תעודת זהות:
מספר מחברת:

מבוא ללמידה חישובית – מבחן מועד א׳
סמסטר א׳ תשע״ה $(2014/5)$

בית הספר למדעי המחשב, אוניברסיטת תל־אביב

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

3.2.2015

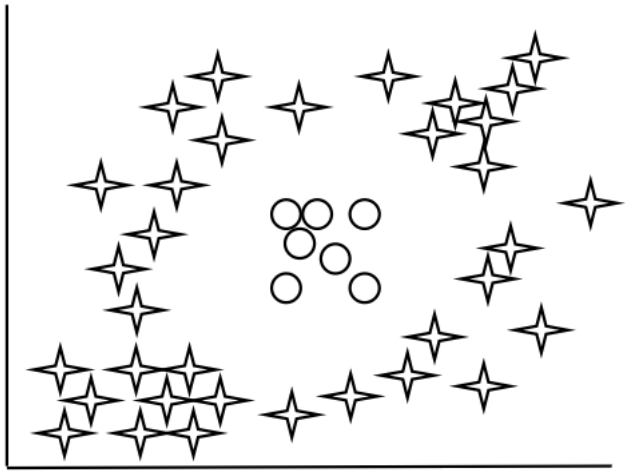
<u>הוראות</u>:

1. מומלץ לקרוא את כל ההנחיות והשאלות בתחילת המבחן לפני תחילת כתיבת התשובות.
2. משך הבחינה – **שלוש שעות**. לא תינתן כל הארכה נוספת.
3. חומר עזר מותר: דף נוסחאות בגודל A4.
4. **יש לענות על השאלות במקום המיועד לכך בטופס השאלון (טופס זה)**. מחברות הבחינה לא ייקראו, ותשמשנה כטיוטה בלבד.
5. יש למלא בכל דף של השאלון מספר ת.ז. ומספר מחברת.
6. במבחן 6 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 - 15 נקודות

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

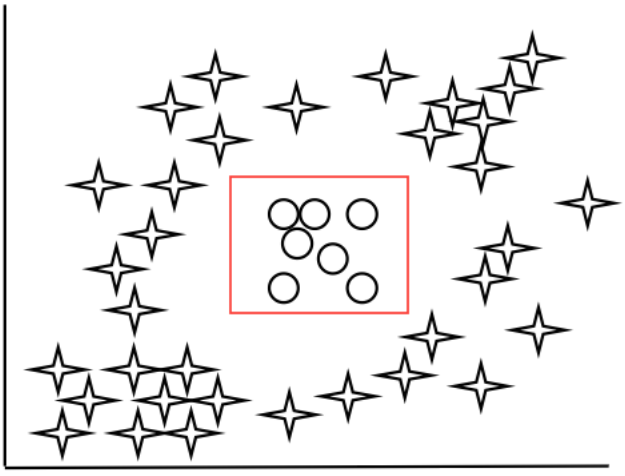


עבור כל אחד מהאלגוריתמים הבאים, קבע/י האם ניתן להריץ אותו עד לקבלת מסווג עם שגיאת למידה אפס, על המדגם הנתון. אם כן, ציין/י קו הפרדה מתאים למסווג המתקבל. אם לא, הסבר/הסבירי מדוע.

1. AdaBoost, כאשר ה־weak classifiers מתוכם האלגוריתם בוחר הם כל ה־decision stumps המקבילים לצירים (כלומר, עבור כל מימד i, וסף a, ניתן לבחור classifier שמסווג על פי תוצאת ההשוואה $x_i < a$).

