תעודת זהות:
מספר מחברת:

מבוא ללמידה חישובית – מבחן מועד ב׳
סמסטר א׳ תשע״ו (2015/6)

בית הספר למדעי המחשב, אוניברסיטת תל־אביב

מרצים: פרופ׳ ליאור וולף, פרופ׳ ערן הלפרין
מתרגל: רגב שוייגר

8.4.2016

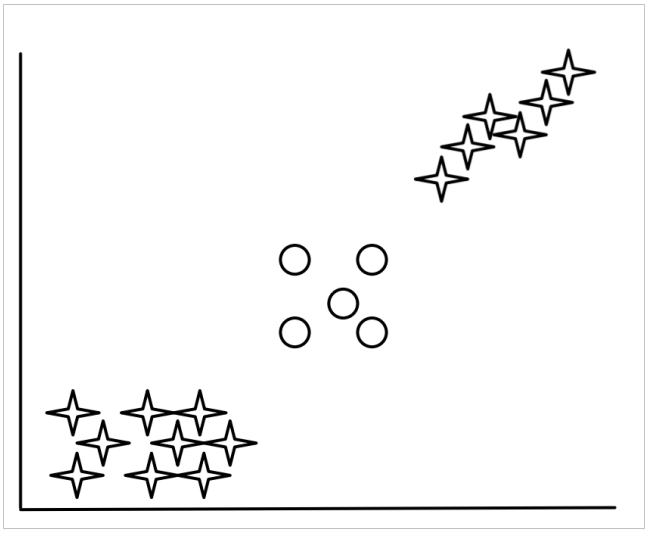
הוראות:

1. מומלץ לקרוא את כל ההנחיות והשאלות בתחילת המבחן לפני תחילת כתיבת התשובות.
2. משך הבחינה – **שלוש שעות**.
3. חומר עזר מותר: דף נוסחאות בגודל A4.
4. **יש לענות על השאלות במקום המיועד לכך בטופס השאלון (טופס זה)**. מחברות הבחינה לא תקראנה, ותשמשנה כטיוטה בלבד.
5. יש למלא בכל דף של השאלון מספר ת.ז. ומספר מחברת.
6. במבחן 5 שאלות:
   - הניקוד לכל שאלה מופיע ליד מספר השאלה.
   - יש לענות תשובות ברורות, ענייניות ותמציתיות.
7. מותר להשתמש בכל טענה שהוכחה בכיתה (בהרצאה, בתרגול, או בתרגיל בית) בתנאי שמצטטים אותה במדויק. טענות אחרות (כאלה שהוכחו בספר, בהרצאות מהסמסטר הקודם, וכו׳) יש להוכיח.
8. אם לא נאמר אחרת, יש להניח שדגימות במדגם נוצרות באופן בלתי תלוי ומאותה התפלגות (i.i.d).

תעודת זהות:
מספר מחברת:

1 שאלה 1 - 20 נקודות

נתון המדגם הדו־מימדי הבא, בו הנקודות מסווגות לשתי מחלקות:



עבור כל אחד מהאלגוריתמים הבאים, קבע/י האם ניתן להריץ אותו עד לקבלת מסווג עם שגיאת למידה אפס, על המדגם הנתון. אם כן, צייר/י קו הפרדה מתאים למסווג המתקבל. אם לא, הסבר/הסבירי מדוע.

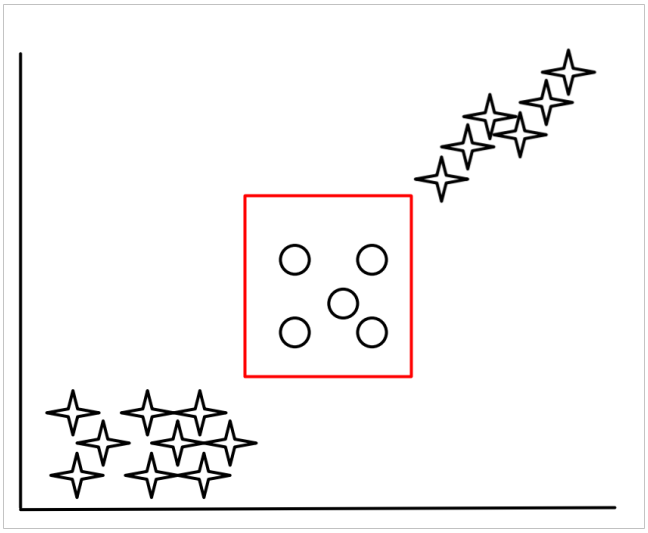
תעודת זהות:
מספר מחברת:

