הטכניון – מכון טכנולוגי לישראל

הפקולטה למדעי המחשב

מבוא למערכות לומדות (236756)

סמסטר אביב תשפ"א – 15 ביולי 2021

מרצה: ד"ר ניר רוזנפלד

מבחן מסכם מועד א' – פיתרון חלקי

הנחיות הבחינה:

- **משך הבחינה:** 3 שעות.
- **חומר עזר:** המבחן בחומר סגור (ללא ספרים, מחברות, דפי נוסחאות).
- אין צורך במחשבון.
- מותר לכתוב בעט או בעיפרון, כל עוד הכתב קריא וברור.
- יש לכתוב את תשובותיכם **על גבי שאלון זה** בכתב יד קריא. תשובה בכתב יד שאינו קריא לא תיבדק.
- במבחן 12 עמודים ממוספרים סה"כ, כולל עמוד שער זה שמספרו 1 ושלושה עמודי טיוטה בסוף הגיליון.
- אין בחירה בין השאלות. יש בסה"כ 104 נקודות והציון המירבי הוא 100.
- נא לכתוב רק את המבוקש ולצרף הסברים קצרים עפ"י ההנחיות.
- **בתום המבחן יש להגיש את שאלון זה בלבד.**
- בצהוב: הבהרות שפורסמו בזמן הבחינה.

מבנה הבחינה:

- **חלק א' [72 נק']:** 4 שאלות פתוחות [כל אחת 18 נק']
- **חלק ב' [32 נק']:** 8 שאלות סגורות (אמריקאיות) [כל אחת 4 נק']

בהצלחה!

חלק א' – שאלות פתוחות [72 נק']

שאלה 1 [18 נק']

נתון סט אימון עם תיוגים בינאריים ולומדים עליו מסווג. בסט האימון אין שתי דוגמאות זהות.

כעת, בוחרים באקראי 3 דוגמאות אימון שונות ומשכפלים אותן (כל אחת עם התיוג שלה) כך שיופיעו פעמיים בסט האימון.

לבסוף, מאמנים מסווג חדש על סט האימון המעודכן.

לכל אחד מאלגוריתמי הלמידה הבאים סמנו האם גבולות ההחלטה (decision boundaries) של המסווג החדש לאחר השכפול זהים בהכרח לאלה של המסווג המקורי. רק אם סימנתם שהגבולות **לא בהכרח זהים**, הסבירו בקצרה מדוע.

הניחו שאין צעדים סטוכסטיים (אקראיים) בריצת האלגוריתמים.

א. ID3 המשתמש באנטרופיה ובונה עץ בעומק מירבי 4 — גבולות ההחלטה: **זהים בהכרח** / **לא בהכרח זהים**

הסבר (אם סימנתם "לא בהכרח"): ______________________________

____

ב. k-NN כאשר $k = 1$ — גבולות ההחלטה: **זהים בהכרח** / **לא בהכרח זהים**

הסבר (אם סימנתם "לא בהכרח"): ______________________________

____

ג. k-NN כאשר $k = 3$ — גבולות ההחלטה: **זהים בהכרח** / **לא בהכרח זהים**

הסבר (אם סימנתם "לא בהכרח"): ______________________________

____

ד. Perceptron בהנחה שהדאטה פריד ליניארית — גבולות ההחלטה: **זהים בהכרח** / **לא בהכרח זהים**

הסבר (אם סימנתם "לא בהכרח"): יש אפשרות לטעות על אותה דוגמה מספר פעמים, ולכן הכפלת דוגמאות יכולה

לשנות את סדר העדכון והתוצאה הסופית

ה. AdaBoost with decision stumps — גבולות ההחלטה: **זהים בהכרח** / **לא בהכרח זהים**

הסבר (אם סימנתם "לא בהכרח"): ______________________________

____

ו. Hard-SVM בהנחה שהדאטה פריד ליניארית — גבולות ההחלטה: **זהים בהכרח** / **לא בהכרח זהים**

הסבר (אם סימנתם "לא בהכרח"): ______________________________

____

שאלה 2 [18 נק']

א. [6 נק'] הוכיחו את תכונת המונוטוניות של VC-dimension:

לכל שתי מחלקות היפותזות, אם $\mathcal{H}_1 \subseteq \mathcal{H}_2$ אזי $\text{VCdim}(\mathcal{H}_1) \leq \text{VCdim}(\mathcal{H}_2)$.