   ☐ לא יכול להגיע לשגיאה אפס. הסבר:

   ✓ יכול להגיע לשגיאה אפס. קו ההפרדה של המסווג:

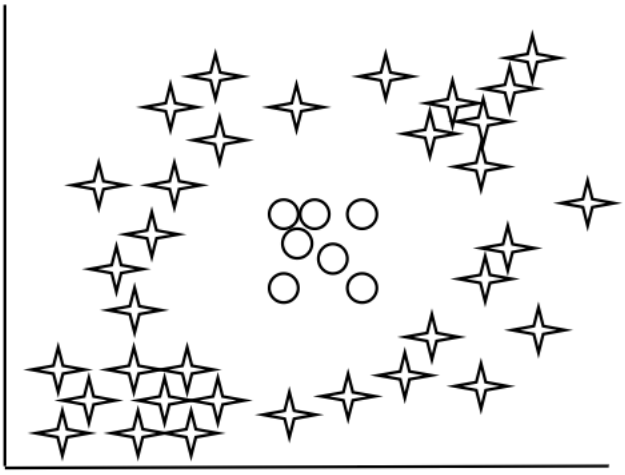


ניתן להגיע לקו מפריד כזה על ידי שימוש נכון במשקולות של 5 מסווגים - שניים בכל ציר, שהסף שלהם מתאים לצלעות של המרובע, ועוד מסווג קבוע.

2. Perceptron (כאשר הדגימות מוזנות לו אחת אחרי השניה בסדר שרירותי כלשהו).

✓ לא יכול להגיע לשגיאה אפס. הסבר: הדגימות לא ניתנות להפרדה על ידי מפריד לינארי. זאת מכיוון שיש נקודה חיובית (עיגול) שנמצאת בקמור של הנקודות השלילית (כוכב).

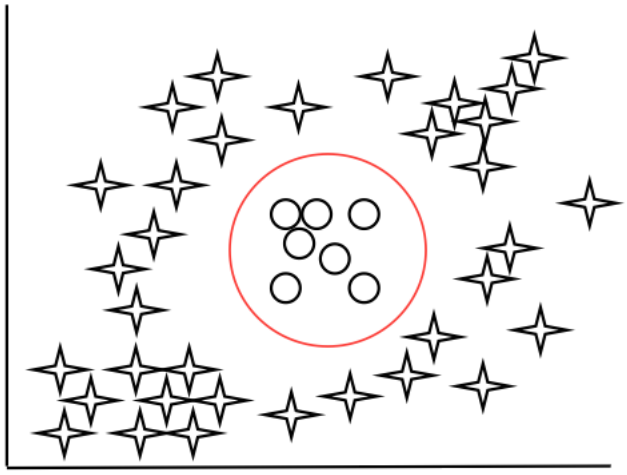
☐ יכול להגיע לשגיאה אפס. קו ההפרדה של המסווג:



3. SVM עם Gaussian kernel.

☐ לא יכול להגיע לשגיאה אפס. הסבר:

✓ יכול להגיע לשגיאה אפס.



תעודת זהות:
מספר מחברת:

2 שאלה 2 - 15 נקודות

מאמנים מסווג א' (classifier) באמצעות אלגוריתם A ללימוד מסווגים, על מדגם אימון מסוים. כעת מפעילים על כל נקודה במדגם טרנספורמציה נתונה T, ועל המדגם החדש מאמנים מסווג ב' באמצעות אותו האלגוריתם. תהי $\mathbf{x}$ נקודה חדשה. מפעילים את מסווג א' על $\mathbf{x}$, ומפעילים את מסווג ב' על $T(\mathbf{x})$. האם הם בהכרח יחזירו את אותו הסיווג? נמק/י את תשובתך בקצרה.

1. A הוא Hard Margin SVM.

(א) $T(\mathbf{x}) = \mathbf{x} + \mathbf{x}_0$, כאשר $\mathbf{x}_0$ הוא וקטור.

