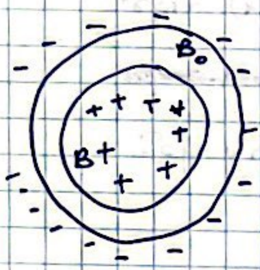
למידה חישובית - תרגיל 2

(1) בהינתן מדגם בגודל n בלתי תלוי נמצא את העיגול הכי קטן שעקבי עם המדגם

~~[illegible]~~ ונסמן אותו ב B_0 כלומר עבור המדגם $\{u_1,\dots,u_n\}$

לכל u_i $\|u_i\|_2 \le R$. נסמן ב B את הכדור ה"מושלם" קיים כזה כיוון שאנו במקרה realizable

נשים לב ~~[illegible]~~ שהטעות היחידה שעשויה היא לסמן נקודות חיוביות כשליליות.



הטעות של האלגוריתם תהיה $e_p(h_R) = P(B_0 \setminus B)$

אם $P[B_0 \setminus B] \le e$ אז $e_p(h_R) = P[B_0 \setminus B] \le P[B_0] \le e$

נניח אם כן: $P[B_0] \ge e$

נבחר לפי ~~[illegible]~~ מרכז B ~~[illegible]~~ ונסמן את T כך ש $P(T)=\varepsilon$

T מתחילה מקצוות B_0 ומתקדמת פנימה לכיוון המרכז.

אם נניח כי קיימת דגימה u_i כך ש $u_i \in T$ אז נקבל כי $B_0 \setminus B_n \subseteq T$

ולכן $e_p(h_n) = P(B_0 \setminus B_n) \le P(T) = \varepsilon$

באופן שווה אם $e_p(h_R) > \varepsilon$ אזי ~~[illegible]~~ כל n הדגימות לא נמצאות ב T

הסיכוי שדגימה לא ב T היא $1-\varepsilon$ ולכן:

$$P(e_p(h_R) > \varepsilon) \le P(\forall i : u_i \notin T) \le \prod_{i=1}^{n} P(u_i \notin T) = (1-\varepsilon)^n$$

נקבל ש $(1-\varepsilon)^n \le e^{-n\varepsilon} \le \delta \Rightarrow n \ge \frac{1}{\varepsilon} \cdot \ln\frac{1}{\delta}$

לכן נסמן $N(\varepsilon,\delta) = \frac{1}{\varepsilon}\ln\frac{1}{\delta}$

עבור כל ε, δ וכל realizable התפלגות P מעל R^d labeling $B_0 \in H_{ball}$

כאשר מריצים אלגוריתם שמחזיר למדגם של $n \ge N(\varepsilon,\delta)$ נקודות שנדגמו בצורה בת"ל מ P היפותזה h_R מתוך H_{ball}

כך ש $P[e_p(h) > \varepsilon] \le (1-\varepsilon)^n \le \delta$. ולכן H_{ball} היא למידה PAC.

~~[illegible]~~ נשים לב כי סיבוכיות המדגם לא תלויה במימד d.

בכיוון הראשון נניח כי H היא למידה PAC בתוחלת. ②

יודעים כי $0 \le e_p(A(S))$, ולכן לפי אי שוויון מרקוב מקבלים כי

$$P\left(e_p(A(S)) > \varepsilon\right) \le \frac{E[e_p(A(S))]}{\varepsilon}.$$

עבור $a \in (0,1)$ יהיו δ, ε עבורם $a = \varepsilon \cdot \delta$

$N(a) = N(\varepsilon \cdot \delta)$

ולכן מתקיים:

$$P(e_p(A(S)) > \varepsilon) \le \frac{E[e_p(A(S))]}{\varepsilon} \le \frac{a}{\varepsilon} = \frac{\varepsilon \cdot \delta}{\varepsilon} = \delta$$

כלומר לכל $n > N(a)$ זה מתקיים ולכן H היא למידה PAC.

בכיוון השני נניח כי H למידה PAC. $a \in (0,1)$ לכל $n > N(\varepsilon, \delta)$: (כך ש $a = \varepsilon + \delta$)

$$E[e_p(A(S))] \overset{\text{לפי הגדרת התוחלת}}{=} E\Big[\overbrace{e_p(A(S)) \mid e_p(A(S)) > \varepsilon}^{\le 1}\Big] P\big(\overbrace{e_p(A(S)) > \varepsilon}^{\le \delta}\big) +$$

$$+ E\Big[\overbrace{e_p(A(S)) \mid e_p(A(S)) < \varepsilon}^{\le \varepsilon}\Big] P\big(\overbrace{e_p(A(S)) < \varepsilon}^{\le 1}\big) =$$

$$\overset{\text{לפי PAC}}{\le} \delta + \varepsilon = a$$

לכן עבור $a = \varepsilon + \delta$ H למידה PAC בתוחלת עם אלגוריתם A.

למידה חישובית - המשך

(3) נוכיח כי $VC\text{-}dimension(H_k) = 2k$

נראה כי $VC(H_k) \geq 2k$:

נבחר קבוצה בגודל 2k $x_i = \frac{i}{2k}$ ויהי $s_1, \dots, s_{2k} \in \{0,1\}$ כלשהו.

