בית ספר "אפי ארזי" למדעי המחשב המרכז הבינתחומי
The Efi Arazi school of computer science
The Interdisciplinary Center

סמסטר ב׳ תשע"ז
Spring 2017

פתרון מבחן מועד ב בלמידה ממוכנת 2017

תשובה 1

א. היתרון בשיטת instance based learning הוא שאין צורך לרוב בלמידה כלל אלא פשוט שומרים את קבוצת האימון ומבצעים כלל החלטה בעזרתה. לעיתים זה נקרא lazy learning. החיסרון הוא שעלינו לשמור את כל קבוצת האימון כדי לסווג (סיבוכיות מקום גבוהה) וכן מסווגים כאלה נוטים ל-overfitting.

ב. כלל ההחלטה של KNN בסיווג הוא לחפש את קבוצת K השכנים (המופעים הקרובים ביותר) למופע שאותו מסווגים וקובעים את הסיווג לפי הרוב בקבוצה זו.

ג. כלל ההחלטה של KNN ברגרסיה הוא לחפש את קבוצת K השכנים (המופעים הקרובים ביותר) למיקום בו רוצים לחשב את ערך הפונקציה ואז מחשבים ממוצע כלשהו של ערכי הפונקציה של הקבוצה (ע"פ רוב ממוצע משוקלל לפי מרחק מהמיקום).

ד. ניתן לשמור את המופעים במבנה נתונים של חיפוש רב מימדי כמו KD-tree. עץ כזה נבנה ברקורסיה. בכל צמות מחלקים את המרחב לשניים בעזרת ערך סף (threshold) במימד אחד שהוא הערך האמצעי של המופעים שהיו בצומת ומחלקים את המופעים לשני צמתי הבנים כך מספר המופעים בכל בן תהיה שווה (פחות או יותר). בכל רמה בעץ בוחרים מימד אחר לחלק לפיו וממשיכים באותו סדר ברקורסיה עד שיש מספר קטן כרצוננו בכל עלה (עד מופע בודד).
כדי למצא את השכן בכל צומת נסווג אותו לתת עץ לפי ערך הסף בצומת עד שנגיע לעלה ואז נעבור על כל המופעים למצא את המופע הקרוב ביותר.

ה. ניתן לבחור את K לפי validation set: מחלקים את קבוצת האימון המקורית לשני חלקים זרים שאיחודם הקבוצה המקורית: קבוצת בעזרתה מסווגים TC וקבוצה שעליה מעריכים את הטעות TV. מאחר ואין פרוצדורת למידה יש לסווג את המופעים בקבוצת ה-validation TV בעזרת הקבוצה TC. בוחנים מספר ערכים של K ועבור כל ערך של K סופרים את מספר הטעויות שנעשו בסיווג המופעים ב-TV ובוחרים את K שעבורו מספר הטעויות קטן ביותר.

ו. הוצאת הדוגמאות שמסווגות נכון נקראת filtering ומטרתה להקטין את מספר המופעים שיש לשמור כדי להקטין את סיבוכיות המקום. שיטה זו מתאימה לקבוצת אימון גדולה שאין בה הרבה רעש.

ז. הוצאת הדוגמאות שמסווגות לא נכון נקראת noise removal ומטרתה להוציא מופעים שהם רעש מקבוצת האימון. שיטה זו מתאימה לקבוצת אימון שאנו מאמינים שיש בה רעש וטעויות.

תשובה 2

א. מכיוון שלא אפשרי להגיע למספר דגימות למידה כרצוננו (משיקולים של הבאת המידע, זמן למידה, וכד'..), אנו צריכים לדעת את המינימום של דגימות הנדרש כדי להשיג את 'המטרה'. המינימום הזה נקבע ע"פ גודל הטעות שאנו מוכנים לספוג וע"פ הביטחון שאנחנו רוצים שזה אכן יהיה גודל הטעות.

ב. ε – החלק של הקירוב ב-PAC והוא מייצג את הטעות האמיתית שאנו שואפים אליה לאחר הלימוד.

δ – החלק של ההסתברות ב-PAC $(1-\delta)$ המייצג את הוודאות שאנו שואפים אליה שהטעות תהיה קטנה מ-ε.