✓ בהכרח אותו סיווג

☐ לא בהכרח אותו סיווג

הסבר: אינטואיבית, המפריד יזוז יחד עם הנקודות. הוכחה מפורטת: פורמלית, הבעיה המקורית היא:

$$\arg\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \text{ s.t. } y_i(\mathbf{w}\cdot \mathbf{x_i} + b) \geq 1$$

לאחר הפעלת T, נקבל את הבעיה החדשה:

$$\arg\min_{\mathbf{w}^*,b^*} \frac{1}{2}\|\mathbf{w}^*\|^2 \text{ s.t. } y_i(\mathbf{w}^*\cdot \mathbf{T}(\mathbf{x_i}) + b^*) \geq 1$$

אם $\mathbf{w}, b$ הוא מפריד אפשרי בבעיה המקורית, אז $\mathbf{w}^* = \mathbf{w}, b^* = b - \mathbf{w}\cdot \mathbf{x_0}$ הוא מפריד אפשרי בבעיה החדשה - כדי לראות זאת, פשוט מציבים ורואים שמקבלים את הבעיה המקורית. מכאן, שקיימת התאמה חד־חד ערכית בין המפרידים בבעיה המקורית לבעיה החדשה. מעבר לכך, כיוון ש־$\frac{1}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\|\mathbf{w}^*\|^2$, הרי שאם המפריד האופטימלי בבעיה המקורית היה $\mathbf{w}_{opt}, b_{opt}$, אזי המפריד האופטימלי בבעיה החדשה יהיה:

$$\mathbf{w}^*_{opt} = \mathbf{w}_{opt}, b^*_{opt} = b_{opt} - \mathbf{w}_{opt}\cdot \mathbf{x}_0$$

כעת, עבור נקודה חדשה כלשהי $\mathbf{x}$, מתקבל:

$$\mathbf{w}^*_{opt}\cdot T(\mathbf{x}) + b^*_{opt} = \mathbf{w}_{opt}\cdot (\mathbf{x} + \mathbf{x_0}) + b_{opt} - \mathbf{w}_{opt}\cdot \mathbf{x}_0 = \mathbf{w}_{opt}\cdot \mathbf{x} + b_{opt}$$

ובפרט יוחזר אותו סיווג.

(ב) $T(\mathbf{x}) = a\mathbf{x}$, כאשר $a \neq 0$ הוא סקלר.

✓ בהכרח אותו סיווג

☐ לא בהכרח אותו סיווג

הסבר: אינטואיטיבית, השינוי הוא רק בקנה המידה של הבעיה באופן סימטרי בכל הצירים. הניתוח דומה לסעיף א', רק שהפעם ההתאמה המתקבלת היא:

$$\mathbf{w}^*_{opt} = \frac{1}{a} \cdot \mathbf{w}_{opt}, b^*_{opt} = b_{opt}$$

בנוסף, הבאה למינימום של $\frac{1}{2}\|\mathbf{w}\|^2$ שקולה להבאה למינימום של $\frac{1}{2}\|\mathbf{w}^*\|^2 = \frac{1}{2a^2}\|\mathbf{w}\|^2$ ולכן עבורה נקודה חדשה, יתקבל אותו סיווג, שכן:

$$\mathbf{w}^*_{opt} \cdot T(\mathbf{x}) + b^*_{opt} = \frac{1}{a} \cdot \mathbf{w}_{opt} \cdot a \cdot \mathbf{x} + b_{opt} = \mathbf{w}_{opt} \cdot \mathbf{x} + b_{opt}$$

(ג) $T(\mathbf{x}) = D\mathbf{x}$, כאשר D היא מטריצה אלכסונית, ללא 0 על האלכסון.

☐ בהכרח אותו סיווג

✓ לא בהכרח אותו סיווג

הסבר: הפעם אותו נימוק לא עובד, שכן מינימיזציה של הנורמה לא שקולה. דוגמה נגדית: עבור המדגם S

$$S = \{\langle (1,1), 1\rangle, \langle (-1,-1), -1\rangle\}$$

יתקבל המפריד $\mathbf{w}_{opt} = (\frac{1}{2}, \frac{1}{2})^T, b_{opt} = 0$. אבל לאחר הפעלת הטרנספורמציה $D = \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix}$ נקבל את המפריד $\mathbf{w}^*_{opt} = (\frac{10}{101}, \frac{1}{101})^T, b^*_{opt} = 0$ שמסווג חלק מהמישור באופן שונה, לדוגמה את הנקודה

$$\mathbf{x} = (-1, 2), T(\mathbf{x}) = (-10, 2)$$

(ד) $T(\mathbf{x}) = U\mathbf{x}$, כאשר U היא מטריצה אורתונורמלית.