הוכחה:

נסמן $\text{VCdim}(\mathcal{H}_1) = k$.

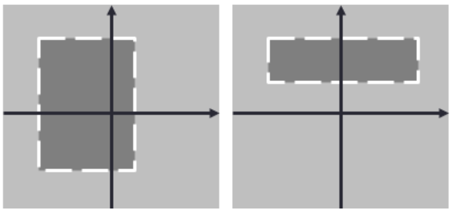
קיימות k נקודות שלכל תיוג שלהן קיימת היפותזה ב-$\mathcal{H}_1$ שמתאימה את התיוג באופן מושלם.

אותן ההיפותזות קיימות ב-$\mathcal{H}_2$ ויכולות לנתץ את אותן k נקודות, ולכן $k \leq \text{VCdim}(\mathcal{H}_2)$.

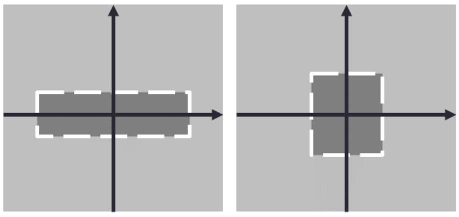
בתרגול הגדרנו את מחלקת ההיפותזות $\mathcal{H}_{\text{aligned}}$ של מלבנים מקבילים לצירים בדו-מימד והראינו שמתקיים $\text{VCdim}(\mathcal{H}_{\text{aligned}}) = 4$. הבהרה: האיזור שבתוך המלבן מסווג כחיובי והאיזור החיצוני כשלילי.

כעת נגדיר את מחלקת $\mathcal{H}_{\text{centered}}$ של מלבנים מקבילים לצירים בדו-מימד **שמרכזם** בדיוק בראשית הצירים.

שתי היפותזות מתוך $\mathcal{H}_{\text{aligned}}$



שתי היפותזות מתוך $\mathcal{H}_{\text{centered}}$



ב. [6 נק'] מהו מימד ה-VC של המחלקה החדשה? מלאו.

$\text{VCdim}(\mathcal{H}_{\text{centered}}) = \boxed{2}$

ג. [6 נק'] הוכיחו את תשובתכם לסעיף הקודם.

הוכחה:

יש להראות **שקיימות** שתי נקודות שאפשר לנתץ (להתאים כל תיוג שלהן).

יש להראות ש**כל** שלוש נקודות, **קיים** תיוג שלא ניתן להתאים.

שאלה 3 [18 נק']

נתון סט אימון עם דוגמאות חד-ממדיות $x_1, \dots, x_m \in \mathbb{R}$ ותיוגים רציפים $y_1, \dots, y_m \in \mathbb{R}$.

פותרים בעיית רגרסיה ליניארית עם רגולריזציה (עם $\lambda = 1$): $\hat{w} = \operatorname{argmin}_{w\in\mathbb{R}}\left(\sum_{i=1}^{m}(wx_i - y_i)^2 + R(w) \right)$

בפיתרון שהתקבל, המשקל שמתאים למאפיין היחיד מקיים $\hat{w} \neq 0$.

כעת, משכפלים את המאפיין היחיד כך שכל דוגמה מעודכנת היא וקטור $\boldsymbol{x}'_i = \begin{bmatrix} x_i \\ x_i \end{bmatrix} \in \mathbb{R}^2$.

לבסוף, פותרים את הבעיה המעודכנת: $\hat{\boldsymbol{u}} = \operatorname{argmin}_{\boldsymbol{u}\in\mathbb{R}^2}\left(\underbrace{\sum_{i=1}^{m}(\boldsymbol{u}^\top \boldsymbol{x}'_i - y_i)^2}_{\triangleq \mathcal{L}'(\boldsymbol{u})} + R(\boldsymbol{u}) \right)$

הבהרה: הפונקציה R מקבלת סקלר או וקטור ומחזירה את הנורמה שלו (בסעיפים א'-ב': L2 בריבוע, בסעיף ג': L1).

א. [6 נק'] כאשר $R(\boldsymbol{v}) = \|\boldsymbol{v}\|_2^2 \triangleq \sum_j (v_j)^2$, נציע כפיתרון לבעיה המעודכנת את הווקטור $\boldsymbol{u} = \begin{bmatrix} \hat{w} \\ 0 \end{bmatrix}$.