ככל שנרצה טעות קטנה יותר, אז נצטרך יותר דגימות שיאפשרו לנו ללמוד מודל שקרוב יותר למודל האמיתי. ולכן כאשר מקטינים את הטעות הרצויה, אז צריך להגדיל את כמות הדגימות. אותו הסבר נכון גם לוודאות δ, כאשר ה-δ מייצגת למעשה את חוסר הוודאות (ו-$1-\delta$ מייצג את הוודאות) ולכן גם פה, כאשר נרצה וודאות גדולה יותר (או חוסר ודאות קטנה יותר) נצטרך להגדיל את כמות הדגימות.

ג. עץ החלטה בעומק 2 יכול המקסימום לבצע 'מבחן' על 3 תכונות (אחד בשורש ושניים בשני הילדים), לכן קונספט המטרה אינו נמצא במרחב ההיפותזות שלנו. אם נניח כי ההגבלה היחידה היא שאין מסלול בעץ שבודק את אותה תכונה פעמיים, אז גודל מרחב ההיפותזות שלנו יהיה 80=(4*4)*5 (5 אפשרויות לשורש ולכל אחד מהבנים 4 אפשרויות ולכן סה"כ יש 16 צירופי תכונות שניתנות להיבחר בצמתי הבנים).

נציב בנוסחה השנייה, מכיוון שמרחב ההיפותזות סופי, אך קונספט המטרה לא נמצא במרחב ההיפותזות שלנו:

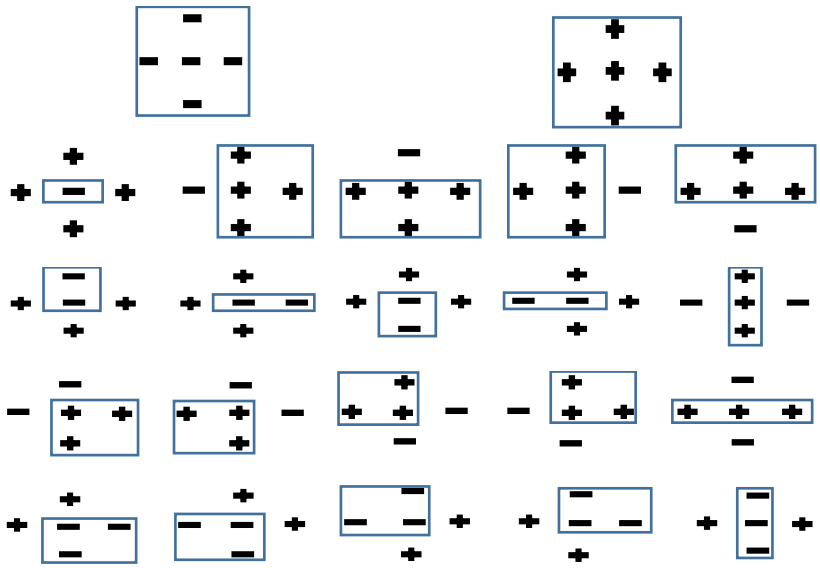
$$m \geq \frac{1}{2(0.1)^2}\left(\ln 80 + \ln \tfrac{1}{0.1}\right)$$

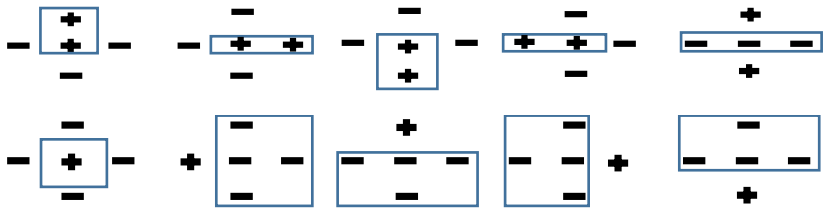
ד. גודל מרחב ההיפותזות שלנו זהה לסעיף ג' – 80, אך הפעם קונספט המטרה נמצא בתוך המרחב, כאשר השורש בודק את התכונה X_1, ושני הילדים בודקים את התכונות X_4, X_3 (הצומת שבודקת את X_4 היא הצומת שקיבלה את הדוגמאות להן $X_1 = 1$). מכיוון שהמרחב סופי והקונספט נמצא בתוכו אזי נשתמש בנוסחה הראשונה:

$$m \geq \frac{1}{0.1}\left(\ln 80 + \ln \tfrac{1}{0.1}\right)$$

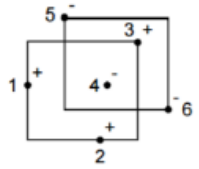
ה. מכיוון שערכים של כל תכונה הם רציפים, אזי מרחב ההיפותזות שלנו הוא אינסופי. לכן, נצטרך לחשב את ממד ה-VC של מרחב ההיפותזות שלנו שהוא מלבן שיודע לסווג או את הדוגמאות הפנימיות או את החיצוניות ל-+.