✓ בהכרח אותו סיווג

☐ לא בהכרח אותו סיווג

הסבר: אינטואיטיבית, מטריצה אורתונורמלית היא הרכבה של סיבובים ושיקופים סביב הראשית, ולכן המפריד האופטימלי יסתובב/ישוקף באופן מתאים. פורמלית, גם פה הנימוק הוא כמו בסעיף א', רק שהפעם ההתאמה המתקבלת היא:

$$\mathbf{w}^*_{opt} = U\mathbf{w}_{opt}, b^*_{opt} = b_{opt}$$

בנוסף, הבאה למינימום של $\frac{1}{2}\|\mathbf{w}\|^2$ שקולה להבאה למינימום של $\frac{1}{2}\|\mathbf{w}^*\|^2 = \frac{1}{2}\|\mathbf{U}\mathbf{w}\|^2 = \frac{1}{2}\|\mathbf{w}\|^2$ ולכן עבורה נקודה חדשה, יתקבל אותו סיווג, שכן:

$$\mathbf{w}^*_{opt} \cdot T(\mathbf{x}) + b^*_{opt} = \langle U\mathbf{w}_{opt}, U\mathbf{x} \rangle + b_{opt} = \mathbf{w}_{opt} \cdot \mathbf{x} + b_{opt}$$

2. A הוא עץ החלטה עם Decision stumps.

(א) $T(\mathbf{x}) = \mathbf{x} + \mathbf{x}_0$, כאשר $\mathbf{x}_0$ הוא וקטור.

✓ בהכרח אותו סיווג

☐ לא בהכרח אותו סיווג

הסבר: נביט בצומת מסוים בעץ החלטה. בבחירת המסווג בצומת, בבעיה המקורית בחרנו במסווג הטוב ביותר מבין כל המסווגים מהצורה:

$$h^i_{v,c_0,c_1}(\mathbf{x}) = \begin{cases} c_0 & \text{IF } \mathbf{x}_i > a \\ c_1 & \text{IF } \mathbf{x}_i \leq a \end{cases}$$

עבור $c_0, c_1 \in \{1, -1\}$. נניח שהמסווג הטוב ביותר היה h^i_{v,c_0,c_1} אז בבעיה החדשה, המסווג הטוב ביותר יהיה בהכרח:

$$h^i_{v+(\mathbf{x}_0)_i,c_0,c_1}$$

ועל נקודה חדשה נקבל את אותו הסיווג.

(ב) $T(\mathbf{x}) = a\mathbf{x}$, כאשר $a \neq 0$ הוא סקלר.

✓ בהכרח אותו סיווג

☐ לא בהכרח אותו סיווג

הסבר: בדומה לסעיף הקודם, רק שכעת המסווג הטוב ביותר יהיה

$$h^i_{a\cdot v,c_0,c_1}$$

(ג) $T(\mathbf{x}) = D\mathbf{x}$, כאשר D היא מטריצה אלכסונית, ללא 0 על האלכסון.

✓ בהכרח אותו סיווג

☐ לא בהכרח אותו סיווג

הסבר: בדומה לסעיף הקודם, רק שכעת המסווג הטוב ביותר יהיה

$$h^i_{D_{i,i}\cdot v,c_0,c_1}$$

(ד) $T(\mathbf{x}) = U\mathbf{x}$, כאשר U היא מטריצה אורתונורמלית.

☐ בהכרח אותו סיווג

✓ לא בהכרח אותו סיווג

הסבר: מטריצת סיבוב יכולה לשנות את אוסף המסווגים האפשריים. כך לדוגמה, אם המדגם הוא

$$S = \{\langle (1,1),1\rangle, \langle (1,-1),1\rangle, \langle (-1,1),-1\rangle, \langle (-1,-1),-1\rangle\}$$

אז קיים מפריד עם 0 שגיאה. אבל לאחר הפעלת סיבוב של 45 מעלות, לא קיים מפריד כזה.