הוכיחו שזהו פיתרון לא אופטימלי על ידי הצעת פיתרון אחר $\boldsymbol{z} \in \mathbb{R}^2$ המקיים $\mathcal{L}'(\boldsymbol{u}) + R(\boldsymbol{u}) > \mathcal{L}'(\boldsymbol{z}) + R(\boldsymbol{z})$.

$\boldsymbol{z} = \boxed{\frac{1}{2}\begin{bmatrix} \hat{w} \\ \hat{w} \end{bmatrix} \text{ (notice that } \mathcal{L}'(\boldsymbol{u}) = \mathcal{L}'(\boldsymbol{z}) \text{ and } R(\boldsymbol{u}) > R(\boldsymbol{z}))}$

ב. [6 נק'] כאשר $R(\boldsymbol{v}) = \|\boldsymbol{v}\|_2^2 \triangleq \sum_j (v_j)^2$, סמנו את הטענה הנכונה <u>בהכרח</u> והסבירו <u>בקצרה</u>.

a. אחד המשקלים שווה לאפס (משמע: $\hat{u}_1 = 0 \vee \hat{u}_2 = 0$)

b. <u>שני המשקלים שונים מאפס (משמע: $\hat{u}_1 \neq 0 \wedge \hat{u}_2 \neq 0$)</u>

c. שני המקרים a, b אפשריים

הסבירו בקצרה:

לכל פיתרון שמכיל אפס, קיים פיתרון טוב ממנו (הראינו דוגמה בסעיף הקודם).

ג. [6 נק'] כאשר $R(\boldsymbol{v}) = \|\boldsymbol{v}\|_1 \triangleq \sum_j |v_j|$, סמנו את הטענה הנכונה <u>בהכרח</u> והסבירו <u>בקצרה</u>.

a. אחד המשקלים שווה לאפס (משמע: $\hat{u}_1 = 0 \vee \hat{u}_2 = 0$)

b. שני המשקלים שונים מאפס (משמע: $\hat{u}_1 \neq 0 \wedge \hat{u}_2 \neq 0$)

c. <u>שני המקרים a, b אפשריים</u>

הסבירו בקצרה:

כל עוד $|u_1| + |u_2| = |\hat{w}|$ וגם $u_1 + u_2 = \hat{w}$, יתקבל ערך זהה למינימום של הבעיה המקורית.

שאלה 4 [18 נק']

היזכרו בבעיות האופטימיזציה של ה-SVM במקרה הלא הומוגני:

Soft SVM

$$\underset{\boldsymbol{w}\in\mathbb{R}^d, b\in\mathbb{R}}{\operatorname{argmin}} \left(\frac{1}{m}\sum_{i=1}^{m}\max\{0, 1-y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)\}+\lambda\|\boldsymbol{w}\|_2^2\right)$$

Hard SVM

$$\underset{\boldsymbol{w}\in\mathbb{R}^d, b\in\mathbb{R}}{\operatorname{argmin}} \|\boldsymbol{w}\|_2^2$$

$$\text{s.t. } y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)\geq 1, \ \forall i\in[m]$$

א. [4 נק'] כתבו וקטור שמהווה subgradient לפי $\boldsymbol{w}$ לפונקציית ה-loss של ה-Soft-SVM.

תשובה סופית (לרשותכם עמודי טיוטה בסוף השאלון):

$$\nabla_{\boldsymbol{w}}\left(\frac{1}{m}\sum_{i=1}^{m}\max\{0, 1-y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)\}+\lambda\|\boldsymbol{w}\|_2^2\right) = 2\lambda w+\frac{1}{m}\sum_{i=1}^{m}f\big(y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)\big)y_i\boldsymbol{x}_i$$

כאשר מסמנים

$$f(z)=\begin{cases}-1, & z<1\\ 0, & z\geq 1\end{cases}$$

ב. [6 נק'] לכל אחת מהטענות הבאות, סמנו האם היא (בהכרח) **נכונה / שגויה / לא ניתן לדעת** והסבירו בקצרה.

a. עבור סט אימון פריד ליניארית, **Hard**-SVM מגיע ל-accuracy **train** גבוה מזה של **פרספטרון**.

הטענה: **נכונה** / **שגויה** / **לא ניתן לדעת** (שניהם מגיעים לשגיאה אפס).

b. עבור סט אימון פריד ליניארית, **Soft**-SVM מגיע ל-accuracy **test** גבוה מזה **Hard**-SVM.

