מבוא למערכות לומדות (236756)

סמסטר אביב תשע"ו

מבחן מסכם מועד ב' , 21 בספטמבר 2016

מספר סטודנט:

משך המבחן: 3 שעות.

חומר עזר: אין להשתמש בכל חומר עזר. בעמוד הבא לרשותכם דף נוסחאות והגדרות.

הנחיות כלליות:

- המבחן כתוב בלשון זכר ומיועד לנשים ולגברים כאחד.
- מלאו את הפרטים בראש דף זה ובדף השער המצורף, בעט בלבד.
- במבחן 18 דפים ממוספרים סהכ, כולל עמוד זה שמספרו 1. ודאו שיש לכם כל הדפים.
- במבחן 5 שאלות. יש לענות על **כל** השאלות.
- משך המבחן 3 שעות (180 דקות).
- כל התשובות יכתבו על טופס הבחינה, **ויש להחזירו בתום הבחינה**.
- אנא כתבו בכתב יד קריא וברור. **תשובה בכתב יד שאינו קריא לא תיבדק.**
- נא לא לתלוש עמודים ממחברת הבחינה.
- נא לכתוב רק את מה שהתבקשתם ולצרף הסברים קצרים רק כפי שמבוקש בשאלה—אין צורך בהסברים או פרטים נוספים על אלו שהתבקשתם במפורש.

כל המוסיף גורע

1. המילה העברית ל-feature היא תכונה או מאפיין. המילה העברית ל-label היא תיוג.

בהצלחה!

דף נוסחאות

1. אלא אם כן מצוין אחרת, מרחב התיוגים $\mathcal{Y}$ הוא $\{-1,+1\}$.
2. $\binom{n}{k} \leq n^k$
3. כפי שהיה מקובל בקורס, סוגריים מרובעים [..] מסמנים קואורדינטה של וקטור. למשל x[3] היא הקואורדינטה השלישית של וקטור x כלשהי.
4. שגיאת הכללה = true error = $L_{\mathcal{D}}^{01}$.
5. שגיאה אמפירית = L_S^{01} (ממוצע השגיאות על מדגם)
6. שגיאת אסטימציה = $L_{\mathcal{D}}^{01} - L_S^{01}$
7. $e \approx 2.72$
8. התפלגות גאומטרית עם פרמטר p

$$P(x=n) = p \quad (1-p)^{n-1}$$

9. תהא $\mathcal{H}$ מחלקת היפותזות של בעיית למידה כלשהי, ו- S קבוצת אימון שנבחרת באקראי. נסמן

$$\hat{h} = argmin_{h\in\mathcal{H}} L_D^{01}(h)$$
$$h^* = argmin_{h\in\mathcal{H}} L_S^{01}(h)$$

אזי, לכל $\delta > 0$: בהסתברות של לפחות $1-\delta$ מתקיים:

$$L_D^{01}(\hat{h}) \leq L_D^{01}(h^*) + O\left(\sqrt{\frac{VCDIM(\mathcal{H}) + \log 1/\delta}{|S|}}\right)$$

ובמקרה ה- Realizable (שמשמעותו $L_D^{01}(h^*) = 0$),

$$L_D^{01}(\hat{h}) \leq O\left(\frac{VCDIM\left(\frac{\mathcal{H}}{\delta}\right)}{|S|}\right)$$

חלק א : שאלות קצרות (27 נק') – ניקוד שווה לכל השאלות

1. נתונה סידרה של מחלקות היפוזות סופיות ... $\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3$ על אותו מרחב דוגמאות אינסופי $\mathcal{X}$. נתון שלכל i המחלקה $\mathcal{H}_i$ היא לפחות בגודל $2^{(2^{(2^i)})}$. אז $\mathcal{H}_i$ אינה efficiently, properly PAC learnable
אבל יתכן שהיא efficiently, improperly PAC learnable.