תחילה נראה ש- $VC(H) \geq 5$:



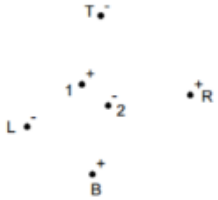


עכשיו נראה ש- $VC(H) < 6$:
נניח 6 נקודות בפיזור כלשהו. עכשיו נחלק את הנקודות לשתי קבוצות, V,W שהמלבן המינימלי שסוגר את הנקודות ב-V מכיל לפחות נקודה אחת מ-W והמלבן המינמלי שסוגר את הנקודות מ-W מכיל לפחות נקודה אחת מ-V. לדוגמא:



V היא הנקודות 1,2,3 ו-W היא 4,5,6. עכשיו אם נסווג את הנקודות ב-V כ + ואת הנקודות ב-W כ -, לא נוכל למצוא מלבן שידע לסווג כמו שצריך את הנקודות.
נראה שבכל 6 נקודות תמיד נוכל לבנות את 2 תתי הקבוצות הנ"ל ולמעשה למצוא דיכוטומיה שבכל פיזור לא מאפשרת סיווג נכון:
נגדיר,
$X_L \triangleq \min_i x_i$ – הערך בציר ה-X של הנקודה השמאלית ביותר L.
$X_R \triangleq \max_i x_i$ – הערך בציר ה-X של הנקודה הימנית ביותר R.
$Y_B \triangleq \min_i y_i$ – הערך בציר ה-Y של הנקודה התחתונה ביותר B.
$Y_T \triangleq \max_i y_i$ – הערך בציר ה-Y של הנקודה העליונה ביותר T.
את 4 ההגדרות האלו מגדירות 4 נקודות מתוך ה-6 (אם יש שוויון אז מחליטים לטובת אחת הנקודות). לשתי הנקודות האחרות נקרא 1,2. נניח בלי הגבלת הכלליות ש נקודה 1 בעלת ערך קטן יותר בציר ה-X, כלומר $x_1 \le x_2$. אם $y_1 \le y_2$ אז $V = \{L, B, 2\}, W = \{R, T, 1\}$:

אחרת, כאשר $y_1 > y_2$ אז $V = \{R, B, 1\}, W = \{L, T, 2\}$:



הראנו שבכל פיזור של 6 נקודות תמיד קיימת דיכוטומיה שלא ניתנת לסיווג ולכן $VC(H) < 6$.
סה"כ קיבלנו ש- $VC(H) = 5$.

תשובה 3 (6 סעיפים)

א. פונקציית הבדלה (discriminant function) היא כל פונקציה מפרידה בין אזורים (לדוגמא חיוביים ושליליים) במרחב המופעים (instance space). בהינתן מופע ניתן בעזרתה לעשות לו סיווג על ידי בדיקה לאיזה איזור המופע שייך ולפי זה להחליט כיצד לסווג אותו.

ב. פונקציית ההבדלה באלגוריתם הפרספטרון מייצרת גבול החלטה לינארי (גאומטרית זהו היפר-מישור n מימדי) כי כלל ההחלטה תלוי בצרוף לינארי של משקולות הפרספטרון במרחב $\mathbb{R}^n$. נניח כי w הוא וקטור המשקלות שמגדיר הפרספטרון אזי הפונקציה f מוגדרת בתור $f(\vec{x}) = \vec{w}\vec{x}$ וכלל ההחלטה הוא:

$$f(x) = \begin{cases} 1 \; if \; \vec{w}\vec{x} > 0 \\ -1 \; if \; \vec{w}\vec{x} < 0 \end{cases}$$

