ב' - 25/5/20 תרגול 8#

נדבר על Boosting: טכניקה שמאפשרת לשפר ביצועים של דאטה באלגוריתמים קיימים, ונציג 2 טריקים עבורה - לשיפור ה-confidence ולשיפור ה-accuracy.

(1) Boosting the confidence: (שיטת ועדה)

נתונים לנו T מומחים, $\{X_i\}_{i=1}^{T} \in \{\pm 1\}$ שמחזירים תשובה נכונה בסיכוי p, ותשובה שגויה בסיכוי $1-p$.

נניח ש- $p > \frac{1}{2}$. נשאלת השאלה; אם נצביע לפי רוב התשובות, מה הסיכוי שצדקנו? תחת הסימונים שלנו, חישוב הצבעת הרוב שקולה למציאת: $\text{sign}(\bar{X}) := \text{sign}\left(\sum_{t=1}^{T} X_t\right)$.

אנו רוצים לחסום את ההסתברות שטעה נכונה, כלומר את $P(\bar{X} > 0) = P(\text{sign}(\bar{X}) = 1)$.

לשם כך נחשב את הסיכוי של המאורע המשלים:

$$P(\bar{X} \le 0) = P(-\bar{X} \ge 0) \underset{(a\in\mathbb{R}_+)}{=} P(-a\bar{X} \ge 0) \underset{(\text{מונו' האקספוננט})}{=}$$

$$= P(e^{-a\bar{X}} \ge e^{0}) \underset{(\text{א"ש מרקוב})}{\le} \frac{\mathbb{E}[e^{-a\bar{X}}]}{e^{0}} = \mathbb{E}\left[e^{-a\sum_{t=1}^{T} X_t}\right] =$$

$$= \mathbb{E}\left[\prod_{t=1}^{T} e^{-aX_t}\right] \underset{(\text{מניחים } X_t \text{ בת"ל ושווי התפלגות})}{=} \left(\mathbb{E}\left[e^{-aX_1}\right]\right)^{T} \underset{(\text{הגדרת תוחלת ל-}\{X_i\}\in\{\pm 1\})}{=}$$

$$= \left(pe^{-a} + (1-p)e^{a}\right)^{T} = \left[e^{a}(1 - p + pe^{-2a})\right]^{T} \underset{(1+x \le e^{x})}{\le}$$

$$\le \left(e^{a}e^{-p+pe^{-2a}}\right)^{T} \underset{(\text{בחרנו: } a=\frac{1}{2}\ln(2p) \text{ מקיים } a\in\mathbb{R}_+ \text{ ל-} p>\frac{1}{2})}{=} \exp\left(T\left(\frac{1}{2}\ln(2p) - p + p\cdot\frac{1}{2p}\right)\right) =$$

$$= \exp\left(Tp\left(-\frac{1}{2p}\ln\left(\frac{1}{2p}\right) - 1 + \frac{1}{2p}\right)\right) \underset{(\text{בחרנו: } x=\frac{1}{2p} \Leftarrow \forall x\in[0,1]: x\ln(x) \ge \frac{x^2}{2} - \frac{1}{2})}{\le}$$

$$\le \exp\left(Tp\left(-\frac{1}{2(2p)^2} + \frac{1}{2} - 1 + \frac{1}{2p}\right)\right) =$$

$$= \exp\left(-\frac{Tp}{2}\left(\frac{1}{(2p)^2} - \frac{1}{p} + 1\right)\right) = \exp\left(-\frac{Tp}{2}\left(\frac{1}{2p} - 1\right)^{2}\right) =$$

$$= \exp\left(-\frac{T}{2p}\left(\frac{1}{2} - p\right)^{2}\right) \Rightarrow P(\bar{X} > 0) = 1 - P(\bar{X} \le 0) \ge$$

$$\ge 1 - e^{-T/2p\,(p-1/2)^2}$$

כלומר הסיכוי לטעות קטן בקצב $O(e^{-T})$ מהיר! ומספיקים מומחים מועטים כדי להקטין את ה-confidence משמעותית.