1. אלגוריתם המוצא מלבן דו־מימדי המוגדר על ידי ארבעה מספרים (x_1, x_2, y_1, y_2).
כלומר, נקודה (x, y) תסווג עם סיווג מסוים אם $x_1 \leq x \leq x_2$ וגם $y_1 \leq y \leq y_2$,
ועם הסיווג השני אחרת.

☐ לא יכול להגיע לשגיאה אפס. הסבר:

☐ יכול להגיע לשגיאה אפס. קו ההפרדה של המסווג:

יכול. ניתן בקלות למצוא מלבן המקיף את הנקודות המסווגות כעיגול.

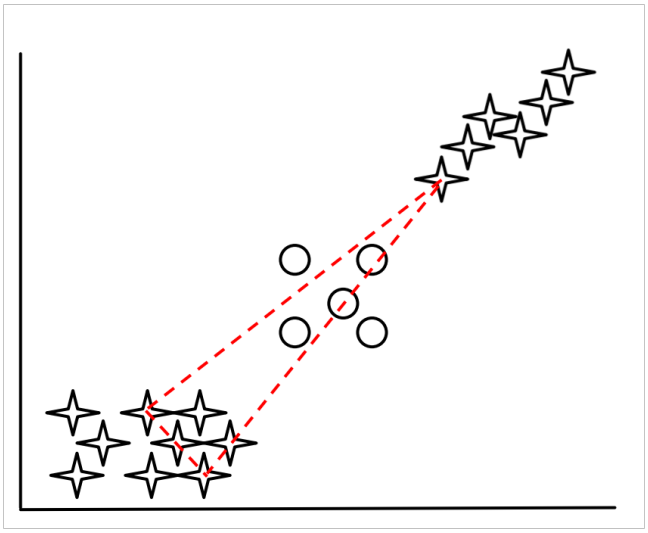


2. רגרסיה לוגיסטית, כאשר הסיווג נקבע לפי העיגול של התוצאה ל־0 או ל־1.

☐ לא יכול להגיע לשגיאה אפס. הסבר:

☐ יכול להגיע לשגיאה אפס. קו ההפרדה של המסווג:

לא יכול. רגרסיה לוגיסטית מעוגלת היא דוגמה לאלגוריתם שמחזיר מסווג לינארי. אבל הדגימה לא ניתנת להפרדה על ידי מסווג לינארי. ניתן לראות זאת על ידי כך שיש נקודה המסווגת כעיגול בצירוף הקמור של שלוש נקודות המסווגות ככוכב. לו היה ניתן להפריד לינארית את המדגם, כל מה שבקמור של נקודות מסווג באותו האופן.

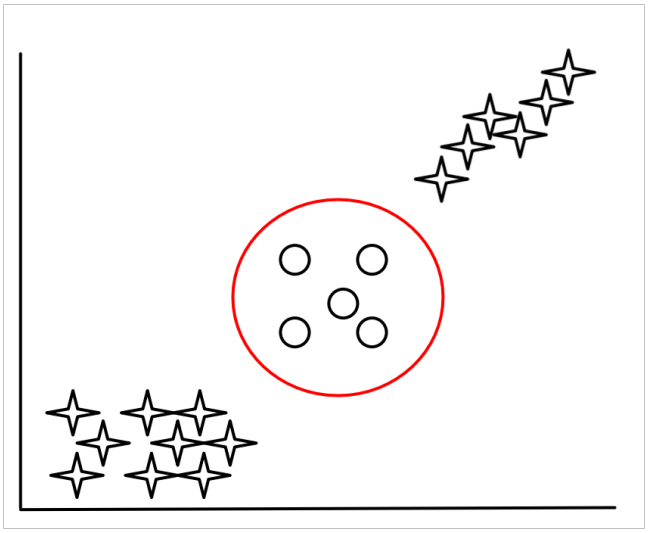


3. SVM עם Polynomial kernel ריבועי.

☐ לא יכול להגיע לשגיאה אפס. הסבר:

☐ יכול להגיע לשגיאה אפס. קו ההפרדה של המסווג:

יכול. ניתן בקלות למצוא עיגול או אליפסה המקיף את הנקודות המסווגות כעיגול.