ג. נניח כי t_d הוא ערך המטרה (1 או 1-) עבור מופע x_d בקבוצת האימון. w_i הוא המשקל של מימד i (תכונה i) במרחב, η הוא מספר קטן (קצב הלמידה). כלל הלמידה ישנה את וקטור המשקולות בכל איטרציה באופן הבא:

$$w_i = w_i + \Delta w_i = w_i - \eta \sum_{d \in D} (f(x_d) - t_d) x_{id}$$

כלל זה עובד בגלל שאם הסיווג של המופע $f(x_d)$ הוא נכון אז הוא לא ישפיע על תוספת השינוי. ואם הוא לא נכון אז קיימים ארבעה מקרים:
אם $f(x_d)$ הוא 1 אבל t_d הוא 1- אז צריך להקטין את המכפלה $w_i x_{id}$ כדי שהערך של כל הסכום $f(x_d)$ יקטן.
בגלל שההפרש $f(x_d) - t_d$ הוא חיובי אז הסימן של המכפלה $(f(x_d) - t_d)x_{id}$ תלוי רק בסימן של x_{id}. אם הוא חיובי אז אנו מקטינים את w_i (בגלל שמכפילים ב- $-\eta$) ולכן גם המכפלה $w_i x_{id}$ תקטן. אם הוא שלילי אז אנו מגדילים את w_i אבל המכפלה $w_i x_{id}$ עדיין תקטן לפי מה שרצינו.
באותו אופן אם $f(x_d)$ הוא 1- אבל t_d הוא 1 אז צריך להגדיל את המכפלה $w_i x_{id}$ כדי שהערך של כל הסכום $f(x_d)$ יגדל.
בגלל שההפרש $f(x_d) - t_d$ הוא שלילי הפעם אז אם x_{id} הוא חיובי הסימן של המכפלה $(f(x_d) - t_d)x_{id}$ יהיה שלילי ואז אנו מגדילים את w_i (בגלל שמכפילים ב- $-\eta$) ולכן גם המכפלה $w_i x_{id}$ תגדל. אם x_{id} הוא שלילי אז הסימן של המכפלה $(f(x_d) - t_d)x_{id}$ יהיה חיובי ואז אנו מקטינים את w_i אבל המכפלה $w_i x_{id}$ תגדל כי x_{id} שלילי לפי מה שרצינו.

ד. הבעיה החמורה בשימוש בכלל זה הוא שיתכן ואלגוריתם למידה הפועל לפי הכלל הזה לא יתכנס במקרה ואין הפרדה לינארית בין המופעים בקבוצת האימון.

ה. נניח כי D^+ היא קבוצת המופעים החיוביים בקבוצת האימון ו-D^- היא קבוצת המופעים השליליים בקבוצת האימון אז פונקציית המטרה באלגוריתם LMS מקטינה עבור המופעים החיוביים את ההפרש בין הערך של הפונקציה הלינארית של המופע ו-1+, ועבור מופעים שליליים את ההפרש בין הערך של הפונקציה הלינארית של המופע ו-1- והנוסחה שלה היא:

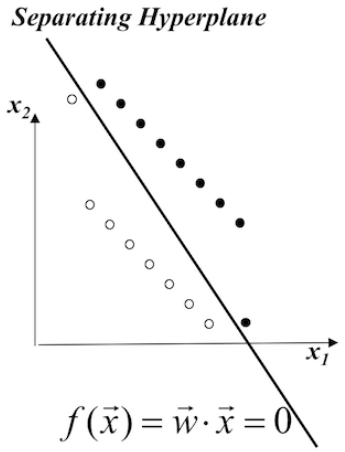
$$E(w) = \frac{1}{2} \sum_{d \in D} (\vec{w}\vec{x} - t_d)^2 = \frac{1}{2} \sum_{d \in D^+} (\vec{w}\vec{x} - 1)^2 + \frac{1}{2} \sum_{d \in D^-} (\vec{w}\vec{x} + 1)^2$$

ו. כלל הלמידה באלגוריתם LMS הוא:

$$w_i = w_i + \Delta w_i = w_i - \eta \sum_{d \in D} (\vec{w}\vec{x} - t_d) x_{id}$$