3 שאלה 3 - 20 נקודות

נסמן ב-$\mathbf{e}_i \in \mathbb{R}^d$ את וקטור היחידה ה-i של הבסיס הסטנדרטי (כלומר, וקטור בן d איברים שכולם 0 מלבד האיבר ה-i שהוא 1). נגדיר את מדגם האימון הבא:

$$S = \cup_{i=1}^{d} \{\langle \mathbf{e}_i, 1\rangle, \langle -\mathbf{e}_i, -1\rangle\}$$

1. כתב/י את בעיית האופטימיזציה הפרימאלית של SVM, עבור המקרה $d = 2$. מהו פתרון הבעיה, ומהם ה-$\mathbf{w}, b$ האופטימליים? כמה Support Vectors ישנם?

הפתרון זהה עבור שני הסעיפים. הבעיה הפרימאלית היא:

$$\arg\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$

$$\begin{aligned} \text{s.t.} \quad & 1 \cdot (\mathbf{w} \cdot \mathbf{e}_i + b) \geq 1 \\ & -1 \cdot (\mathbf{w} \cdot -\mathbf{e}_i + b) \geq 1 \\ & \forall i \in \{1, \ldots, d\} \end{aligned}$$

כאשר $\mathbf{w} = (w_1, \ldots, w_d)$. אי השוויונות מתרגמים לאי השוויונות:

$$w_i + b \geq 1, w_i - b \geq 1, \forall i \in \{1, \ldots, d\}$$

כלומר, ה-$\mathbf{w}$ האופטימלי הוא זה שעבורו $w_i = \max(1 + b, 1 - b)$ לכל i. ה-b שמביא את הביטוי מצד ימין למינימום הוא 0, ומכאן שהערכים האופטימליים הם:

$$\mathbf{w}^* = (1, \ldots, 1), b^* = 0$$

נשים לב, שלכל דגימה במדגם האימון מתקיים:

$$\begin{aligned} 1 \cdot (\mathbf{w}^* \cdot \mathbf{e}_i + b*) &= 1 \\ -1 \cdot (\mathbf{w}^* \cdot -\mathbf{e}_i + b*) &= 1 \end{aligned}$$

ולכן כל הוקטורים במדגם האימון $(2d)$ הם Support Vectors.

2. במקרה הכללי, עבור d כלשהו, מהם $\mathbf{w}, b$ האופטימליים? כמה Support Vectors ישנם?

ראו סעיף 1.

4 שאלה 4 - 20 נקודות

מדגם $\mathbf{x_1}, \ldots, \mathbf{x_m} \in \mathbb{R}^n$ מיוצר כך: נתונה מטריצה W בת n שורות ו־d עמודות, מדרגה מלאה d, כאשר $d < n$. עבור כל $i = 1, \ldots, m$, מגרילים וקטור מקדמים $\mathbf{z}_i \in \mathbb{R}^d$ מההתפלגות הרב נורמלית:

$$\mathbf{z}_i \sim N(\mathbf{0}, \sigma^2 I_d)$$

בנוסף, מגרילים וקטור רעש $\mathbf{e_i} \in \mathbb{R}^n$ מההתפלגות הרב נורמלית:

$$\mathbf{e}_i \sim N(\mathbf{0}, \tau^2 I_n)$$

לבסוף, יוצרים את הנקודה ה־i כך:

$$\mathbf{x}_i = W\mathbf{z}_i + \mathbf{e}_i$$

במודל שתיארנו לעיל, נתייחס ל־$\mathbf{z}_i$ בתור הפרמטרים, בעוד ש־W, σ, τ קבועים וידועים. זהו מודל בייסיאני, כאשר $\mathbf{z}_i \sim N(\mathbf{0}, \sigma^2 I_d)$ מתאר את ה־Prior.