הטענה: **נכונה** / **שגויה** / **לא ניתן לדעת** (אין הבטחה כזו ואפשר לחשוב על מקרים נגדיים).

ג. [8 נק'] נתון דאטה פריד ליניארית עם תיוגים בינאריים.

פותרים עליו Hard-SVM ומקבלים כפיתרון $\boldsymbol{w}_1 \in \mathbb{R}^d$ ו-$b_1 \in \mathbb{R}$. ידוע שהפיתרון שהתקבל הוא הפיתרון היחיד לבעיה.

עכשיו פותרים בעיית Hard-SVM מעודכנת עם **margin=2**. משמע, עם אילוצים מעודכנים $y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 2$.

הערה בדיעבד: שימו לב שהשאלה לא עוסקת ב-margin במובן של $\left(\frac{\boldsymbol{w}^\top \boldsymbol{x}_i + b}{\|\boldsymbol{w}\|}\right)$, אלא ב-$(\boldsymbol{w}^\top \boldsymbol{x}_i + b)$, כפי שכתוב במפורש בנוסחה.

את הפיתרון החדש נסמן ב-$\boldsymbol{w}_2 \in \mathbb{R}^d$ ו-$b_2 \in \mathbb{R}$.

בכל אחד מהסעיפים הבאים סמנו את התשובה הנכונה ומלאו את החסר (היכן שנדרש).

הסבר מפורט: הפתרון המקורי מקיים $y_i(\boldsymbol{w}_1^\top \boldsymbol{x}_i + b_1) \geq 1, \forall i$. כל מה שנדרש כדי לקיים את האילוצים החדשים זה להכפיל ב-2 את שני צידי אי-השוויון:

$$y_i\left(\underbrace{(2\boldsymbol{w}_1)^\top}_{\triangleq \boldsymbol{w}_2} \boldsymbol{x}_i + \underbrace{2b_1}_{\triangleq b_2}\right) = 2y_i(\boldsymbol{w}_1^\top \boldsymbol{x}_i + b_1) \geq 2, \qquad \forall i$$

כפל של פתרון בסקלר כמובן לא משנה את אופי הפתרון (=המפריד).

הדבר הזה בדיוק מדגים את הצורך בנירמול של ה-margin כפי שמראים בהרצאה של ה-SVM, כי ניתן לקבוע כל קבוע שרירותי (1, 2 או משהו אחר) ולשנות את הסקאלה של הפתרון בהתאם.

a. ייתכן שלבעיה המעודכנת אין פיתרון, למשל כאשר המרחק בין המחלקות לא מספיק גדול למרות שהדאטה פריד ליניארית. הטענה: **נכונה** / **שגויה**.

b. אם קיים פיתרון לבעיה המעודכנת אזי הפרדיקציות הבינאריות זהות. משמע, $\forall \boldsymbol{x} \in \mathbb{R}^d: h_{\boldsymbol{w}_1, b_1}(\boldsymbol{x}) = h_{\boldsymbol{w}_2, b_2}(\boldsymbol{x})$.

הטענה: **נכונה** / **שגויה**.

c. אם קיים פיתרון לבעיה המעודכנת אזי קיים סקלר α עבורו $\boldsymbol{w}_2 = \alpha \boldsymbol{w}_1$.

הטענה: **נכונה והסקלר שווה ל-**___2___ / **שגויה**.

d. אם קיים פיתרון לבעיה המעודכנת אזי קיים סקלר β עבורו $b_2 = \beta b_1$.

הטענה: **נכונה והסקלר שווה ל-**___2___ / **שגויה**.

חלק ב' – שאלות אמריקאיות [32 נק']

א. [4 נק'] תזכורת: $\text{Precision}=\frac{TP}{TP+FP}$, $\text{Recall}=\frac{TP}{TP+FN}$.

נתון דאטה עם תיוגים בינאריים (P או N, לפחות דוגמה אחת מכל תיוג).

סמנו את **כל** הטענות **השגויות**.

a. ערך ה-Area under the curve (AUC) האופטימלי של עקומת ROC הוא 1

b. מתקיים Precision+Recall=1

c. תמיד קיים מודל שמשיג Recall=1

d. ניתן ליצור Confusion matrices רק בבעיות עם תיוגים בינאריים (שתי מחלקות)

e. בדאטה מאוזן ($\Pr[y=\text{P}]\approx 0.5$), מדדי ה-Precision וה-Recall נהיים שקולים למדד ה-Accuracy

ב. [4 נק'] נרצה להרחיב את מודל ה-k-NN לבעיות **רגרסיה** ($\boldsymbol{x}_i\in\mathbb{R}^d,\ y_i\in\mathbb{R}$).