הטענה האחרונה הינה (יש לסמן אפשרות אחת):

☐ **אמת**

☐ **שקר** דוגמא נגדית (חובה לספק במקרה שסימנתם "שקר"):

אינטרוולים בקטע [0,1] שהקצוות שלהם שייכות לקבוצה סופית בגודל $2^{(2^{(2^i)})}$ של נקודות בתוך [0,1]. זו קבוצת היפותזות סופית אבל היא efficiently properly PAC learnable ולא כפי שכתוב

2. נניח שקבוצת אימון $(x_1,y_1),\dots,(x_m,y_m), x_i \in \mathbb{R}^d, y_i \in \{\pm 1\}$ היא ניתנת להפרדה לינארית, ואנו מריצים מינימיזציה של פונקציית מטרה כלשהי באמצעות stochastic gradient descent על משתנה המסווג $w \in \mathbb{R}^d$ באופן שבכל איטרציה בוחרים נקודת אימון $i \in \{1..m\}$ באופן אקראי ואוניפורמי ומשתמשים בהערכת גרדיאנט הסטוכסטי (stochastic gradient):

$$-[2 - y_i\langle w, x_i\rangle]_+ \cdot y_i \cdot x_i$$

אז בחירה זהירה של גודל הצעד ומספר איטרציה סופי, בסיכוי גבוה, יביא מסווג שמפריד את קבוצת האימון באופן לינארי.

הערה: הכוונה היא שבכל צעד מעדכנים $w \leftarrow w + \eta[2 - y_i\langle w, x_i\rangle]_+ \cdot y_i \cdot x_i$ עבור η מספיק קטין.

הטענה האחרונה הינה (יש לסמן אפשרות אחת):

☒ **אמת**

☐ **שקר**

הסבר : ה(תת)-גרדיאנט הסטוכסטי מתאים לפונקציית המטרה $\sum_{i=1}^{m}[2 - y_i\langle w, x_i\rangle]_+$. פונקציה זו קמורה, לא-שלילית, עבור w מספיק גדול נותנת ערך 0, וכאשר היא נותנת ערך 0 אז היא מפרידה לינארית את הנקודות.

2. ענו בקצרה: מהו ה- ~~Representer Theorem~~? מדוע הוא שימושי בהקשר של SVM?

המשפט אומר שהפיתרון האופטימלי של Soft-SVM (* וכן של למשפחה רחבה של בעיות אופטימיזציה אחרות) הוא צירוף לינארי של נקודות הקלט $x_1 \dots x_m$. מסקנה זו מאפשרת להחליף את משתני האופטימיזציה המקוריים (המקדמים של וקטור ההפרדה w) במשתנים $\alpha_1 .. \alpha_m$, שמבטאים את התרומה הלינארית של x_i בייצוג של w. לאחר החלפת המשתנים מתברר ש- SVM מוגדר באופן מוחלט ע"י מטריצת ה Gram של הקלט (יחד עם וקטור הסיווגים y), מה ש"פותח את הדלת" לשימוש בקרנלים.

3. חיתוך של שתי קבוצות קמורות (ב $\mathbb{R}^d$) היא קבוצה קמורה

☒ **אמת**

☐ **שקר** דוגמא נגדית (חובה לספק במקרה שסימנתם "שקר"):

4. פונקציית המטרה של Soft-SVM היא $\lambda \cdot L_S^{hinge}(w) + \|w\|^2$

ככל שמגדילים את λ, כך צפוי ששגיאת ההכללה (יש לסמן אפשרות אחת):

☐ **תרד**

☒ **תעלה**

☐ **תישאר ללא שינוי**

הסבר: הגדלה של λ מקטינה את התרומה של הרגולריזציה, ולכן צפויה להגדיל את שגיאת ההכללה.

5. טכניקות Feature Selection נועדו כדי לשפר את יכולת ההכללה (generalization), לשפר את יעילות החישוב, את משאבי התיקשורת והזיכרון.

הטענה האחרונה הינה (יש לסמן אפשרות אחת):

☒ **אמת**

☐ **שקר**

6. הסבירו מהו Feature Selection מסוג Greedy, ותנו דוגמא שנילמדה בכיתה לשיטה מסוג זה

שיטת greedy feature selection מגדילה את קבוצת המאפיינים באופן הדרגתי, כאשר בכל צעד מוסיפים מאפיין יחיד לקבוצה מהצעד הקודם באופן greedy (חמדני). לדוגמא, אפשר להוסיף את המאפיין שיחד עם הקבוצה מהצעד הקודם מביא למינימום את שגיאת האימון.