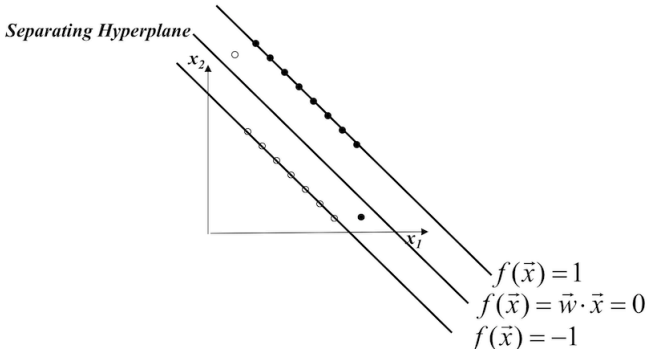
הכלל הזה נגזר פשוט משיטת gradient descent שבה אנו מביאים למינימום את הפונקציה (בסעיף הקודם) בעזרת הליכה בצעדים קטנים בכיוון הגרדיאנט של הפונקציה בכל פעם. אם נגזור את הפונקציה מסעיף ה נקבל כי הנגזרת החלקית לפי כל תכונה i היא $\sum_{d \in D} (\vec{w}\vec{x} - t_d) x_{id}$ ולכן כלל העידכון מתקדם קצת בכיוון הנגזרת לכל תכונה i.

ז. בגלל שקיימת הפרדה לינארית בקבוצת האימון אזי הפרספטרון ימצא גבול החלטה כזה והטעות שלו תהיה 0.



לעומת זאת אלגוריתם LMS מחפש להקטין את המרחק של הדוגמאות החיוביות מקו 1+ של פונקציית ההחלטה הלינארית ואת המרחקים של הדוגמאות השליליות מקו 1- של פונקציית ההחלטה הלינארית. לכן בדוגמא זו מאחר וכמעט כל הדוגמאות החיוביות נמצאות על קו והוא מקביל לקו שעליו נמצאות כמעט כל הדוגמאות השליליות, ניתן להניח כי LMS ימצא קו בדיוק ביניהם כדי להפריד את הדוגמאות כי אז הטעות של כל הדוגמאות (מלבד שתיים) תהיה – 0 והיא הטעות הקטנה האפשרית (לכל קו אחר תהיה טעות לכל הדוגמאות ולכן הוא לא הקו המינימאלי). אבל לרוע המזל קו ההפרדה במקרה הזה עובר בין שתי הדוגמאות יוצאות הדופן וגם מסווג אותם לא נכון ולכן בהכרח תהיה טעות.

$$E[\vec{w}] \equiv \frac{1}{2}\sum_{d\in D}(t_d - o_d)^2 = \frac{1}{2}\left[\sum_{d\in D^+}(\vec{w}\cdot\vec{x}_d - 1)^2 + \sum_{d\in D^-}(\vec{w}\cdot\vec{x}_d + 1)^2\right]$$



תשובה 4 (6 סעיפים)

א. רגרסיה לינארית Linear Regression

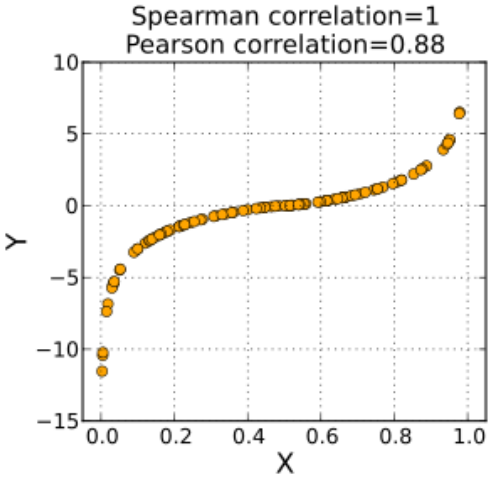
1. הפלט של אלגוריתם הלמידה ב linear regression הוא המשקלים של פונקציית ההחלטה שהם גם המקדמים של הפונקציה הלינארית במרחב התכונות (*features*), כלומר $\theta = (\theta_0, \theta_1, \dots, \theta_n)$.
2. הערך החזוי מתקבל ע"י מכפלה פנימית של התכונות של הדוגמא החדשה, x, עם וקטור המשקלים:
$$\hat{y} = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n = \theta \cdot x$$
3. הפונקציה שאותה רוצים להביא למינימום היא ממוצע הטעויות החיזוי בריבוע:
$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(\theta \cdot x^{(i)} - y^{(i)}\right)^2$$
4. *gradient descend* הוא תהליך שבו אלגוריתם מחפש מינימום של פונקציה גזירה (רב מימדית) על ידי תזוזה בצעדים קטנים. בכל צעד מחושב הגרדיאנט (וקטור הנגזרות החלקיות לכל כיוון) בנקודה נוכחית במרחב, והאלגוריתם זז לנקודה הבאה בכיוון הפוך לוקטור הגרדיאנט בצעדים קטנים.
מאחר ובבעיית רגרסיה לינארית יש פונקציה גזירה שהוגדרה בסעיף ג, ניתן להביא אותה למינימום בעזרת תהליך *gradient descend* בעזרת הנוסחה הבאה:
$$\theta_j := \theta_j - \alpha \frac{1}{m}\sum_{i=1}^{m}\left(\theta \cdot x^{(i)} - y^{(i)}\right) x_j^{(i)}$$