נגדיר את $\text{NN}(\boldsymbol{x})$ להיות קבוצת האינדקסים של k השכנים הקרובים ביותר ל-$\boldsymbol{x}$.

סמנו את **שתי** פונקציות החיזוי (של y בהינתן $\boldsymbol{x}$) המתאימות ביותר לבעיה.

a. $h(\boldsymbol{x})=\frac{1}{k}\sum_{i\in\text{NN}(\boldsymbol{x})}\boldsymbol{x}_i$

b. $h(\boldsymbol{x})=\frac{1}{k}\sum_{i\in\text{NN}(\boldsymbol{x})}y_i$

c. $h(\boldsymbol{x})=\sum_{i\in\text{NN}(\boldsymbol{x})}(\|\boldsymbol{x}-\boldsymbol{x}_i\|_2\, y_i)$

d. $h(\boldsymbol{x})=\sum_{i\in NN(\boldsymbol{x})}\left(\frac{exp\{-\|\boldsymbol{x}-\boldsymbol{x}_i\|_2\}}{\sum_{j\in NN(\boldsymbol{x})}exp\{-\|\boldsymbol{x}-\boldsymbol{x}_j\|_2\}}\, y_i\right)$

e. $h(\boldsymbol{x})=\sum_{i\in\text{NN}(\boldsymbol{x})}(\max\{0,\ \|\boldsymbol{x}-\boldsymbol{x}_i\|_2^2\}\, y_i)$

ג. [4 נק'] נתונים סט אימון וסט מבחן הנדגמים מהתפלגות כלשהי $\mathcal{D}$ ומחלקת היפותזות כלשהי $\mathcal{H}$.

בוחרים היפותזה מתוך המחלקה ע"י ERM.

עתה מוסיפים לסט האימון מספר דוגמאות חדשות (הנדגמות מ-$\mathcal{D}$), ולומדים היפותזה חדשה ע"י ERM על $\mathcal{H}$.

לכל סוג שגיאה **סמנו** את האפשרות המתאימה בהכרח.

a. שגיאת האימון: **תעלה / תרד / לא תשתנה / לא ניתן לדעת**

b. שגיאת המבחן: **תעלה / תרד / לא תשתנה / לא ניתן לדעת**

ד. [4 נק'] מה התפקיד של פונקציית ה-sigmoid במסווג logistic regression? סמנו את התשובה הנכונה.

a. להכניס non-linearity ולאפשר ללמוד גבולות החלטה לא ליניאריים

b. להפוך את ה-margin להסתברות

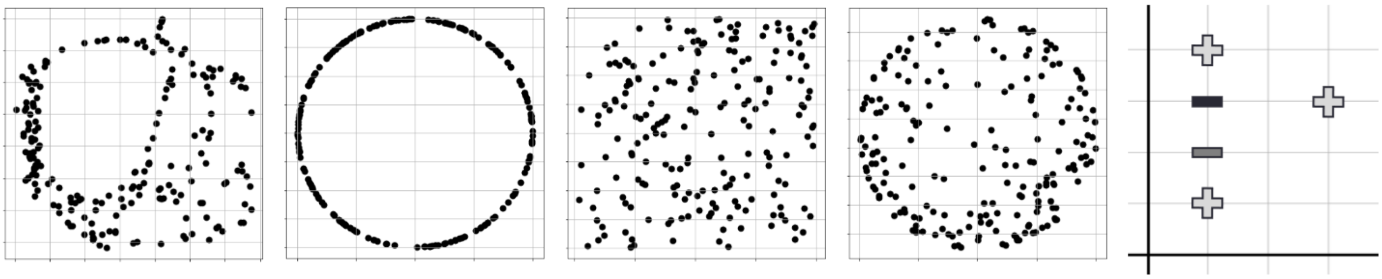
c. לעשות רגרסיה פולינומיאלית בין $\boldsymbol{x}\in\mathbb{R}^d$ ל-$y\in\mathbb{R}$

d. להוסיף רגולריזציה למסווג שנלמד

e. למקסם את האנטרופיה

ה. [4 נק'] נתון דאטה שנדגם מהתפלגות אחידה על ספֵירה (sphere) תלת-ממדית סביב ראשית הצירים.
מפעילים PCA כדי להוריד ממד מ-3 ל-2. הקיפו את האות שמתאימה לתרשים של הדאטה לאחר הורדת הממד.

