$AdaBoost$

שרון מלטר, אתגר 17

10 באוגוסט 2024

1 מוטיבציה

לרוב קשה לעצב מעריך שיש לו דיוק גבוה עבור $test\_set$ והוא גם מכליל באופן טוב (כלומר הוא לא עושה $underfitting$ וגם לא $underfitting$) עם זאת, ניתן למצוא בקלות מעריכים הפועלים לפי 'כלל אצבע', אך הם רק מעט יותר טובים מניחוש רנדומלי.
אבל - מה אם נוכל לשלב בין מספר מעריכים חלשים כדי למצוא אחד מדויק?
(שילוב אמיתי, לא הצבעה)

2 $AdaBoost$

הרעיון המרכזי הוא לשקול את החיזויים של המעריכים עם אחוז השגיאה שלו ולאחר מכן נתמקד בדוגמאות אשר הוערכו לא נכון. כלומר, מדובר באלגוריתם איטרטיבי.
נתחיל מהמקרה של סיווג בינארי, כך שניתן לסווג $1-/1$. במקרה זה נקבל את פונקציית האפליה הבאה;

$$g(x)=\sum_{t=1}^{T}\alpha_t f_t(x),\quad \alpha_t\geqslant 0$$

כאשר T הוא מספר המעריכים שיש לנו ו- $f_t(x)$ הוא המעריך החלש ה- i (ששוב, הוא קצת יותר טוב מהטלת מטבע) ו- α_i הוא המשקל שנבחר עבורו.
נתאר את האלגוריתם באופן כללי:

- תחילה, כל המשקלים זהים.
- נשתמש ב- $g(x)$ כדי להעריך את דוגמאות האימון.
- נמצא אילו דוגמאות סווגו נכון ואילו לא. בפונקציית השגיאה נגדיל את המשקל של דוגמאות שסווגו לא נכון ונקטין את המשקל של אלו שסווגו נכון- כך שהלמדן יתמקד בדוגמאות הקשות.
  עם זאת, משקל המסווגים שפעלו לא נכון עם אותן דוגמאות יקטן עבורן.

באופן יחסי, פשוט לממש את האלגוריתם (כל עוד יש לנו מימוש של האלגוריתמים החלשים יותר) כמו כן, ניתן להשתמש באלגוריתם כדי לחזק מעריכים בסיסיים (שחייבים להיות לפחות מעט יותר טובים מניחוש רנדומלי)

לסיום, נציג את האלגוריתם המלא:

1. המשקל $d_0(x_i)=\frac{1}{n}$ הוא ההתחלתי עבור הדוגמה x_i. (בכל איטרציה תמיד מתקיים $\sum d(x_i)=1$)

2. בכל איטרציה $t\in T$:

3. נמצא את המסווג החלש הכי טוב $f_i(x)$ בעזרת הדוגמאות והמשקלים שלהם.

4. נחשב את השגיאה של אותו מסווג;

$$\epsilon_t=\sum_{i=1}^{N}d_t(x_i)\cdot\delta(y_t\neq f_t(x))$$

כאשר δ היא פונקציית $0-1$ (משקל השגיאה)

5. נגדיר את α_t, משקל השגיאה של הרצה זו, להיות־

$$\alpha_t = \frac{1}{2} \cdot ln(\frac{1-\epsilon_t}{\epsilon_t})$$

הערה: אכן מתקיים $\alpha_t > 0$ מכיוון ש־ $\epsilon_t < \frac{1}{2}$.
זוהי פונקציה יורדת ממש כך שככל שהשגיאה ϵ_t גדלה, משקלה קטן. באופן זה המסווג ישפיע על הדוגמאות פחות ככל ששגא יותר.

6. לכל x_i, נעדכן את ההתפלגות של משקלי הדוגמאות עבור האיטרציה הבאה־

$$d_{t+1}(x_i) = d_t(x_i) \cdot exp(-\alpha_t \cdot y_i \cdot f(x_i))$$

כלומר;

$$d_{t+1}(x_i) = d_t(x_i) \cdot e^{-\alpha_t \cdot y_i \cdot f_t(x_i)}$$

$$d_{t+1}(x_i) = d_t(x_i) \cdot e^{-\frac{1}{2} \cdot ln(\frac{1-\epsilon_t}{\epsilon_t}) \cdot y_i \cdot f_t(x_i)}$$

$$d_{t+1}(x_i) = d_t(x_i) \cdot e^{-(ln(\frac{1-\epsilon_t}{\epsilon_t}))^{\frac{1}{2} \cdot y_i \cdot f_t(x_i)}}$$

$$d_{t+1}(x_i) = d_t(x_i) \cdot (ln(\frac{1-\epsilon_t}{\epsilon_t}))^{-\frac{1}{2} \cdot y_i \cdot f_t(x_i)}$$