עבור כל רצף של $x_i, x_{i+1}, \dots, x_j$ עבורם $y_i = \dots = y_j = 1$ נגדיר אינטרבל $[x_i, x_j]$

ולכן לכל $i \leq m \leq j$: $h(x_m) = 1$ ולכל נקודה שלא שייכת לרצף תקבל $h(x_m) = 0$.

נוכיח שיש לכל היותר k רצפים בכל 2k נקודות. נניח בשלילה כי יש k+1.

כל 2 רצפים מופרדים לפחות על ידי 0 אחד. [ ]0[ ]0[ ]

ולכן במקרה הטוב שבו כל רצף מכיל רק איבר אחד נקבל שהרצפים מכילים k+1 איברים

ובנוסף יש k אפסים. סה"כ לפחות 2k+1 נקודות בסתירה לכך שהקבוצה בגודל 2k.

נראה כי $VC(H_k) < 2k+1$:

יהי קבוצה בגודל 2k+1 $x_1, \dots, x_{2k+1}$ נניח בה"כ $x_1 < x_2 < \dots < x_{2k+1}$.

ניתן לסירוגין לכל איבר 1 ו-0. פורמלית $x_i = (-1)^{i+1}$ $(1,0,1,0,\dots,1)$

נניח והצלחנו למצוא k אינטרבלים. אז כיוון שיש k+1 אחדים מעקרון היונים יש

2 אחדים בתוך אינטרבל, נסמן אותם x_i, x_j. אבל מהגדרת הקבוצה הם ∃ עבורו

$x_i < x_m < x_j$ ולכן האינטרבל $[x_i, x_j]$ מכיל נקודה שהלייבל שלה 0 בסתירה.

סה"כ קיבלנו כי $VC(H_k) = 2k$

למידה חישובית

(4) נוכיח כי $VC(H_{poly})=\infty$

נראה כי לכל n קיים קבוצה C של n משתנים שמנותצת ע"י H_{poly}.

נבחר $C=\{(x_1,0),\dots,(x_n,0)\}$ כאשר ה-x_i בסדר עולה.

לפי הרמז לכל $x_1,\dots,x_n\in\mathbb{R}$ שונים קיימים $z_1,\dots,z_n\in\mathbb{R}$ שונים

קיים פולינום p מדרגה $n-1$ כך שלכל מתקיים: $p(x_i)=z_i$

יהי $s_1,\dots,s_n\in\{0,1\}$ קומבינציה.

לכל מתקיים אם $s_i=0$ נבחר $(x_i,-1)$ אחרת $(s_i=1)$ $(x_i,0)$

יהי הפולינום p המקיים $p(x_i)=\begin{cases}0 & s_i=1\\ -1 & s_i=0\end{cases}$ (לפי הרמז)

ולכן לכל מתקיים

$$h_p(x_i,0)=\begin{cases}1 & p(x_i)\ge 0\\ 0 & \text{else}\end{cases}=$$

$$=\begin{cases}1 & s_i=1\\ 0 & s_i=0\end{cases}=s_i \Rightarrow |H_{poly_C}|=2^{|C|}$$

כלומר קיבלנו כי $VC\dim(H_{poly})\ge n$ וכיוון ש-n לא קבוע וניתן

להגדיל כרצוננו אז $VC\dim(H_{poly})=\infty$

חלק מעשי

הערה: אנו מסתכלים על הקטע $[0,1]$ כלומר אם $x\notin H$ אז $x\in\{(0.2,0.4),(0.6,0.8)\}$

1c) $e_p(h)=E[\Delta(Y,h(x))]$ נחשב תחילה בנפרד: $H=\{(0,0.2),(0.4,0.6),(0.8,1)\}$

$$(*)\quad P(h(x)=1,Y=0)=P(Y=0\mid h(x)=1)P(h(x)=1)=\overbrace{P(Y=0\mid x\in H)}^{0.2}P(x\in H,h(x)=1)+$$

$$+\overbrace{P(Y=0\mid x\notin H)}^{0.9}P(x\notin H,h(x)=1)=0.2\,P(x\in H,h(x)=1)+0.9\,P(x\notin H,h(x)=1)$$

$$(**)\quad P(Y=1,h(x)=0)=P(Y=1\mid h(x)=0)P(h(x)=0)=\overbrace{P(Y=1\mid x\in H)}^{0.8}P(h(x)=0,x\in H)+$$

$$+\overbrace{P(Y=1,x\notin H)}^{0.1}P(x\notin H,h(x)=0)=0.8\,P(h(x)=0,x\in H)+0.1\,P(x\notin H,h(x)=0)$$

אנחנו רוצים להביא למינימום את הביטויים הללו כדי להקטין כמה שיותר את ההסתברות.