הדאטה המקורי **(a)** **(b)** **(c)** **(d)**



ו. [4 נק'] נתונה בעיית multiclass.
אימנו מסווג One vs. All (**שיטה א'**) ומסווג multinomial logistic regression (**שיטה ב'**).
לאחר האימון, נוספה לבעיה מחלקה חדשה עם דוגמאות חדשות.
על מנת להתמודד עם המחלקה החדשה (סמנו את הטענה הנכונה ביותר):

a. בשיטה א' אנחנו נדרשים לאמן רק מסווג אחד נוסף ואין שינוי במסווגים שכבר אומנו
b. בשיטה ב' אנחנו נדרשים לאמן רק מסווג אחד נוסף ואין שינוי במסווגים שכבר אומנו
c. שתי הטענות a+b נכונות
d. כל הטענות הקודמות שגויות

הערה: חלק מהסטודנטים הבינו שכל המסווגים בשיטה ב' נחשבים מסווג אחד, ולכן התקבלה גם תשובה b

ז. [4 נק'] נסתכל על רשת Multi-Layer Perceptron לרגרסיה שלומדת פונקציה $h: \mathbb{R}^d \to \mathbb{R}$.
כפונקציית אקטיבציה נשתמש בפונקציית הזהות $\sigma(z) = z$. לסיום, נשתמש ב-loss ריבועי (MSE) ללא רגולריזציה.
סמנו את הטענה הנכונה.

a. במחלקת מודלים, ה-capacity שקול לזה של מודלים ליניאריים (linear regression)
b. השכבה האחרונה ברשת צריכה להיות שכבת Softmax
c. לבעיית האופטימיזציה יש מינימום גלובאלי יחיד
d. כל הטענות הקודמות שגויות

ח. [4 נק'] נתונות דוגמאות $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m \in \mathbb{R}^d$ מתוייגות באופן בינארי, משמע $y_1, \dots, y_m \in \{-1, +1\}$.
רוצים ללמוד הורדת ממד ליניארית (בעזרת מטריצה $\mathbf{W}$) ל-k מימדים כאשר $k \ll d$.
רוצים שלאחר ההטלה, דוגמאות מתיוגים זהים תהיינה קרובות אחת לשנייה ורחוקות מהדוגמאות של התיוג השני.
סמנו את בעיית האופטימיזציה (היחידה) שמתאימה לפיתרון הבעיה.

a. $\underset{\mathbf{W} \in \mathbb{R}^{k \times d}}{\operatorname{argmin}} \left(\sum_{i,j} \left\| \mathbf{W}\boldsymbol{x}_i - \mathbf{W}\boldsymbol{x}_j \right\|_2^2 \right)$

b. $\underset{\mathbf{W} \in \mathbb{R}^{k \times d}}{\operatorname{argmin}} \left(\sum_{i,j} \left\| \mathbf{W}\boldsymbol{x}_i - \mathbf{W}\boldsymbol{x}_j \right\|_2^2 + \sum_{i,j: y_i \neq y_j} \left\| y_i - y_j \right\|_2^2 \right)$

c. $\underset{\mathbf{W} \in \mathbb{R}^{k \times d}}{\operatorname{argmin}} \left(\sum_{i,j: y_i = y_j} \left\| \mathbf{W}\boldsymbol{x}_i - \mathbf{W}\boldsymbol{x}_j \right\|_2^2 + \sum_{i,j: y_i \neq y_j} \max\left\{ 0, 1 - \left\| \mathbf{W}\boldsymbol{x}_i - \mathbf{W}\boldsymbol{x}_j \right\|_2^2 \right\} \right)$

d. $\underset{\mathbf{W}\in\mathbb{R}^{k\times d}}{\operatorname{argmin}}\left(\sum_{i,j:y_i\neq y_j}\ln\left(1+\left\|\mathbf{W}\boldsymbol{x}_i-\mathbf{W}\boldsymbol{x}_j\right\|_2^2\right)+\sum_{i,j}\left(\mathrm{W}_{i,j}\right)^2\right)$