1. כתב/י את פונקציית ה־Log Prior של הפרמטרים, כלומר את:

$$\log \Pr(\mathbf{z_1}, \ldots, \mathbf{z_m})$$

$$\begin{aligned}
\log \Pr(\mathbf{z_1}, \ldots, \mathbf{z_m}) &= \log \prod_{i=1}^{m} \Pr(\mathbf{z_i}) = \sum_{i=1}^{m} \log \Pr(\mathbf{z_i}) \\
&= \sum_{i=1}^{m} \log \left(\frac{1}{(\sqrt{2\pi\sigma^2})^d} \cdot \exp\left(-\frac{\|\mathbf{z}_i\|^2}{2\sigma^2} \right) \right) \\
&= \frac{md}{2} \cdot \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{m} \|\mathbf{z}_i\|^2
\end{aligned}$$

2. כתב/י את הפונקצית ה־Log Likelihood של המדגם עם ה־Prior המתאים, כלומר את:

$$\log\left(\Pr\left(\mathbf{x_1},\ldots,\mathbf{x_m};\mathbf{z_1},\ldots,\mathbf{z_m}\right)\cdot\Pr\left(\mathbf{z_1},\ldots,\mathbf{z_m}\right)\right)$$

$$\begin{aligned}
& \log\left(\Pr\left(\mathbf{x_1},\ldots,\mathbf{x_m};\mathbf{z_1},\ldots,\mathbf{z_m}\right)\cdot\Pr\left(\mathbf{z_1},\ldots,\mathbf{z_m}\right)\right) \\
= \; & \log\Pr\left(\mathbf{x_1},\ldots,\mathbf{x_m};\mathbf{z_1},\ldots,\mathbf{z_m}\right)+\log\Pr\left(\mathbf{z_1},\ldots,\mathbf{z_m}\right) \\
= \; & \sum_{i=1}^{m}\log\Pr(\mathbf{x_i};\mathbf{z_i})+\log\Pr\left(\mathbf{z_1},\ldots,\mathbf{z_m}\right) \\
= \; & \sum_{i=1}^{m}\log\left(\frac{1}{(\sqrt{2\pi\tau^2})^n}\cdot\exp\left(-\frac{\|\mathbf{x}_i-W\mathbf{z}_i\|^2}{2\tau^2}\right)\right)+\log\Pr\left(\mathbf{z_1},\ldots,\mathbf{z_m}\right) \\
= \; & \frac{mn}{2}\cdot\log\left(2\pi\tau^2\right)-\frac{1}{2\tau^2}\sum_{i=1}^{m}\|\mathbf{x}_i-W\mathbf{z}_i\|^2+\frac{md}{2}\cdot\log\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\|\mathbf{z}_i\|^2
\end{aligned}$$

3. מהו אומד ה־MAP של $\mathbf{z_i}$?

אומד ה־MAP הוא זה שמביא למקסימום את הביטוי מסעיף 2. נשים לב שהבאה למקסימום שקולה להבאה למינימום של הביטוי:

$$\frac{1}{2\tau^2}\sum_{i=1}^{m}\|\mathbf{x}_i-W\mathbf{z}_i\|^2+\frac{1}{2\sigma^2}\sum_{i=1}^{m}\|\mathbf{z}_i\|^2$$

תחילה נשים לב, שאם נכפול ב־$2\tau^2$, נקבל בדיוק את פונקציית המטרה של Ridge Regression, עם $\lambda=\tau^2/\sigma^2$, ולכן הפתרון ידוע. ניתן גם לקבל אותו ישירות - נגזור (וקטורית) לפי $\mathbf{z_i}$, נשווה ל־0 ונקבל:

$$\begin{aligned}
& \frac{1}{2\tau^2}\cdot(-2W^T)\cdot(\mathbf{x}_i-W\mathbf{z}_i)+\frac{1}{2\sigma^2}\cdot 2\mathbf{z}_i=\mathbf{0}\Rightarrow \\
& \left(\frac{1}{\tau^2}\cdot W^TW+\frac{1}{\sigma^2}I_d\right)\mathbf{z}_i-\frac{1}{\tau^2}\cdot W^T\mathbf{x}_i=\mathbf{0}\Rightarrow \\
& (\mathbf{z}_i)_{MAP}=\left(W^TW+\frac{\tau^2}{\sigma^2}I_d\right)^{-1}W^T\mathbf{x}_i
\end{aligned}$$