ב. Bias – הוא הבדל השגיאות בין הקונספט האמיתי שאותו מנסים ללמוד להיפותזה הטובה ביותר שמרחב היפותזות יכול לייצר. אפשר לחשוב על ההיפותזה הטובה ביותר בתור ההיפותזה "הממוצעת" על כל קבוצות האימון האפשריות.
Variance – מדד השונות של היפותזות שונות שנוצרות מקבוצות אימון שונות בגודל זהה.
כאן אפשר להסתכל על זה בתור ההבדל בין היפותזה מסוימת שנוצרה בעזרת קבוצת אימון מסוימת לבין ההיפותזה הממוצעת על כל קבוצות האימון בגודל זהה (שהיא גם הטובה ביותר האפשרית).
הדילמה באה לידי ביטוי בעובדה שניתן לפרק את הטעות האמיתית של היפותזה לשני מרכיבים חיבוריים: ה-bias וה-variance. מאחר והטעות קבועה (מוגדרת אחרי שבחרנו היפותזה מסוימת) אז אם נקטין מרכיב אחד (bias) לרוב המרכיב השני יגדל (variance) או ההיפך.
כלומר, ה-bias הוא טעות שנובעת מבחירת מרחב ההיפותזות (או אלגוריתם הלמידה) וה-variance נובע ע"פ רוב מהתאמת יתר (overfit) לקבוצת אימון.

ג. ניתן להשפיע על ה-bias וה-variance בבעיית הרגרסיה ע"י בחירה בפונקציה אחרת למדל את הקשר בין המשתנים המסבירים למוסבר. למשל ניתן לעשות רגרסיה פולינומית בו הפונקציה היא עם מעריך גבוה מ-1. בפונקציות כאלה, ה-*bias* יהיה נמוך יותר יחסית לרגרסיה לינארית כי מרחב ההיפוטזות גדול יותר ויש סיכוי טוב יותר שההיפותזה המתקבלת תתקרב לקונספט האמיתי. אבל מצד שני בגלל המרחב הגדול יותר זה יכול לגרום ל variance גבוה משום שבמצב זה ההיפותזה יכולה להיות 'מורכבת' מדי (יש בה יותר פרמטרים ללמוד) יחסית לקבוצת האימון, ואז כאשר תשתנה קבוצת האימון ישתנה מאוד הפלט של אלגוריתם הלמידה. לעומת זאת, ככל שהמעריך נמוך או הפונקציות פשוטות יותר ה-variance נמוך אבל ה-bias גבוה משום שבמצב זה סביר שמרחב ההיפותזות לא יוכל לייצג את קונספט המטרה.

ד. טעות האימון תהיה 0, מכיוון שיש תלות לינארית מלאה, כלומר על כל שינוי ב-x יש שינוי קבוע ב-y ויש בינהם קשר לינארי מדוייק שניתן לבטא אותו בעזרת פונקציה לינארית שאותה הרגרסיה תלמד.

ה. כן, ייתכן מצב בו Spearman correlation=1, אבל Pearson correlation≠1 ואז בוודאות לפחות לאחת מדוגמאות האימון תהיה פרדיקציה שונה מהערך האמיתי שלה (מה שיביא לטעות אימון גדולה מ-0). לדוגמא:



הערה: כמובן שייתכן מצב בו שניהם שווים ל-1 ואז טעות האימון תהיה שווה 0.

<u>תשובה 5 (5 סעיפים)</u>

א.

1. confusion matrix למסווג בינארי מוגדרת כך:

Predicted Class / True Class	Positive	Negative
Positive	#TP	#FN
Negative	#FP	#TN

2. Area Under the Curve – AUC

זהו מדד המאפשר השווה בין עקומות ROC שונות. המדד, כשמו כן הוא, בודק את גודל השטח מתחת עקומת ה-ROC, ככל שהשטח גדול יותר המדד גבוה יותר וזה טוב יותר.