דוגמא (שהסביר למה): עבור אותם $\{X_i\}_{i=1}^T$, מהי התוחלת ומהי השונות לכל X_t ($t\in[T]$), ועבור הממוצע: $\bar{X} := \frac{1}{T}\sum_{t=1}^T X_t$?

★ הפעם הגדרנו את $\bar{X}$ להיות הממוצע, בשונה מההגדרה הקודמת שלו. מכיוון שאנו למעשה מעוניינים ב- sign, 2 ההגדרות שקולות.

$\Leftarrow$

- $E[X_t] = 1\cdot p - 1(1-p) = 2p-1$

  וכאן אם $p>\frac{1}{2}$ אז $E[X_t]>0$

- $Var(X_t) = E[(X_t - E[X_t])^2] = E[X_t^2] - (E[X_t])^2 =$
  $= [p\cdot 1^2 + (1-p)\cdot(-1)^2] - [2p-1]^2 = 1-4p^2+4p-1 =$
  $= 4p(1-p)$

- $E[\bar{X}] = E[\frac{1}{T}\sum_{t=1}^T X_t] \underset{\text{(לינאריות)}}{=} \frac{1}{T}\sum_{t=1}^T E[X_t] \underset{\text{(ב"ת } X_t)}{=} E[X_1] = 2p-1$

- $Var(\bar{X}) = Var(\frac{1}{T}\sum_{t=1}^T X_t) = \frac{1}{T^2} Var(\sum_{t=1}^T X_t) \underset{\text{(ב"ת } X_t)}{=} \frac{1}{T^2}\sum_{t=1}^T Var(X_t) =$
  $= \frac{1}{T} Var(X_1) = \frac{1}{T}\cdot 4p(1-p)$

שיטת הוועדה שבחרנו משמרת את התוחלת החיובית כמו שהיא, אך מקטינה את השונות בקצב של $O(\frac{1}{T})$, ולכן ה-confidence עולה: הביטחון ברמת הדיוק ($(\delta-1)$ בהגדרת למידות-PAC) מוגדר ע"פ מדגם אקראי בגודל m. לכן אם נדגום T מדגמים כאלו ונמצע את התשובות אז נגדיל את ה-confidence שלנו!

__אבל__: למה שלא נאחד T m-מדגמים למדגם אחד גדול, ונפעל עליו?
כי במה שהראינו דרשנו יש __מעלה__ אחרת:

Bootstrap: דגימה ב"ת ועם חזרות מאותו סט, כדי לייצר m-מדגמים שונים מאותו הסט.

★ אפשר להשתמש במה שחישבנו בשביל לממש את ה-bootstrap! אבל יש בעיה אחת: הדגימות אינן ב"ת. __נעדכן את הבעיה__:

נניח ש: $\{X_i\}_{i=1}^T$ הם מ"מים אקראיים עם $Var(X_t)=\sigma^2$, והם תלויים: לכל $i\neq j\in[T]$: $Corr(X_i,X_j)=\rho$. __נראה ש__: $Var(\bar{X}) = \rho\sigma^2 + (1-\rho)\frac{\sigma^2}{T}$

נזכיר: $Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$. __נתון__: $E[X_t]=\mu$ אז: $E[\sum_{t=1}^T X_t] = T\mu$

$\Leftarrow$ $Var(\bar{X}) = Var(\frac{1}{T}\sum_{t=1}^T X_t) = \frac{1}{T^2} Var(\sum_{t=1}^T X_t) =$

$$= \frac{1}{T^2}\left[\mathbb{E}\left[\left(\sum_{t=1}^{T} X_t\right)^2\right] - (T\mu)^2\right] = \frac{1}{T^2}\mathbb{E}\left[\sum_{s,t=1}^{T} X_s X_t\right] - \mu^2 \underset{(\text{לינאריות } \mathbb{E})}{=}$$

$$= \frac{1}{T^2}\sum_{s,t=1}^{T} \mathbb{E}[X_s X_t] - \mu^2 =$$

(⋆) $\rho = Corr(X_s, X_t) = \frac{Cov(X_s, X_t)}{\sigma^2} = \frac{\mathbb{E}[X_s X_t] - \mu^2}{\sigma^2} \Rightarrow$