שאלה 2 - 20 נקודות

נתון לנו מדגם מסווג $(\mathbf{x}_i, y_i)$, כאשר $\mathbf{x}_i \in \mathbb{R}^2$ ו- $y_i \in \{+1, -1\}$. נרצה לפתור את בעיית הרגרסיה הבאה:

$$y_i = w^2(\mathbf{x}_i)_1 + (\mathbf{x}_i)_2$$

1. מהי פונקציית האופטימיזציה המתאימה לבעיה החדשה?

נרצה להביא למינימום את סכום פונקציות ה-loss של כל הנקודות:

$$w_{opt} = \arg\min_{w \in \mathbb{R}} \sum_{i=1}^{n} (w^2(\mathbf{x}_i)_1 + (\mathbf{x}_i)_2 - y_i)^2$$

2. מהו ה־w שמביא למינימום את פונקציית המטרה שהוגדרה בסעיף 1?

אם נסמן $W = w^2$, ניתן לראות כי פונקציית האופטימיזציה היא משוואה ריבועית ב־W:

$$\begin{aligned}
&\sum_{i=1}^{n}(W(\mathbf{x}_i)_1 + (\mathbf{x}_i)_2 - y_i)^2 \\
=&\sum_{i=1}^{n} W^2(\mathbf{x}_i)_1^2 + 2W(\mathbf{x}_i)_1((\mathbf{x}_i)_2 - y_i) + ((\mathbf{x}_i)_2 - y_i)^2 \\
=&W^2\sum_{i=1}^{n}(\mathbf{x}_i)_1^2 + W\sum_{i=1}^{n}2(\mathbf{x}_i)_1((\mathbf{x}_i)_2 - y_i) + \sum_{i=1}^{n}((\mathbf{x}_i)_2 - y_i)^2
\end{aligned}$$

אם לכל i, $(\mathbf{x}_i)_1 = 0$, אז ברור כי ל־w אין השפעה על האופטימיזציה ואפשר לבחור כל ערך. אחרת, נקבל כי המקדם של W^2 הוא חיובי, ופונקציית האופט' היא פרבולה. המינימום של פרבולה $f(x) = ax^2 + b + c$ עם $a > 0$ מתקבל בנקודה $x = -b/2a$, ובמקרה זה הוא:

$$\frac{-\sum_{i=1}^{n}2(\mathbf{x}_i)_1((\mathbf{x}_i)_2 - y_i)}{2\sum_{i=1}^{n}(\mathbf{x}_i)_1^2} = \frac{\sum_{i=1}^{n}(\mathbf{x}_i)_1(y_i - (\mathbf{x}_i)_2)}{\sum_{i=1}^{n}(\mathbf{x}_i)_1^2}$$

עלינו להתחשב בכך ש־W חייב להיות אי־שלילי. כלומר, אם המינימום מתקבל בנקודה שלילית, אז המינימום תחת האילוץ הוא 0. כלומר בסך הכל,

$$w_{opt} = \max\left(0, \sqrt{\frac{\sum_{i=1}^{n}(\mathbf{x}_i)_1(y_i - (\mathbf{x}_i)_2)}{\sum_{i=1}^{n}(\mathbf{x}_i)_1^2}}\right)$$

ניתן להגיע לאותה המסקנה באמצעות פיתוח וגזירה מפורשים של המשוואה.

שאלה 3 - 20 נקודות

האלגוריתם שראינו בכיתה ל־Soft Margin SVM מניח שלכל הנקודות יש אותה חשיבות בעינינו. בשאלה זו, נניח כי לכל נקודה $\mathbf{x}_n$ נתונה משקולת $0 \leq v_n \leq 1$ שמגדירה את החשיבות שלה. נרצה כעת להביא למינימום את סכום החריגות מהשוליים, המשוקלל לפי החשיבות. כלומר, נתונה הבעיה הבאה:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \ & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N} v_n\xi_n \\ \text{s.t.} \ & y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n \qquad \forall n = 1,\ldots,N \\ & \xi_n \geq 0 \end{aligned}$$

כאשר $\mathbf{x}_n \in \mathbb{R}^d$ ו־ $y_n \in \{+1,-1\}$.