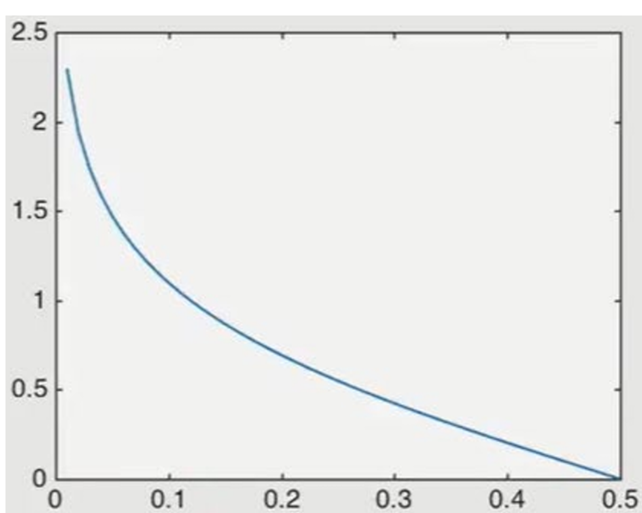
כך שמשקל הדוגמה גדל אם המסווג שגא בסיווגה, ובאיטרציות הבאות נמצא מסווג המתאים לה יותר.
לאחר מכן ננרמל את $d_{t+1}(x_i)$ כך שסכומם יהיה 1.

הערה: מאחר שמדובר במקרה הבינארי, מתקיים $y_i \cdot f_t(x_i) \in \{-1, 1\}$

7. בסוף האיטרציות, נסווג את הדוגמאות כך;

$$f_{FINALE} = sign(\sum_{t=1}^{T} \alpha_t \cdot f_t(x))$$

ולהמחשה, הגרף $\alpha_t = \frac{1}{2} ln(\frac{1-\epsilon_t}{\epsilon_t})$;



הערות !

- אלגוריתם זה עובד רק במקרה של סיווג בינארי.
- אם המעריכים שלנו לא לוקחים דוגמיות ממושקלות, נוכל לקחת דוגמיות מנתוני האימון לפי התפלגות $d_t(x)$ (כלומר, להשפיע על משקל הדוגמאות עבור המעריכים באמצעות כמות המבוססת על הסתברות עם $d_t(x)$)
- מאחר שכל מעריך יותר טוב ממעריך רנדומלי, אחוז השגיאה ϵ_t קטן מ־ $\frac{1}{2}$.
- ניתן להוכיח שאחוז השגיאה באימון קטן באופן אקספוננציאלי; $Error_{train} \leqslant exp(-2\sum_t(\epsilon_t - \frac{1}{2})^2)$

יתרונות :)

- האלגוריתם מהיר באופן יחסי.
- הוא פשוט.
- לאלגוריתם יש רק פרטמר אחד שיש לבחור־ מספר המעריכים T.
- האלגוריתם גמיש־ ניתן לשלב אותו עם כל משלב (או משלבים)
- ניתן למצוא $outliers$ (חריגים) שהינן הדוגמאות הקשות ביותר למציאה.
- האלגוריתם אפקטיבי.
- האלגוריתם מותאם למניעת $overfitting$.
- השגיאה עבור $test$ קטנה לאחר ששגיאת האימון היא 0.
- האלגוריתם ממקסם את המרחק מהשוליים, כפונקציה של T.

חסרונות :(

- האלגוריתם תלוי במעריך חלש־ כך שאם הוא חלש **מדי** הוא יכול להיכשל. אך אם הוא חזק מדי, עלול להתרחש $overfitting$.
- באופן אמפירי, האלגוריתם רגיש מאוד לרעש. זאת מכיוון שהוא מנסה לסווג את כל הנקודות שהוא מקבל.

עד כאן דיבורים.
נבין עד כמה הבנו, עם **שאלות!**

שאלות נכון / לא נכון:

1. אם למסווג בינארי יש שגיאה ממושקלת $\epsilon \leqslant \frac{1}{3}$, אזי הוא יכול לסווג לא נכון רק $\frac{1}{3}$ מנקודות האימון?
2. כאשר מעדכנים משקלים, נקודת האימון בעלת המשקל הקטן ביותר באיטרציה הקודמת תמיד תגדיל את משקלה?
3. $AdaBoost$ מתחשבת בחריגות בכך שהיא מקטינה את משקלי נקודות האימון שמסווגים באופן לא נכון פעמים רבות?

תשובות:

1. לא נכון.
ייתכן שהמסווג עונה נכונה על יותר משליש מנקודות האימון, אך משקלם נמוך.
למשל: אם נקודות האימון הן x_1, x_2, x_3 והוא מסווג נכונה רק את x_1 אך המשקלים הם; $d_t(x_3) = \frac{1}{6}$ $d_t(x_2) = \frac{1}{6}$ $d_t(x_1) = \frac{2}{3}$

2. לא נכון.
משקלה יגדל רק אם היא תסווג לא נכון במסווג הטוב ביותר באיטרציה הבאה.

3. לא נכון.
משקל של נקודת אימון גדל כאשר היא מסווגת לא נכון.