לכל $s \neq t$: $\boxed{\mathbb{E}[X_s X_t] = \rho\sigma^2 + \mu^2}$

אחרת: $\sigma^2 = Cov(X_s, X_t) = Cov(X_t, X_t) = Var(X_t) =$

$= \mathbb{E}[X_t^2] - \mu^2 \Rightarrow \boxed{\mathbb{E}[X_t^2] = \sigma^2 + \mu^2}$

$$= \frac{1}{T^2}\left[\sum_{s,t\in[T],\, s\neq t}(\rho\sigma^2 + \mu^2) + \sum_{s,t\in[T],\, s=t}(\sigma^2 + \mu^2)\right] - \mu^2 =$$

$$= \frac{1}{T^2}\left[(T^2 - T)(\rho\sigma^2 + \cancel{\mu^2}) + T(\sigma^2 + \cancel{\mu^2})\right] - \cancel{\mu^2} =$$

$$= \frac{1}{T^2}\left[(T^2 - T)\rho\sigma^2 + T\sigma^2(\rho - 1 - \rho)\right] = \frac{\cancel{T^2}}{\cancel{T^2}}\rho\sigma^2 + (1-\rho)\frac{\sigma^2}{T}$$

■

(2) Boosting the accuracy (Adaboost):

משתמש בעקרון של weak learner כדי להגביר את הדיוק.

למה צריכים לומד חלש? נסתכל על 2 דוגמאות:

1. בעיית ה- Bias-Complexity Tradeoff

נסמן את הפתרון שהחזיר האלג' ב- $h_S = A(S)$, ואת הפתרון הטוב ביותר ממחלקת ההיפותזות $\mathcal{H}$ ב- $h^* = \arg\min_{h\in\mathcal{H}} L_D(h)$ מבחינת שגיאת ההכללה, אז:

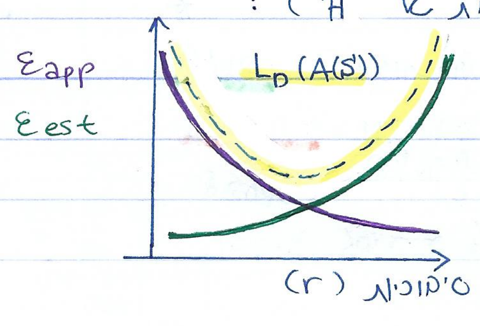
$$L_D(h_S) = \underbrace{L_D(h^*)}_{\varepsilon_{app}} + \underbrace{L_D(h_S) - L_D(h^*)}_{\varepsilon_{est}}$$

ε_{app} – שגיאת הקירוב; ε_{est} – שגיאת השערוך

במקרים רבים שגיאות אלו באות על חשבון השנייה. לדוג':

ברגרסיה פולינומיאלית: $\mathcal{H}_r = \{p : p \text{ פולינום בדרגה } \leq r\}$

ננתח את בחירת r (שאותו נפרש כמידת הסיבוכיות של $\mathcal{H}$):



⇐ לא תמיד פשוט למצוא את נק' האיזון, אז בבואנו לבחור r נשאל את עצמנו היכן הנו מעדיפים להיות: בצד ה"שמרן" (r נמוך) או בצד ה"מתבדר" (r גבוה). לרוב נעדיף להיות שמרנים כי נצפה שדרגות פול' נמוכות יהיו מקדמים גדולים יותר ונקבל לומד חלש.

2. סיבוכיות זמן ריצה:

מי אומר שניתן להריץ ERM בזמן סביר? לא התייחסנו לסיבוכיות הזו.

נתייחס לחישוב זמני הריצה עבור 2 המטלות -

* מציאת ההיפותזה הטובה ביותר (אימון).
* שערוך (פרדיקציה) על דגימה חדשה, בהינתן ההיפותזה שמצאנו.