1. מהו הלגרנז'יאן בבעיה זו?

לפי כללי KKT שלמדנו, הלגרנז'יאן הוא:

$$L(\boldsymbol{\alpha},\mathbf{r},\mathbf{w},b,\boldsymbol{\xi}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N} v_n\xi_n - \sum_{n=1}^{N}\alpha_n(y_n(\mathbf{w}^T\mathbf{x}_n+b)-1+\xi_n) - \sum_{n=1}^{N} r_n\xi_n$$

2. מהי הבעיה הדואלית במקרה זה? יש להציג את צורתה הסופית, שלא תלויה במשתנים $\mathbf{w}, b, \boldsymbol{\xi}$.

עלינו לפתור את הבעיה

$$\max_{\boldsymbol{\alpha},\mathbf{r}} \min_{\mathbf{w},b,\boldsymbol{\xi}} L(\boldsymbol{\alpha}, \mathbf{r}, \mathbf{w}, b, \boldsymbol{\xi})$$

בכפוף לאילוצים $\alpha_n \geq 0$ ו־$r_n \geq 0$. נגזור:

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n = 0 \Rightarrow \mathbf{w}(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n$$

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

$$\frac{\partial L}{\xi_n} = C v_n - \alpha_n - r_n = 0$$

נציב בביטוי של הלגרנז'יאן ונקבל, בדומה לפיתוח של Soft Margin SVM:

$$\begin{aligned} L(\boldsymbol{\alpha}) &= \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N} v_n\xi_n - \sum_{n=1}^{N} \alpha_n(y_n(\mathbf{w}^T\mathbf{x}_n + b) - 1 + \xi_n) - \sum_{n=1}^{N} r_n\xi_n \\ &= \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N} v_n\xi_n - \mathbf{w}^T(\sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n) - b\sum_{n=1}^{N} \alpha_n y_n + \sum_{n=1}^{N} \alpha_n - \sum_{n=1}^{N} \alpha_n\xi_n - \sum_{n=1}^{N} r_n\xi_n \\ &= -\frac{1}{2}(\sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i)^T(\sum_{j=1}^{N} \alpha_j y_j \mathbf{x}_j) + \sum_{n=1}^{N} \alpha_n + \sum_{n=1}^{N} \xi_n(C v_n - \alpha_n - r_n) \\ &= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j + \sum_{n=1}^{N} \alpha_n \end{aligned}$$

המשך בדף הבא.

אם אנחנו מגדירים מטריצה M שמקיימת

$$M_{i,j} = y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

אז הבעיה הדואלית היא:

$$\max_{\boldsymbol{\alpha}} -\frac{1}{2}\boldsymbol{\alpha}^T M \boldsymbol{\alpha} + \mathbf{1}^T \boldsymbol{\alpha}$$

בכפוף לאילוצים

$$\begin{aligned} \sum_{n=1}^{N} \alpha_n y_n &= 0 \\ \alpha_n &\geq 0 \\ r_n &\geq 0 \\ Cv_n - \alpha_n - r_n &= 0 \end{aligned}$$

מכיוון ש־r_n לא מופיע בפונקצית האופטימיזציה, והוא אי־שלילי, אפשר להיפטר ממנו ולהחליף את שלושת האילוצים האחרונים, ולקבל:

$$\max_{\boldsymbol{\alpha}} -\frac{1}{2}\boldsymbol{\alpha}^T M \boldsymbol{\alpha} + \mathbf{1}^T \boldsymbol{\alpha}$$

בכפוף לאילוצים

$$\begin{aligned} \sum_{n=1}^{N} \alpha_n y_n &= 0 \\ 0 \leq \alpha_n &\leq Cv_n \end{aligned}$$

שאלה 4 - 20 נקודות

באלגוריתם PCA נתון לנו מדגם $\mathbf{x}_i$, כאשר $\mathbf{x}_i \in \mathbb{R}^d$, עבור $i = 1, \ldots, n$. בשאלה זו נרצה להרחיב את האלגוריתם ל-Kernel PCA.