חלק ב: איחוד של מקטעים (30 נק)

<u>נקודות מתחלקות אחיד בין הסעיפים, ובתוך כל סעיף אחיד בין תתי-הסעיפים וכו</u>

נתונה בעיית למידה שבה $\mathcal{X} = [0,1]$, כלומר האינטרוול הממשי בין 0 ל- 1, ו- $\mathcal{D}$ היא ההתפלגות האוניפורמית הסטנדרטית על האינטרוול. לכל מספר טבעי k מוגדרת מחלקת היפותזות $\mathcal{H}_k$ כאיחוד של לכל היותר k אינטרוולים פתוחים. למשל, $\mathcal{H}_1$ היא מחלקת ההיפותזות שמוגדרת ע"י

$$\mathcal{H}_1 = \{h_{a,b} | 0 \le a \le b \le 1\}, h_{a,b}(x) = \begin{cases} 1 & a < x < b \\ -1 & otherwise \end{cases}$$

$$\mathcal{H}_2 = \{h_{a,b,c,d} | a \le b \le c \le d\}$$

$$h_{a,b,c,d}(x) = \begin{cases} 1 & a < x < b \ or \ c < x < d \\ -1 & otherwise \end{cases}$$

וכו'

1. מהו מימד ה- VC של $\mathcal{H}_k$? תנו הסבר קצר, הכולל קבוצה בגודל המתאים שאפשר לנתץ, והסבר (קצר) מדוע לא ניתן לנתץ קבוצה יותר גדולה.

מימד ה VC הוא 2k

ניתן לנתץ 2k נקודות (לא חשוב אילו) פשוט כי כל סיווג של 2k נקודות מכיל לכל היותר k רצפים של +1 (כשעוברים על הנקודות משמאל לימין, למשל)

לעומת זאת, 2k+1 נקודות מסווגות לסרוגין +1,-1,+1,...-1,+1 דורשות k+1 רצפים של +1, מה שלא ניתן להשיג עם $\mathcal{H}_k$/

2. נאדר החליט ללמוד באמצעות Nearest-Neighbors. במילים אחרות, בהינתן קבוצת אימון $(x_1, y_1) \dots (x_m, y_m)$ הוא מחזיר את פונקציית הפרדיקציה g המוגדרת ע"י:

$$g(x) = y_{i(x)}, where\ i(x) \in argmin_{j=1..m}|x - x_j|$$

א. הניחו את המחלקה $\mathcal{H}_1$ ואת המקרה ה- realizable, כלומר קיים $0 \le a \le b \le 1$ כך שלפונקציה $h_{a,b}$ יש שגיאה 0. אם נאדר דוגם נקודה יחידה (m=1), מה תוחלת שגיאת ההכללה שלו? שימו לב: יש להביע את תוחלת השגיאה כפונקצייה של a, b. כמו כן, התוחלת היא ביחס להתפלגות על נקודת האימון הבודדת.

תוחלת שגיאת ההכללה היא $2(b-a)(1-b+a)$ הסבר: בסיכוי (b-a) נקודות האימון בתוך האינטרוול [a,b] ובמקרה זה נאדר נותן סיווג 1+ לכל הקבוצה [0,1]. הוא טועה על הנקודות מחוץ לאינטרוול [a,b] שההסתברות שלהן היא $(1-b+a)$, וזה גם הביטוי לשגיאת ההכללה שלו. בסיכוי (1-b+a) הוא מתאמן על נקודה מחוץ לאינטרוול [a,b] ואז שגיאת ההכללה שלו היא בידיוק הסיכוי שהנקודה הבאה נופלת באינטרוול *[a,b]* כלומר *(b-a)*. מכאן שתוחלת שגיאת ההכללה היא כפי שנכתב.

ב. לנאדר מידע מוקדם ש $a > 0.1$. כדי לנצל פיסת אינפורמציה זו, הוא החליט להוסיף באופן מלאכותי לקבוצת האימון את הזוג $(x_0, y_0) = (0.1, -1)$.
בתנאים של הסעיף הקודם (כלומר: nearest neighbors אבל עכשיו עם שתי נקודות אימון (x_0, y_0) יחד עם (x_1, y_1) שמוגרלת כבסעיף הקודם), האם האסטרטגיה של נאדר מורידה תמיד את תוחלת שגיאת ההכללה, או שמא יש מקרים ספציפיים שבהם האסטרטגיה החדשה דווקא מעלה אותה? אם בחרתם באפשרות הראשונה אנא רישמו הסבר קצר (אך משכנע) לתשובתכם. אם בחרתם באפשרות השנייה, מיצאו דוגמא ספציפית לערכים של a,b והראו שנאדר מפסיד.

