x_1,\dots,x_N\in\mathbb{R}^d$ ולכל $y\in\mathbb{R}^N$:
מהגדרת K_1, $\overline{\overline{K_1}}=0$, לכן $y^T\overline{\overline{K_1}}y=0\geq 0$.
מהגדרת K_2, $\overline{\overline{K_2}}$ היא מטריצה שכל האיברים שלה הם 1. לכן:

$$y^T\overline{\overline{K_2}}y=\sum_{i=1}^{N}\left(y_i\sum_{j=1}^{N}y_j\overline{\overline{K_2}}_{(i,j)}\right)=\sum_{i=1}^{N}\left(y_i\sum_{j=1}^{N}y_j\right)=\left(\sum_{j=1}^{N}y_j\right)\cdot\left(\sum_{i=1}^{N}y_i\right)\geq 0$$

הסכום הפנימי לא תלוי ב- i — כל מכפלה של מספר בעצמו היא אי שלילית

כעת נראה ש- K אינה חיובית: מעל $\mathbb{R}$, נבחר $N=1$, $x_1=1$. אזי:

$$\overline{\overline{K}}=\overline{\overline{K_1}}-\overline{\overline{K_2}}=0-1=-1$$

נבחר $y=1$. וכעת:

$$y^t\overline{\overline{K}}y=1\cdot(-1)\cdot 1=-1<0$$

ב) הפונקציה אינה בהכרח מוגדרת: אם נבחר את K_2 להיות זהה ל- K_1 מסעיף א' (כלומר תמיד מחזירה 0), הפונקציה אינה מוגדרת.

ג) הפונקציה אכן proper kernel. נוכיח:
1) סימטריות: יהיו $x,z\in\mathbb{R}^d$. אזי:

$$K(x,z)=aK_1(x,z)=aK_1(z,x)=K(z,x)$$

כאשר השוויון האמצעי הוא בגלל ש- K_1 היא proper kernel.
2) חיוביות: מהגדרת K, לכל $x_1,\dots,x_N$ מתקיים ש- $\overline{\overline{K}}=a\overline{\overline{K_1}}$. לכן, לכל y:

$$y^T\overline{\overline{K}}y=y^Ta\overline{\overline{K_1}}y=a\left(y^T\overline{\overline{K_1}}y\right)$$

ומה שקיבלנו אכן גדול מאפס כי נתון ש- $a>0$, ו- K_1 היא proper kernel.

ד) הפונקציה אינה בהכרח proper kernel. נראה דוגמה נגדית:
נבחר $K_1(x,z)=1$ לכל $x,z\in\mathbb{R}^d$. נראה ש- $K(x,z)=-aK_1(x,z)$ לא חיובית. מעל $\mathbb{R}$ נבחר $N=1$, $x_1=1$. אזי:

$$\overline{\overline{K}}=\overline{\overline{-aK_1}}=-a\overline{\overline{K_1}}=-a\cdot 1=-a$$

נבחר $y=1$. וכעת:

$$y^t\overline{\overline{K}}y=1\cdot(-a)\cdot 1=-a<0$$

ה) הפונקציה אכן proper kernel. נוכיח:
1) סימטריות: יהיו $x,z\in\mathbb{R}^d$. אזי:

$$K(x,z)=f(x)f(z)=f(z)f(x)=K(z,x)$$

כאשר השוויון האמצעי הוא בגלל קומוטטיביות של כפל ממשיים (נתון ש- f ממשית).
2) חיוביות: יהיו $x_1,\dots,x_N\in\mathbb{R}^d$, $y\in\mathbb{R}^N$. נראה שכל הערכים העצמיים של $\overline{\overline{K}}$ אי שליליים.
נניח בשלילה שקיים ערך עצמי λ של $\overline{\overline{K}}$ כך ש- $\lambda<0$. נשים לב שהעמודה ה- i של $\overline{\overline{K}}$ היא מהצורה:

$$\begin{pmatrix} f(x_1)f(x_i) \\ f(x_2)f(x_i) \\ \dots \\ f(x_N)f(x_i) \end{pmatrix}$$

כלומר היא מהצורה $\begin{pmatrix} f(x_1) \\ f(x_2) \\ \dots \\ f(x_N) \end{pmatrix} f(x_i)$. לכן כל עמודה היא כפולה של $\begin{pmatrix} f(x_1) \\ f(x_2) \\ \dots \\ f(x_N) \end{pmatrix}$, והעמודות תלויות לינארית בזוגות. בהכרח קיים $i \in \{1, \dots, N\}$ כך ש- $f(x_i) \neq 0$ (אחרת $\overline{\overline{K}} = 0$ ונקבל סתירה לכך ש- $\lambda < 0$), לכן הדרגה של המטריצה היא 1, ולכן יש לה $N - 1$ פעמים את הערך העצמי אפס (ופעם אחת את λ).
כעת, נשים לב שעקבת המטריצה היא: $\sum_{i=1}^{N} f(x_i)f(x_i) \geq 0$ (אי שלילי כי כל איבר בסכום אי שלילי). מצד שני, עקבת מטריצה שווה לסכום הערכים העצמיים שלה, והראנו שהערכים העצמיים הם כולם 0, וערך עצמי יחיד שערכו λ. לכן, סכום הערכים העצמיים הוא $\lambda < 0$, בסתירה לכך שהוא שווה לעקבה (שהיא אי שלילית).

ו) הפונקציה אכן proper kernel. נוכיח:
יהי $p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \dots + \alpha_n x^n$ פולינום כך שלכל $i \in \{0, \dots, n\}$, $\alpha_i \geq 0$. נוכיח ש- $p(K_1(x,z))$ הוא proper kernel. נראה באינדוקציה על n.
<u>בסיס</u>: $n = 0$: אזי קיים $a \geq 0$ כך ש- $p(x) = a$, לכן $p(K_1(x,z)) = aK_1(x,z)$. בסעיף ג' ראינו שזה proper kernel.
<u>הנחה</u>: נניח נכונות עבור n.
<u>צעד</u>: נוכיח עבור $n + 1$.
נחלק את הפולינום בצורה הבאה: נסמן:

$$q(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \dots + a_{n-1}x^{n-1}, r(x) = a_n x^n$$

אזי:

$$p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \dots + \alpha_n x^n = q(x) + r(x)$$

ראשית, מהנחת האינדוקציה, $q(K_1(x))$ הוא proper kernel. נראה בהמשך ש- $r(K_1(x))$ הוא proper kernel. כעת נוכל להשתמש בטענה שראינו בשיעור, שסכום של proper kernels הוא proper kernel, ולקבל את הרצוי.

נותר להוכיח שלכל n ולכל $\alpha \geq 0$, $\alpha R_1(x,z)^n$ הוא proper kernel. יהי $\alpha \geq 0$. נראה, שוב, באינדוקציה על n.
<u>בסיס</u>: $n = 0$: זהה לבסיס של האינדוקציה הקודמת.
<u>הנחה</u>: נניח נכונות עבור n.
<u>צעד</u>: נוכיח עבור $n + 1$.

$$\alpha R_1(x,z)^{n+1} = (\alpha R_1(x,z)^n) \cdot R_1(x,z)$$

כאשר צד ימין של השוויון הוא proper kernel לפי טענה שראינו בכיתה- שמכפלה של proper kernels היא proper kernel (הכופל השמאלי proper kernel לפי הנחת האינדוקציה והכופל הימני proper kernel לפי הנחות השאלה).