קיימת פונקציית מיפוי $\phi : \mathbb{R}^d \to H$ (כאשר H הוא מרחב כלשהו אליו ממופות הנקודות) וקיים קרנל K, כך שמתקיים, לכל שתי נקודות $\mathbf{x}, \mathbf{x}'$:

$$K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$$

פונקציית המיפוי לא בהכרח נתונה, אך הקרנל נתון לנו במפורש. כלומר, לכל שתי נקודות $\mathbf{x}, \mathbf{x}'$, אנחנו יכולים לחשב את $K(\mathbf{x}, \mathbf{x}')$, אבל לא את $\phi(\mathbf{x}), \phi(\mathbf{x}')$.

נזכיר, כי באלגוריתם PCA אנו דורשים כי המדגם יהיה ממורכז, כלומר ש-

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i = \mathbf{0}$$

כדי להשיג מטרה זו, מבצעים שלב עיבוד מקדים בו מחסרים מכל קואורדינטה את הממוצע שלה. כיוון שכעת המרחב בו נפעל הוא H, נדרוש במקום זאת ש-

$$\frac{1}{n} \sum_{i=1}^{n} \phi(\mathbf{x}_i) = \mathbf{0}$$

אך נשים לב שכעת לא ניתן לחסר במפורש את הממוצע במרחב H.

(המשך בעמוד הבא)

1. במהלך האלגוריתם, משתמשים במטריצה $\overline{\overline{K}}$ המוגדרת על ידי

$$\overline{\overline{K}}_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$$

הראה/י כיצד ניתן להחליף אותה במטריצה חדשה $\overline{\overline{K'}}$ כך שהאיבר המתאים לכל שתי נקודות מהמדגם $\mathbf{x}_i, \mathbf{x}_j$, יהיה שקול להפעלה של הקרנל המקורי על אותן שתי נקודות, לאחר שהמדגם עבר מרכוז במרחב H. אין להשתמש ישירות במיפויים $\phi(\mathbf{x}_i)$. מה הסיבוכיות של פעולה זו?

$$\begin{aligned}
\overline{\overline{K'}}_{i,j} &= \left\langle \phi(\mathbf{x}_i) - \frac{1}{n}\sum_{k=1}^{n}\phi(\mathbf{x}_k), \phi(\mathbf{x}_j) - \frac{1}{n}\sum_{l=1}^{n}\phi(\mathbf{x}_l) \right\rangle \\
&= \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle - \frac{1}{n}\sum_{k=1}^{n}\langle \phi(\mathbf{x}_k), \phi(\mathbf{x}_j)\rangle - \frac{1}{n}\sum_{l=1}^{n}\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_l)\rangle \\
&\quad + \frac{1}{n^2}\sum_{k,l=1}^{n}\langle \phi(\mathbf{x}_k), \phi(\mathbf{x}_l)\rangle \\
&= \overline{\overline{K}}_{i,j} - \frac{1}{n}\sum_{k=1}^{n}\overline{\overline{K}}_{k,j} - \frac{1}{n}\sum_{l=1}^{n}\overline{\overline{K}}_{i,l} + \frac{1}{n^2}\sum_{k,l=1}^{n}\overline{\overline{K}}_{k,l}
\end{aligned}$$

הביטויים הנסכמים הם סכומי שורות, עמודות וסכום כל איברי המטריצה, והם ניתנים לחישוב מקדים בזמן $O(n^2)$, ואז חישוב כל איבר במטריצה הוא $O(1)$. סך הכל הסיבוכיות היא $O(n^2)$, כאשר יש לכפול אותה בסיבוכיות החישוב של איבר במטריצה $\overline{\overline{K}}$. אם לדוגמה סיבוכיות חישוב קרנל היא $O(d)$, נקבל כי הסיבוכיות הכוללת היא:

$$O(n^2 d)$$

2. נניח שהמדגם ממורכז כנדרש. אנו רוצים להפעיל את אלגוריתם PCA על המיפויים $\phi(\mathbf{x}_i)$. מכיוון שהטווח H של ϕ יכול להיות מרחב ממימד אינסופי או גבוה, לא נחפש את הצירים הראשיים (ה־Principal Components) עצמם, אלא נסתפק באפשרות לבצע מכפלה פנימית של כל ציר עם המיפוי $\phi(\mathbf{x})$ של נקודה חדשה כלשהי, $\mathbf{x}$.