הרעיון של נאדר יכול במקרים מסויימים להזיק. למשל, אם ההיפותזה האופטימלית היא $h_{a,b}$ כאשר a=0.1 ו- b=1 אז תוחלת שגיאת ההכללה ללא שימוש בנקודה x_0 היא $2 \cdot 0.9 \cdot 0.1 = 0.18$ (כפי שחושב בסעיף הקודם).

אם הוא מצרף את הנקודה x_0 אז תוחלת שגיאת ההכללה גדולה יותר. להלן דרך להראות זאת:

אם נקודת האימון נופלת באינטרוול [0,0.1] (בסיכוי 0.1) אז שגיאת ההכללה היא 0.9. התרומה לתוחלת שגיאת ההכללה: 0.09.

אם נקודת האימון נופלת מימין ל 0.6 (קורה בסיכוי 0.4), אז המסווג של נאדר טועה (בין השאר) באינטרוול (0.1,0.35] שההסתברות שלו 0.25. התרומה לתוחלת שגיאת ההכללה לפחות $0.6 \cdot 0.25 = 0.15$.

סה"כ קיבלנו תוחלת שגיאת הכללה לפחות $0.09 + 0.15 = 0.24$ כלומר עדיף במקרה זה להתעלם מ x_0.

ג. <u>טענה:</u> לכל k, במקרה ה realizable ביחס למחלקה $\mathcal{H}_k$, ולכל $\varepsilon > 0$, אם נאדר יפעיל את אלגוריתם ה nearest-neighbors עם מספר דוגמאות m שמקיים

$$m \geq \frac{c \cdot VC(\mathcal{H}_k)}{\varepsilon}$$

עבור קבוע אוניברסאלי כלשהו c, אז בסיכוי לפחות 0.99 שגיאת ההכללה שלו תהייה לכל היותר ε.

האם הטענה נכונה או לא? כיתבו את תשובתכם במסגרת, נמקו בקצרה את תשובתכם.

כאן יש לשים לב שבבעייה זו אלגוריתם nearest neighbor הוא גם אלגוריתם CONSISTENT, ולכן חלים משפטי הכללה PAC למקרה ה Realizable. בגלל טעות בניסוח השאלה היה אמור להיות כתוב $m/\log m \geq \frac{c \cdot VC(\mathcal{H}_k)}{\varepsilon}$, ולכן קיבלנו גם תשובה "לא" וגם תשובה "כן" כל עוד כתבתם את הנימוק לעיל).

3. גם לנטאשה יש אינפורמציה נוספת, והיא גם עובדת בתנאים של Realizabile.
היא יודעת שהפרדיקטור האופטימלי בא מתוך $\mathcal{H}_1$. נסמן אותו $h_{a,b}$. בנוסף, יש
לה מידע בייזיאני (Bayesian): היא יודעת ש- b=1 ושהפרמטר a מתפלג
אוניפורמי בקטע [0,1], ושנקודות המדגם מתפלגות באופן בלתי תלוי ב a. היא
מחליטה להשתמש בשיטת MAP כדי לעדכן את הידע על הפרמטר a בהינתן
קבוצת אימון. בהינתן נקודת אימון בודדת (x_1, y_1) (שמתפלגת מ $\mathcal{D}$ לעיל) כיצד
עליה לעדכן את ההתפלגות על הפרמטר a?
(אם הינכם מתקשים לעבוד עם התפלגות רציפה, מותר לכם להניח ש $\mathcal{X} =$
$\{0.01, 0.02, \ldots 1\}$, ההתפלגות $\mathcal{D}$ היא אוניפורמית על $\mathcal{X}$ והפרמטר a מתפלג
אוניפורמי על $\mathcal{X}$ גם כן.) שימו לב: אין צורך לתת הוכחה, רק לתאר את
ההתפלגות המעודכנת של a.

אם $y_1 = -1$ אז ההתפלגות המעודכנת היא אוניפורמית על $[x_1, 1]$. אם $y_1 = 1$ אז ההתפלגות אוניפורמית על $[0, x_1]$.

4. אביגדור העצלן הוריד פעם תוכנת SVM חינמית מהאינטרנט, ולכן החליט להשתמש בה. כמו כן הוא שמע שקרנלים פולינומיאלים (Polynomial Kernels) יכולים לעזור במיקרים כאלה .