נשים לב שהמטריצה $\left(W^TW + \frac{\tau^2}{\sigma^2}I_d\right)$ היא positive definite. כדי לראות זאת, ניקח וקטור שרירותי $\mathbf{v} \neq \mathbf{0}$, ואז:

$$\mathbf{v}^T \left(W^TW + \frac{\tau^2}{\sigma^2}I_d\right) \mathbf{v} = \|W\mathbf{v}\|^2 + \frac{\tau^2}{\sigma^2}\|\mathbf{v}\|^2 > 0$$

מכאן שהמטריצה הפיכה (ולכן הפתרון חוקי), וגם שההסיאן של הפונקציה הוא positive definite, ומכאן שמדובר במינימום.

5 שאלה 5 - 15 נקודות

נתון מדגם הכולל את הנקודות (x,y) הבאות:

$$(-2,10),(-1,5),(0,0),(1,5),(2,10)$$

1. מהו השיפוע של הקו המתאים לרגרסיה לינארית? נמק/י. אין הכרח לחשב את הקו.

מכיוון שמדובר בבעיית רגרסיה, השיפוע חייב להיות סופי (הקו לא יכול להיות מאונך). משיקולי סימטריה, הקו הנכון יהיה עם שיפוע 0. ניתן גם לפתור את בעיית הרגרסיה: במקרה הזה,

$$X=\begin{pmatrix}-2\\-1\\0\\1\\2\end{pmatrix},\mathbf{y}=\begin{pmatrix}10\\5\\0\\5\\10\end{pmatrix},$$

ואז המקדם המתאים הוא:

$$\beta=(X^TX)^{-1}(X^T)\mathbf{y}=(4+1+0+1+4)^{-1}\cdot(-20-5+0+5+20)=0$$

2. מהו השיפוע של הקו המתאים ל-PCA של מימד אחד? נמק/י. אין הכרח לחשב את הקו.

עכשיו זו לא בעיית רגרסיה, ולכן גם קווים אנכיים הם אפשריים. שוב משיקולי סימטריה, אפשר לראות שניתן לשקול רק שני קווים - קו אופקי וקו אנכי. נזכור שאלגוריתם PCA מנסה להביא למינימום את סכום ריבועי המרחקים הקצרים ביותר לקו. סכום ריבועי המרחקים שמתאימים לקו $y=0$ הוא $2\cdot(10^2+5^2)=250$, ואילו סכום ריבועי המרחקים שמתאימים לקו $x=0$ הוא $2\cdot(2^2+1^2)=10$. כלומר, השיפוע המתאים הוא ∞. ניתן גם לחשב במפורש. במקרה זה,

$$X=\begin{pmatrix}-2&-1&0&1&2\\10&5&0&5&10\end{pmatrix}$$

נחסר את הממוצע מכל שורה ונחשב את $X_0X_0^T$:

$$X_0=\begin{pmatrix}-2&-1&0&1&2\\4&-1&-6&-1&4\end{pmatrix},X_0X_0^T=\begin{pmatrix}10&0\\0&70\end{pmatrix},$$

המטריצה כבר מלוכסנת, ואפשר לראות שהערך העצמי הכי גדול הוא זה שמתאים לוקטור העצמי:

$$\begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

שמתאים לקו האנכי עם שיפוע ∞.