A הוא אלג' למידה יעיל אם הוא פולינומיאלי ב-3 פרמטרים:

* m גודל המדגם
* $1/\varepsilon$ גודל שגיאת הדיוק (רוצים להגדיל את זמן הריצה ככל ש- $\varepsilon \to 0$ כלומר ככל ש- $1/\varepsilon \to \infty$)
* $1/\delta$ שגיאת ה-confidence.

דוגמה: קלט: תמונה בגודל 19x19 פיקסלים, כל מקום מקבל ערך בין 0 ל- 255. פלט: להחזיר 1 אם יש בתמונה פנים, -1 אם אין. כלומר: $X=[256]^{19\times 19}$, $Y=\{\pm 1\}$.

⇐ באופן נאיבי: נחפש היפותזה $h:[256]^{19\times 19}\to\{\pm 1\}$ ששייכת למחלקת היפותזות סופית אבל גדולה מאוד! נעדיף לבחור לומד חלש שאולי לא יחזיר את ההיפותזה הטובה ביותר, אבל יחזיר משהו בזמן סביר.

γ-לומד חלש: אלג' למידה A ייקרא γ-לומד חלש עבור מחלקת היפותזות H כשפא אם קיימת $m_H:(0,1)\to\mathbb{N}$ עבורה לכל $\delta\in(0,1)$ ולכל D התפלגות מעל X, אזי לכל פונק' תיוג $f:X\to\{\pm 1\}$ המקיימת את הנחת הריאליזביליות ביחס ל- H,D,f אזי אם ניקח $m\geq m_H(\delta)$ דגימות iid מ-D שמתויגות ע"י f יתקיים:

$$L_D(A(S))\leq 1/2-\gamma$$

מחלקת היפותזות H γ-למידה חלש: H היא γ-למידה חלש אם קיים לה γ-לומד חלש.

★ ההבדל בין הגדרה זו להגדרת למידות-PAC הוא שכאן קבענו את ε.

טענה: למידות-חלשה גוררת למידות-PAC.

הוכחה: מהמשפט היסודי: אם H ממימד VC סופי אז דרושות לנו לפחות $m_H(\varepsilon,\delta)\geq C_1\frac{d+\log(1/\delta)}{\varepsilon}$ דגימות עבור $C_1\in\mathbb{R}$ קבוע. ובפרט עבור $\varepsilon=1/2-\gamma$ מצאנו זה מתקיים. אחרת, נניח שמימדה ∞

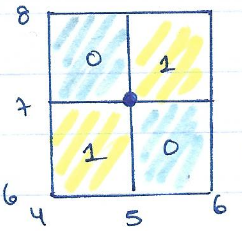
נאמר הבא אינה למידת-PAC אז H לא יכולה להיות למידה חלש כי זה יסתור את הדרישה לגבי כמות דגימות סופית. ☒

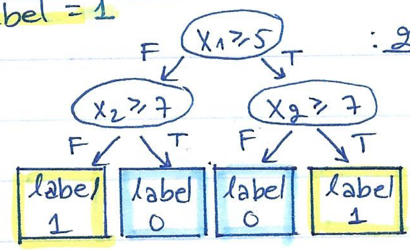
נאמר ניתן לאפיין למידה חלשה ע"י מימד-VC, כמו בלמידות-PAC אך מבחינת זמני ריצה - יש לנו יתרון בשימוש בלומד חלש.

דוגמאות ללומדים חלשים:

- עבור אוסף התמונות MNIST (מאגר ספרות בכתב יד, בגודל 28x28). נרצה להפריד בין 0 ל-1. אז נתבונן בפיקסל האמצעי ונבחן האם הוא כהה או בהיר. על-סמך הבחנה זו נבנה לומדים-חלשים: עבור פיקסל (i,j) $(i,j\in[28])$ נסמן $h_{i,j}=\mathbb{1}[pixel(i,j)\geq 127]$. אז $H=\{h_{ij}\mid i,j\in[28]\}\cup\{-h_{ij}\mid i,j\in[28]\}$ מחלקת ההפונקציות.
  סיווג (i,j) ככהה, סיווג (i,j) כבהיר
- דוגמה שלישית: $X=\mathbb{R}^2$, $y=\{0,1\}$.
  {אוסף העצים עם לכל היותר 2 רמות} $=H$. נסמן: $B\subset H$ להיות כל העצים עם רמה אחת בדיוק. האם קיים אלג' ע-לומד חלש שמחזיר היפותזה מ-B?