א. באיזה דרגת פולינום כדאי לאביגדור לבחור, בהנחה של מחלקת היפותזות $\mathcal{H}_k$?
(לצורך סעיף זה, הניחו את המקרה ה Realizable ומיצאו את הדרגה שיכולה תמיד לתת שגיאת אימון 0). אין צורך לתת הסבר.

דרגת הפולינום צריכה להיות 2k, מכיוון שאם רוצים ליצור k אינטרוולים ע"י לקיחת סימן של פולינום במשתנה אחד אז לפולינום יהיו בהכרח 2k אפסים. לכן בוודאי שלא פחות מ 2k.

הערה: הטיעון הנ"ל מספיק כדי לקבל ניקוד מלא על השאלה

בנוסף, לכל בחירה של p אפסים של פולינום, נסמנם $z_1 \dots z_p$ ניתן למצוא פולינום שהסימן שלו הוא בידיוק k אינטרוואלים: $-(x-z_1)(x-z_2)\dots(x-z_p)$

ב. לרוע המדל, גירסת החינמי של תוכנת ה- SVM לא כוללת תמיכה בקרנלים פולינומיאליים, ואביגדור (הקמצן) לא מוכן לשלם על שידרוג. כיתבו פסאודוקוד שהופך את קבוצת האימון $(x_1,y_1),\dots,(x_m,y_m)$ לקבוצת אימון אחרת $(z_1,y_1)\dots(z_m,y_m)$ כך ש $z_i \in \mathbb{R}^d$ לכל i, עבור מימד d כלשהו (לבחירתכם), וכך שהפלט של ה SVM ללא קרנל על קבוצת האימון החדשה יהיה שקול לפלט של ה- SVM עם הקרנל המתאים על קבוצת האימון המקורית.

שימו לב: לצורך סעיף זה, התעלמו בגורם הרגולריזציה של SVM. כמו כן, הכוונה ב"ללא קרנל" היא שימוש בקרנל הטריוויאלי (ידוע גם בכינוי "לינארי"), כלומר מכפלה פנימית רגילה $\langle z_i, z_j\rangle$.

אפשר לקחת: $d = 2k$ ו-

$$z_i[j] = x_i^j$$

5. לג'קי יש הגירסה המלאה של תוכנת ה- SVM, אבל הוא לא שמע (ולא מוכן לשמוע) על קרנלים. למזלו, הוא מכיר את AdaBoost. הוא בדק וגילה שלכל k ולכל $h \in \mathcal{H}_k$ קיים <u>מפריד לינארי</u> ב $\mathbb{R}^1$ שנותן שגיאת הכללה לכל היותר $\frac{1}{2} - \alpha_k$ ביחס ל h וביחס ל $\mathcal{D}$ הספציפי לעיל, עבור מספר גלובלי כלשהו $\alpha_k > 0$ (במקרה ה-Realizable).

א. האם התנאים הנ"ל מספיקים כדי להצדיק (תאורטית) שימוש ב Adaboost במקרה ה-realizable? אם כן, רשימו "כן". אם לא, רישמו מה התנאי החסר.

לא.

צריך להתקיים ששגיאת ההכללה לכל היות $\frac{1}{2} - \alpha_k$ ביחס לכל $h \in \mathcal{H}_k$ וביחס <u>לכל</u> $\mathcal{D}$, ולא רק עבור ה $\mathcal{D}$ הספציפי שהוגדר בשאלה.

ב. ג'קי כתב קוד ל- adaboost בעצמו. לרוע המזל, הוא שכח אם לשים סימן פלוס ('+') או מינוס ('-') במקומות המסומנים בחצים בקוד מטה. אנא עיזרו לו להשלים את החסר ע"י כתיבה של סימן '+' או '–', (יש לבחור סימן אחד בכל משבצת)

Input:

Training set $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$

Weak Learner A, which will be applied to *distributions D over S*

Initialize: $D^{(1)} = \left(\frac{1}{m}, \frac{1}{m}, \dots, \frac{1}{m}\right)$

For t=1, …, T: *// T to be determined later*

$$h_t = A(D^{(t)})$$
$$\epsilon_t = L_{D^{(t)}}(h_t) = \frac{1}{m}\sum_i D_i^{(t)} \cdot [[h_t(x_i) \neq y_i]]$$
$$w_t = \frac{1}{2}\log\left(\frac{1}{\epsilon_t} - 1\right)$$
$$D_i^{(t+1)} = \frac{D_i^{(t)} \exp(-w_t y_i h_t(x_i))}{\sum_j D_j^{(t)} \exp\left(-w_t y_j h_t(x_j)\right)}$$

Output: $h_s(x) = sign(\sum_{t=1}^{T} w_t h_t(x))$

הסבר: יש להגדיל את המשקל של הנקודות שעליהן אנו טועים. כמו כן יש לדאוג לקבל פונקציית התפלגות על הנקודות, לכן הנירמול במכנה הוא זהה.