6 שאלה 6 - 20 נקודות

נסמן ב-$\mathbf{e}_i \in \mathbb{R}^m$ את וקטור היחידה ה-i של הבסיס הסטנדרטי (כלומר, וקטור בן m איברים שכולם 0 מלבד האיבר ה-i שהוא 1). נגדיר את מדגם האימון הבא:

$$S = \{\langle -\mathbf{e}_1, -1\rangle, \ldots, \langle -\mathbf{e}_m, -1\rangle\}$$

נזכיר שבהקשר של אלגוריתם Perceptron, ה-margin של המפריד המוגדר על ידי וקטור $\mathbf{w}^*$ הוא:

$$\gamma = \min_{\mathbf{x}\in S} \frac{|\mathbf{x}\cdot\mathbf{w}^*|}{\|\mathbf{x}\|}$$

1. נגדיר:

$$\mathbf{w}^* = \frac{1}{\sqrt{m}} \cdot (1, 1, \ldots, 1)$$

מהו ה-Margin של המדגם ביחס למפריד המוגדר על ידי $\mathbf{w}^*$?

לכל $-\mathbf{e}_i \in S$, מתקיים $|-\mathbf{e}_i \cdot \mathbf{w}^*| = 1/\sqrt{m}$. בנוסף, $\|-\mathbf{e_i}\| = 1$ ולכן:

$$\gamma = \min_{\mathbf{e_i}} \frac{|-\mathbf{e_i}\cdot\mathbf{w}^*|}{\|-\mathbf{e_i}\|} = \frac{1}{\sqrt{m}}$$

2. מפעילים את אלגוריתם Perceptron על המדגם. מהו החסם העליון התיאורטי על כמות השגיאות שהאלגוריתם יבצע, כפונקציה של m?

כיוון שכל הנקודות שליליות, הרי שהמדגם ניתן להפרדה מלאה על ידי על־מישור. בנוסף לכך, ניתן לראות בקלות כי העל־מישור המוגדר על ידי $\mathbf{w}^*$ הוא מפריד. מכאן, על פי משפט שראינו, מספר הטעויות במקרה זה חסום מלמעלה על ידי:

$$\frac{1}{\gamma^2} = m$$

3. כמה שגיאות יבצע האלגוריתם בפועל על המדגם?

במקרה הכללי ביותר, האלגוריתם יכול לקבל את הנקודות בכל סדר שהוא, ועם חזרות. נטען כי האלגוריתם טועה בדיוק פעם אחת על כל נקודה, בפעם הראשונה שהוא רואה אותה. נאמר שבאיטרציה ה־t, האלגוריתם נתקל לראשונה בנקודה $-\mathbf{e}_i$. בכל זמן נתון, $\mathbf{w}^t$ הוא קומבינציה לינארית של הנקודות שהאלגוריתם ראה עד כה. כיוון שהוא לא ראה את $-\mathbf{e}_i$, הרי שקיים צירוף:

$$\mathbf{w}^t = \sum_{j \neq i} \alpha_j \mathbf{e}_j$$

כיוון שלכל $j \neq i$, מתקיים $\mathbf{e}_i \cdot \mathbf{e}_j = 0$, בהכרח מתקבל כי $\mathbf{w}^t \cdot -\mathbf{e}_i = 0$, ולכן האלגוריתם יסווג את הנקודה כחיובית - על אף שהיא שלילית.

לאחר הטעות, נעדכן את הוקטור:

$$\mathbf{w}^{t+1} = \mathbf{w}^t + \mathbf{e}_i$$

כעת נאמר שבאיטרציה ה־t', האלגוריתם ייתקל שוב בנקודה $\mathbf{e}_i$. כעת נקבל את הסיווג:

$$sign(\mathbf{w}^{t'} \cdot -\mathbf{e}_i) = sign\left(\left(\sum_{j \neq i} \alpha'_j \mathbf{e}_j + \mathbf{e}_i\right) \cdot -\mathbf{e}_i\right) = -1 > 0$$

והאלגוריתם לא ייטעה.

מכאן, שהאלגוריתם יבצע בדיוק m טעויות, ולכן החסם מסעיף 2 הדוק.