המסווג הטוב ביותר לפי AUC הוא זה שיצר את העקומה השמאלית (זאת שגבוהה יותר), משום שהשטח מתחת לעקומה שם גדול יותר.

ב.

0.1:

Predicted Class / True Class	Positive	Negative
Positive	100	0
Negative	100	0

:0.25

Predicted Class / True Class	Positive	Negative
Positive	80	20
Negative	30	70

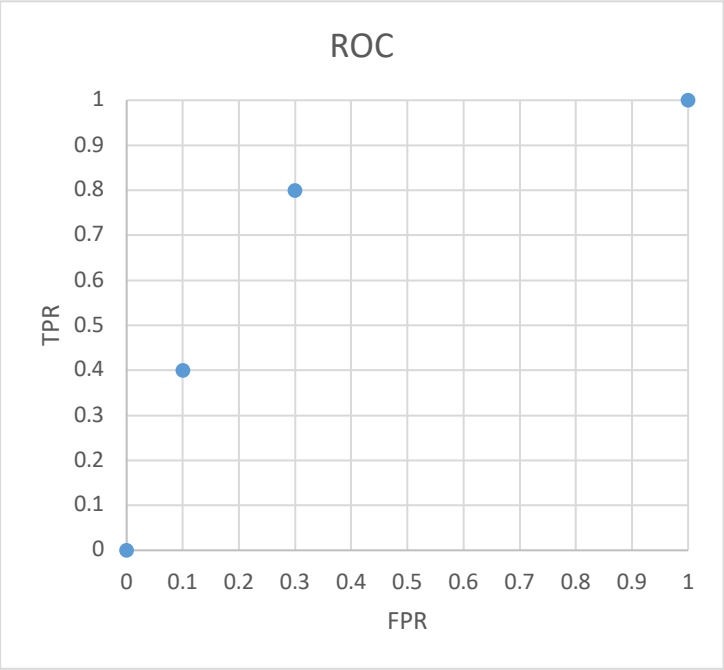
:0.75

Predicted Class / True Class	Positive	Negative
Positive	40	60
Negative	10	90

:0.9

Predicted Class / True Class	Positive	Negative
Positive	0	100
Negative	0	100

ג. כדי לייצר את עקומת ה-ROC, נצטרך לחשב לכל ערך p את ה-TPR וה-FPR:
נשים את הנקודות בגרף ונקבל:



:0.1

$$\mathrm{TPR} = \frac{100}{100} = 1$$

$$\mathrm{FPR} = \frac{100}{100} = 1$$

:0.25

$$\mathrm{TPR} = \frac{80}{100} = 0.8$$

$$\mathrm{FPR} = \frac{30}{100} = 0.3$$

:0.75

$$\mathrm{TPR} = \frac{40}{100} = 0.4$$

$$\mathrm{FPR} = \frac{10}{100} = 0.1$$

:0.9

$$\mathrm{TPR} = \frac{0}{100} = 0$$

$$\mathrm{FPR} = \frac{0}{100} = 0$$

ד. נחשב את ההפסד לכל אחד מהערכים:

:0.1

$$100 * C + 0 * 0.4C = 100C$$

:0.25

$$30 * C + 20 * 0.4C = 38C$$

:0.75

$$10 * C + 60 * 0.4C = 34C$$

:0.9

$$0 * C + 100 * 0.4C = 40C$$

הערך של p שממזער את מחיר ההפסד הוא p=0.75.

ה.

1. מכיוון שהסיווג מתבצע ע"פ אחוז הדוגמאות החיוביות שיש בעלה, אם נחלק את צומת 1 לשני צמתים בהם אחוז הדוגמאות החיוביות זהה לאחוז שהיה בצומת 1, אזי התוצאה של סעיף ב לא תשתנה.
לדוגמא, נחלק 20/10 לכל אחד מהצמתים 11,12.
2. כדי לקבל FPR=0.9 ו-TPR=1, אנו צריכים להעביר 10 דוגמאות מ-FP ל-TN ואז נקבל:

$$\mathrm{TPR} = \frac{100}{100} = 1$$

$$\mathrm{FPR} = \frac{90}{100} = 0.9$$

כדי לעשות זאת, נחלק את הדוגמאות לצמתים 11, 12 באופן הבא:
צומת 11: 40/10
צומת 12 0/10
החלוקה תגרום לסיווג חיובי בצומת 11, וסיווג שלילי בצומת 12, כלומר, 10 דוגמאות שליליות שסווגו כחיוביות בצומת 1, עכשיו יסווגו כשליליות בצומת 12.