אז בעצם נקבל שכדי שהחיתוך בנפרד יהיה הכי קטן, נרצה הפלה כלומר דווקא אמא $x\in H$ $h(x)=1$

ואם $x\notin H$ אז $h(x)=0$, כלומר המקטעים הכי טובים הם בדיוק H ו-$k=3$ היה האופטימלי. והמתאים עבור H_{10}

נסכם הכל ביחד:

$$e_p(h)=E[\Delta(Y,h(x))]=P(h(x)\neq Y)=(*)+(**)=$$

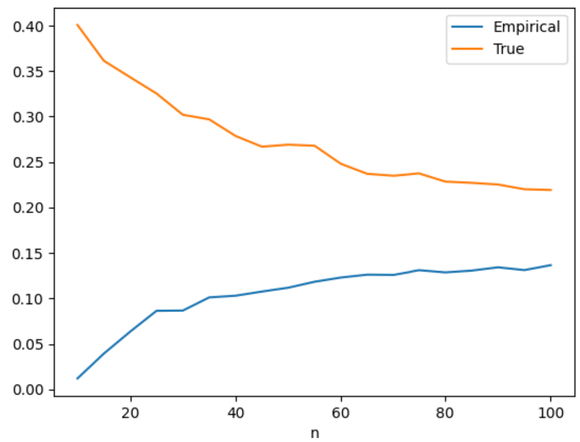
$$=0.2\,P(x\in H,h(x)=1)+0.9\,\cancel{P(x\notin H,h(x)=1)}+0.8\,\cancel{P(x\in H,h(x)=0)}+0.1\,P(x\notin H,h(x)=0)=$$

המקרים שהתאפסו לא יכולים לקרות עבור ההיפותזה הכי טובה כפי שהראנו למעלה

$$=0.2\cdot P(x\in H)+0.1\,P(x\notin H)=0.2\cdot0.6+0.1\cdot0.4=\underline{\underline{0.16}}$$

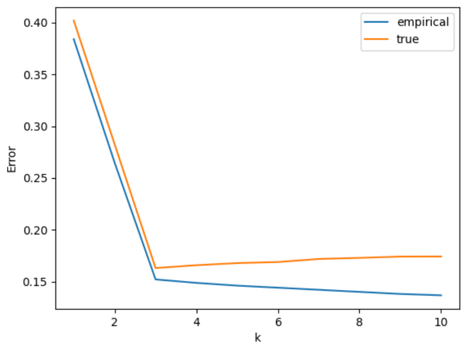
המשך – גרפים

סעיף ב'



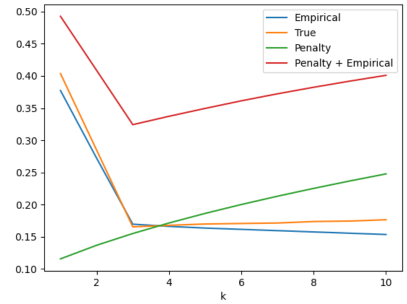
ניתן לראות מהגרף כי השגיאה האמיתית יורדת ככל שגודל המדגם גדל כי ככל שכמות הדוגמאות גדלה אנחנו מתקרבים יותר להסתברות האמיתית של העולם P.
השגיאה האמפירית עולה ככל שגודל המדגם עולה מפני שככל שיש יותר דגימות הסיכוי לקבל תוצאות קיצוניות גדל ולכן השגיאה עולה בהתאם.
ניתן גם לראות שהערכים מתקרבים אחד לשני וזה גם מסתדר כי התוחלת של השגיאה האמפירית היא השגיאה האמיתית ולכן ככל שנגדיל את גודל המדגם הערכים יתקרבו אחד לשני.

סעיף ג'



ניתן לראות מהגרף כי השגיאה האמפירית יורדת ככל ש k גדל, ואילו השגיאה האמיתית מגיעה לתחתית ב k=3 ואז מתחילה לעלות. קיבלתי כי $k^* = 10$, אבל נשים לב כפי שציינתי כבר עבור k=3 מקבלים את השגיאה האמיתית הנמוכה ביותר, ולכן הבחירה של k^* לא מוצלחת. מצב זה נוצר כתוצאה מאימון יתר של הנתונים שלנו שגורם לרגישות יתר לרעשים ותוצאות קיצוניות.

סעיף ד'



נשים לב שהקנס עולה ככל ש k גדל וזה הגיוני כי המימד גדל כלומר ההיפותזה מסובכת יותר ולכן נצטרך לשלם קנס גבוה יותר, כמובן שניתן גם לראות את זה מחישוב הקנס שיש יחס ישיר בין הקנס ל k.
הערך המינימלי של ה penalty + empirical error מתקבל עבור k=3 וזה זהה לתוצאה מהסעיף הקודם.

סעיף ה'

```
[(0.00021516523452441705, 0.19929980245149892), (0.400244140438411, 0.5993662634659548), (0.800076854255245, 0.9977186127349789)]
best k is 3

Process finished with exit code 0
```

לאחר הרצת האלגוריתם קיבלנו כי ההיפותזה הכי טובה התקבלה עבור k=3.
ללא הרצת ה validation היינו אולי יכולים לקבל היפותזה עם שגיאה אמפירית נמוכה יותר כי שראינו בסעיף קודם עם k^* והיינו בוחרים ב k הלא נכון. נשים לב גם שהתוצאה קרובה מאוד להיפותזה האופטימלית שראינו בסעיף א'.