נסמן ב־$\mathbf{u}_1, \ldots, \mathbf{u}_k$ את k הצירים הראשיים הראשונים במרחב H המתאימים לנקודות $\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)$, ותהא $\mathbf{x}$ נקודה כלשהי חדשה. הראה/י כיצד ניתן לחשב את

$$\langle \mathbf{u}_j, \phi(\mathbf{x}) \rangle$$

עבור $j = 1, \ldots, k$. מהי הסיבוכיות של הפתרון?

נסמן ב־X את המטריצה שעמודותיה הן $\phi(\mathbf{x}_i)$. נסמן את פירוק ה־SVD שלה ב:

$$X = U\Sigma V^T$$

כך ש־$\mathbf{u}_1, \ldots, \mathbf{u}_k$ הם k העמודות הראשונות ב־U. אז מתקיים:

$$U = XV\Sigma^+$$

כאשר Σ^+ הוא ה־Pseudo-inverse של Σ. מכאן אפשר לראות שכל עמודה של U היא צירוף לינארי של עמודות X, ובפרט

$$\mathbf{u}_j = \sigma_j^{-1} X \mathbf{v}_j$$

נשים לב שכדי למצוא את V, Σ אנחנו לא יכולים לחשב את ה־SVD ישירות. במקום זה, עלינו לחשב את $X^T X$, שזו המטריצה $\overline{\overline{K}}_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$, בעלות של $O(n^2)$ כפול סיבוכיות חישוב קרנל; ואז עלינו לחשב את הערכים העצמיים והוקטורים העצמיים בעלות של $O(n^3)$. מכיוון ש: $X^T X = V\Sigma^T\Sigma V^T$, ניתן לחלץ כך את V, Σ. ואז:

$$\begin{aligned}
\langle \mathbf{u}_j, \phi(\mathbf{x}) \rangle &= \langle \sigma_j^{-1} X \mathbf{v}_j, \phi(\mathbf{x}) \rangle \\
&= \left\langle \sum_{i=1}^{n} \sigma_j^{-1} V_{ij} \phi(\mathbf{x}_i), \phi(\mathbf{x}) \right\rangle \\
&= \sum_{i=1}^{n} \sigma_j^{-1} V_{ij} \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle \\
&= \sum_{i=1}^{n} \sigma_j^{-1} V_{ij} K(\mathbf{x}_i, \mathbf{x})
\end{aligned}$$

סיבוכיות: בנוסף לחישוב וקטורים וערכים עצמיים, מחשבים את הביטוי לעיל בסיבוכיות של $O(n)$ כפול סיבוכיות חישוב קרנל. סך הכל, אם לדוגמה סיבוכיות חישוב קרנל היא $O(d)$, נקבל כי הסיבוכיות הכוללת היא:

$$O(n^3 + n^2 d)$$

תעודת זהות:
מספר מחברת:

שאלה 5 - 20 נקודות

נתון לנו מדגם $(\mathbf{x}_i, y_i)$, כאשר $\mathbf{x}_i \in \mathbb{R}^d$ ו- $y_i \in \mathbb{R}$. נרצה לשנות את בעיית הרגרסיה הלינארית כדי שתתאים ל- L1 loss.

1. מהי פונקציית האופטימיזציה המתאימה לבעיה החדשה?

נרצה להביא למינימום את סכום פונקציות ה-loss של כל הנקודות:

$$\mathbf{a}_{opt} = \arg\min_{\mathbf{a}\in\mathbb{R}^d} \sum_{i=1}^{n} |\mathbf{a}^T \mathbf{x}_i - y_i|$$

2. כיצד ניתן לפתור את הבעיה החדשה באמצעות Linear Programming?

נגדיר משתני עזר חדשים, r_i, ונפתור את הבעיה הבאה:

$$\arg\min_{\mathbf{a},\boldsymbol{\xi}} \sum_{i=1}^{n} r_i$$

בכפוף לאילוצים:

$$\begin{aligned} r_i \geq & \mathbf{a}^T\mathbf{x}_i - y_i \\ r_i \geq & -\left(\mathbf{a}^T\mathbf{x}_i - y_i\right) \end{aligned}$$

לכל i. ניסוח זה מוודא שכל משתנה עזר גדול או שווה לשארית המתאימה. כיוון שאנחנו מחפשים מינימום, הוא יהיה שווה אליה ממש, ולכן הבעיה שקולה למינימום על סכום הערכים המוחלטים.