<u>תשובה 6 (6 סעיפים)</u>

א. במשימת הסיווג אנחנו לומדים על קבוצת אימון הכוללת את המחלקות של הדוגמאות. לעומת זאת, בקיבוץ, הלמדה היא unsupervised, ואנו מנסים לחלק את הדוגמאות לתתי קבוצות, כאשר הנקודות בכל תת קבוצה כזו קרובות אחת לשנייה.

ב. הפונקציה ש k-means שואף להביא למינימום:

$$J = \sum_{\substack{\mu_i, \\ centroids}} \sum_{xj \in C_i} dist(x_j - \mu_i)$$

ג.

1. לא. אם ניקח איזושהי דוגמא מאחד ה- cluster (שאיננה נמצאת בדיוק במרכז, ובהגדרה יש כזאת כי הנקודות שונות אחת מהשנייה), ונהפוך אותה להיות מרכז של ה-cluster הריק (החמישי), אזי נשפר את ערך פונקציית המטרה משום שמרחק הנקודה הזו ממרכז ה-cluster החדש יהיה 0 (היא היחידה ששם). לכן בנקרה זה אם יש cluster ריק, אזי בהגדרה הפונקציה לא נמצאת במינימום גלובלי.
2. כן. אם נוריד את ההגבלה שכל הנקודות יהיו שונות, אז נוכל ליצור 4 קובצות של 25 נקודות זהות, ובמקרה זה 4 clusters יכולים להביא למינימום גלובלי.

ד. ישנן כמה אופציות אתחול שיביאו למינימום גלובלי (למעשה יש אינסוף כאלו...). ניקח לדוגמא את האתחול הבא: $\mu_1 = d_5$ ו- $\mu_2 = d_7$.
אתחול זה יביא למרכזים הסופיים הבאים: $\mu_1 = (2, 1.5)$ ו- $\mu_2 = (7, 1.5)$.
ערך פונקציית המטרה לאחר התכנסות האלגוריתם:

במרכז הראשון, יש שני מופעים במרחק $\frac{1}{2}$, וארבע מופעים במרחק $\sqrt{1+\left(\frac{1}{2}\right)^2}=\frac{\sqrt{5}}{2}$,

וסה"כ נקבל $1+4\frac{\sqrt{5}}{2}=1+2\sqrt{5}$.

במרכז השני יש ארבעה מופעים ב במרחק $\sqrt{1+\left(\frac{1}{2}\right)^2}=\frac{\sqrt{5}}{2}$, וסה"כ נקבל

$4\frac{\sqrt{5}}{2}=2\sqrt{5}$.

סה"כ נקבל $1+4\sqrt{5}$

ה. גם פה ישנן כמה אופציות אתחול שיביאו למינימום מקומי (למעשה יש אינסוף כאלו...). ניקח לדוגמא את האתחול הבא: $\mu_1 = d_5$ ו- $\mu_2 = d_7$.

אתחול זה יביא למרכזים הסופיים הבאים: $\mu_1 = (4,2)$ ו- $\mu_2 = (4,1)$.

אתחול זה יביא למרכזים אלו מכיוון שכל הנקודות עם y=2 קרובות יותר למרכז הראשון, ואילו כל הנקודות עם y=1 קרובות יותר למרכז השני. לאחר החלוקה הראשונה לתתי קבוצות נקבל את המרכזים שרשמנו כמרכזים סופיים, וגם להם החלוקה נשארת זהה, ולכן האלגוריתם יעצור.

כדי להראות שזהו אינו מינימום גלובלי נחשב את הערך הפונקציה בחלוקה זו:

במרכז הראשון, סכום המרחקים מהמרכז יהיה: $3+2+1+2+4=12$.

במרכז השני, סכום המרחקים מהמרכז יהיה: $3+2+1+2+4=12$.

סה"כ נקבל 24, וכמובן ש- $1+4\sqrt{5}<24$, ולכן זהו מינימום מקומי.

ו. אחת הדרכים להתמודד עם בעיה זו היא לנסות אתחולים שונים. צריך לוודא שהאתחולים שונים מספיק אחד משהני כדי לתת תוצאות שונות. לבסוף נבחר את הקיבוץ שממזער את J.

בהצלחה!