~~חלק~~ ג: מודלים הסתברותיים (27 נק)

1. שחר ודני מנסים לחזות את כמות המשקעים בחיפה בעונת החורף הקרובה באמצעות המודל ההסתברותי Maximum a posteriori (MAP) . המודל מניח כי כמות המשקעים בעונה מתפלגת נורמלית עם ממוצע μ וסטיית תקן 10. שחר משתמש בהסתברות ה-prior הבאה: μ מתפלג נורמלית עם ממוצע 300 וסטיית תקן 40. דני משתמש בהסתברות prior נורמלית עם ממוצע 300 וסטיית תקן 5.

א. לאחר אימון המודלים (על אותה קבוצת אימון) ועידכון התפלגות המשערך ע"י שחר ודני, אחד קיבל התפלגות חדשה עם תוחלת 323.5 והשני 315.8.
איזו תוחלת שייכת לדני ואיזו לשחר?

ב. לאור מחקרים אחרונים שחר הבין כי המודל של כמות המשקעים נכון, אבל ה prior הנורמלי על הפרמטר μ בעייתי. עם זאת אין לו חלופה טובה אחרת.
האם הגדלת כמות דוגמאות האימון אמורה לסייע לשחר או לא? הסבר.

ג. (אין קשר בין סעיף זה לסעיפים הקודמים) נתונה בעיית פרדיקציה שעבורה נתון $\mathcal{X} = \{0,1,2,\dots\}^2$ ו- $\mathcal{Y} = \{0,1\}$. הוחלט לפתור אותה באמצעות שיטת בייז נאיבית (Naive Bayes), תחת ההנחה שלכל $i = 1,2$: $x[i]$ מתפלג גאומטרית עם פרמטר 0.2 בהינתן ש $y = 0$, ואחרת $x[i]$ מתפלג גאומטרית עם פרמטר 0.8

I. תחת ההנחה בשאלה זו, חשבו את ההסתברות הבאה

$$\Pr[(x[1] = 2 \wedge x[2] = 3)|\ y = 1] = \left[\qquad\qquad\qquad\right]$$

II. נתונה קבוצת האימון הבאה

$$(x_1 = (2,5), y_1 = 0), (x_2 = (1,3), y_2 = 1)$$

איזו פרדיקציה מתאימה ל- $x = (1,4)$ אם מניחים ש- $\Pr(y = 1) = \Pr(y = 0) = 0.5$? הראו את החישובים בפתרון.

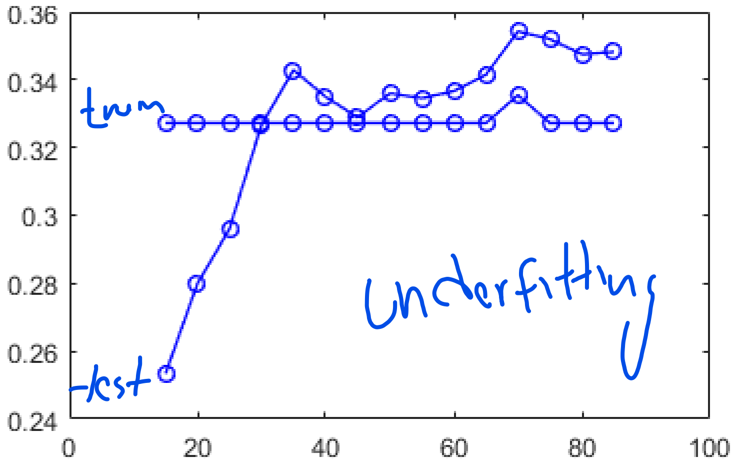
פתרון:

חלק ד: ניתוח פלט של scipy.sklearn.SVM על בעייה אמיתית (16 נק)