לא! נסתכל על ההתפלגות האחידה מתחום $[4,6]\times[6,8]$ שמתפלגת אחיד בריבוע שמוצג בשאלה:
אפשר ללמוד זאת עם עץ בגובה 2:





$\Leftarrow$ f ריאליזבילית.

אך לא ניתן ללמוד אותו עם עצים מגובה 1: נניח בה"כ שהעץ לומד לפי הצבר x_1. כלומר נחליר הפונקציה: $h_{\theta_1}(x)=\mathbb{1}[x_1>\theta_1]$ ל-θ_1 כלשהו. נחשב: (ונצפה לקבל משהו טוב יותר מניחוש גמור $(\theta_1 < x_1)$):

$$P(y=0\mid x_1>\theta_1)=P(y=0,\ x_1<5\mid x_1>\theta_1)+P(y=0,\ x_1\geq5\mid x_1>\theta_1)=$$
$$=P(x_2\geq7,\ x_1<5\mid x_1>\theta_1)+P(x_2<7,\ x_1\geq5\mid x_1>\theta_1)\overset{★}{=}$$

(★ בהינתן θ_1 כלשהו הסיכוי להיות $x_1\leq5$ ו- $x_2\geq7$ שווה לסיכוי להיות $x_1\leq5$ ו- $x_2\leq7$ כי לשני החלקים יש שטח זהה, ועיקרם ההתפלגות אחידה לכן הם שווי התפלגות במרחב $[4,6]\times[6,8]$ כולו.)

$$=P(x_2<7,\ x_1<5\mid x_1>\theta_1)+P(x_2<7,\ x_1\geq5\mid x_1>\theta_1)=$$ (הסתברות שלמה ל-$x_2<7$)

לא משנה אם $\leq$ או $<$ כי אנו בהתפלגות רציפה אז $P(x=t)=0$

$$=P(x_2<7\mid x_1>\theta_1)=\frac{1}{2}$$

(כלומר להיות 2 המרובעים התחתונים והם שווי שטח)

$$\left(=P(y=0\mid x_1<\theta_1)=\ =P(y=0\mid x_2>\theta_2)=\ldots\right)$$

אולי בכ״ז נצליח ללמוד משהו? ל- $X \overset{iid}{\sim} D$:

$$L_{D,f}(h_{\theta_1}) = P(y \neq h_{\theta_1}(x)) = P(y=0, h_{\theta_1}(x)=1) +$$

$$+ P(y=1, h_{\theta_1}(x)=0) = P(y=0, x_1 > \theta_1) +$$

$$+ P(y=1, \theta_1 > x_1) \underset{(\text{בייס})}{=} P(y=0 \mid x_1 > \theta_1) P(x_1 > \theta_1) +$$

$$+ P(y=1 \mid x_1 < \theta_1) P(x_1 < \theta_1) = \frac{1}{2} P(x_1 > \theta_1) + \frac{1}{2} P(x_1 < \theta_1) =$$

$$= \frac{1}{2} [P(x_1 > \theta_1) + P(x_1 < \theta_1)] = 1/2 \cdot 1 = 1/2 \Rightarrow$$ לא

★ ניתן להרחיב דוגמא זו ל- $\mathbb{R}^m$ בקלות.

★ הכשלון נעוץ בהתפלגות הלא טבעית הזו

⇐ Adaboost

— האלגוריתם כתוב בסוף ההרצאה השישית —

★ האלג׳ משתמש בעובדה שלומדים חלשים מחזירים את אותה השגיאה לכל התפלגות. ע״י שינוי ההתפלגות בין האיטרציות, האלג׳ מבטיח לחלק דוגמאות שמש״ג בצורה שגויה קודם.

★ נחשוב על D^i כמו משקולות על הדגימות.