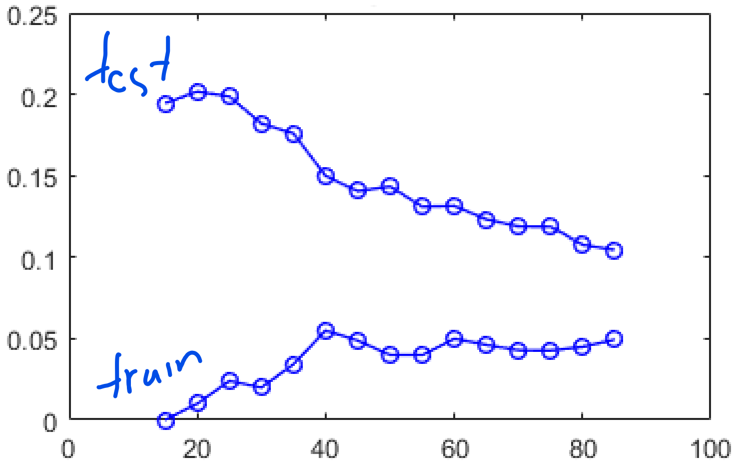
כרמלה מנסה ללמוד באמצעות Kernel-SVM פרדיקטור שבהינתן זוג קוורדינטות $(x[1],x[2]) \in \mathcal{X}=[0,1]^2$ המזהה פיקסל בתמונה, אם צבע הפיקסל המתאים בתמונה להלן הוא שחור (1) או לבן (0). ההתפלגות $\mathcal{D}$ היא אחידה על פני כל הפיקסלים בתמונה.

היא השתמשה בסיפריית SVM עם קרנל פולינומי מדרגה $d=2,10,20$. (לצורך השלמת הפרטים של הניסוי, יש לציין שהיא השאירה את שאר הפרמטרים בערך ה default של הסיפריה, בפרט, מקדם הרגולריזציה C=1 והפרמטר ה"חופשי" של קרנל הפולינום $\gamma=1$). עבור כל דרגה היא הגדילה את גודל קבוצת האימון m מ 15 ל- 85 בדילוגים של 5. עבור כל m וכל d היא הריצה SVM עם קרנל פולינומי מדרגה d וקבוצת אימון בגודל m שלוש פעמים, וחישבה את הממוצע של שגיאת האימון ושגיאת ההכללה. (שגיאת ההכללה מוגדרת כאחוז השגיאה של המסווג על כלל הפיקסלים בתמונה לפי ההתפלגות האחידה על הפיקסלים של התממונה). על התוצאות היא דיווחה בגרפים לפניכם. גרף אחד מתאים ל d=2, אחד ל d=10 ואחד ל d=20. בכל גרף, קו אחד מתאים לממוצע שגיאת ההכללה, והשני לממוצע שגיאת האימון (על פני שלוש הרצות,

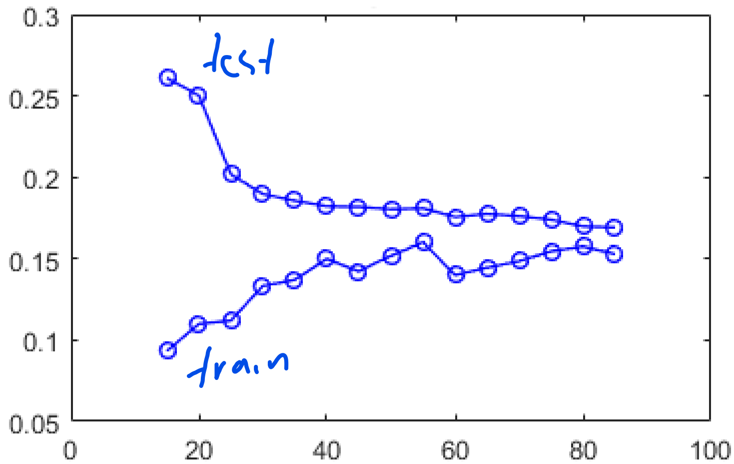
כאמור). לרוע המזל היא לא סימנה איזה d מתאים לכל גרף, וגם שכחה ליצבוע את שני הקווים בכל גרף בצבעים שונים. עליכם להתאים לכל גרף את הערך המתאים של d.



_d=_**2**___



__d=__**20**_______



__d=__ _**10**____