מבוא ללמידה חישובית תרגיל בית 2

ד) נגדיר פונקציה $A(S)$ על $S=\{(x_1,y_1),\dots(x_m,y_m)\}$ באופן הבא:

- אם קיים i כך ש- $y_i=1$ נחזיר את h_{x_i}
- אחרת נחזיר h^-

<u>הוכחה:</u>

נניח בשלילה שקיימת פונקציה שממזערת את $e_S(h)$ יותר מ- $A(S)$.

נניח שאין אף i כך ש- $y_i=1$. אז $A(S)=h^-$

ומתקיים - $e_S(h^-)=\sum_{j=1}^{m}\Delta_{ZO}(0,y_j)=0$ ולכן $e_S(h^-)$ המינימלי
ביותר האפשרית, כנדרש.

נניח שכן קיים i עבורו $y_i=1$. אז $A(S)=h_{x_i}$

ומתקיים: $e_S(h_{x_i})=\sum_{j=1}^{m}\Delta_{ZO}(h_{x_i}(x_j),y_j)=\left[\sum_{\substack{j=1\\j\neq i}}^{m}\Delta_{ZO}(0,y_j)\right]+\Delta_{ZO}(1,1)=$

$$=\sum_{\substack{j=1\\j\neq i}}^{m}\Delta_{ZO}(0,y_j)$$

אבל בגלל שידוע ש- f מחזירה 1 רק על לכל היותר ערך יחיד של x: $\Delta_{ZO}(0,y_j)=\Delta_{ZO}(0,0)=0$ ולכן סה"כ $e_S(h_{x_i})=0$ כנדרש, ולכן בכל מקרה, $e_S(A(S))=0$ בסתירה להנחת השלילה.

כיוון 1 - נניח ש-$E[e_p(A(S))]\le\alpha$ (*) ונוכיח ש-H

PAC למידה. (*) עבור $|S|>N(\alpha)$

יהיו δ,ε. נגדיר $N^*=N(\varepsilon\cdot\delta)$. יהי S (מדגם) כך

ש-$|S|>N^*$ ונקבל:

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

כנדרש.

המעבר הראשון נובע מא"ש מרקוב והשני מכך

שהגדרנו $\delta\cdot\varepsilon=\alpha$.

יקיים $N:(0,1)^2\to R$ כך:

כיוון 2 - נניח שלכל δ,ε ולכל מדגם S, כך

ש-$|S|>N(\varepsilon,\delta)$ מתקיים $P[e_p(A(S))>\varepsilon]<\delta$.

נגדיר $N':(0,1)\to R$ כך ש- $N'(\alpha)=N(\frac{\alpha}{2},\frac{\alpha}{2})$

נקבל: עבור מדגם S כך ש- $|S|>N'(\alpha)$:

$$E[e_p(A(S))]=E[e_p(A(S))\mid e_p(A(S))>\tfrac{\alpha}{2}]\cdot P[e_p(A(S))>\tfrac{\alpha}{2}]+$$
$$+E[e_p(A(S))\mid e_p(A(S))\le\tfrac{\alpha}{2}]\cdot P[e_p(A(S))\le\tfrac{\alpha}{2}]$$

נסתכל על ארבעת הביטויים הנ"ל ונחסום

כל אחד מהם מלמעלה:

1) $E[e_p(A(S))\mid e_p(A(S))>\frac{\alpha}{2}]\le 1$

הסבר - כי זו קבוצה של e_p, תמיד התוחלת

של $e_p(A(S))$ היא בין 0 ל-1 מהגדרת e_p

2) $P[e_p(A(S))>\frac{\alpha}{2}]<\frac{\alpha}{2}$

הסבר - לפי ההנחה של PAC-Learnable עם

$\varepsilon=\delta=\frac{\alpha}{2}$

3) $E[e_p(A(S))\mid e_p(A(S))\le\frac{\alpha}{2}]\le\frac{\alpha}{2}$

הסבר - בהינתן $e_p(A(S))\le\frac{\alpha}{2}$ גורר שגם התוחלת

של $e_p(A(S))$ היא לכל היותר $\frac{\alpha}{2}$ תוחלת של מ"מ

יכולה להיות גדולה יותר מהחסם העליון של אותו מ"מ

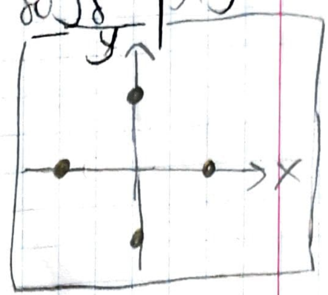
4) $P[e_p(A(S))\le\frac{\alpha}{2}]\le 1$

הסבר - הסתברות, תמיד קטנה או שווה ל-1

לסיכום קיבלנו חסם - $E[e_p(A(S))]\le 1\cdot\frac{\alpha}{2}+\frac{\alpha}{2}\cdot 1=\alpha$

כנדרש.

מימד ה-VC הוא 4. נראה דוגמה של 4 נקודות שניתן לנפץ - $S=\{(-1,0),(0,1),(1,0),(0,-1)\}$. עבור משפחה כזו:

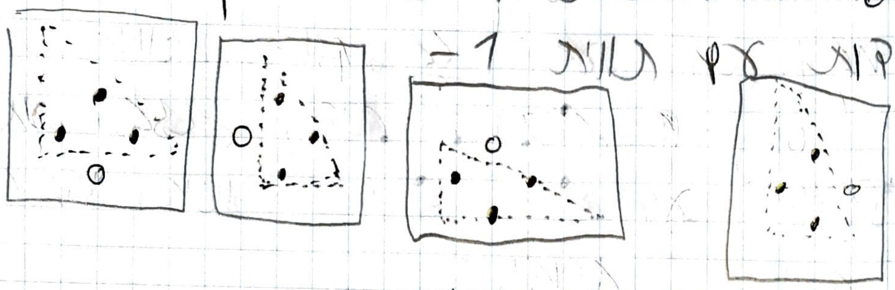


אחיים כל אפשרויות התיוג -,

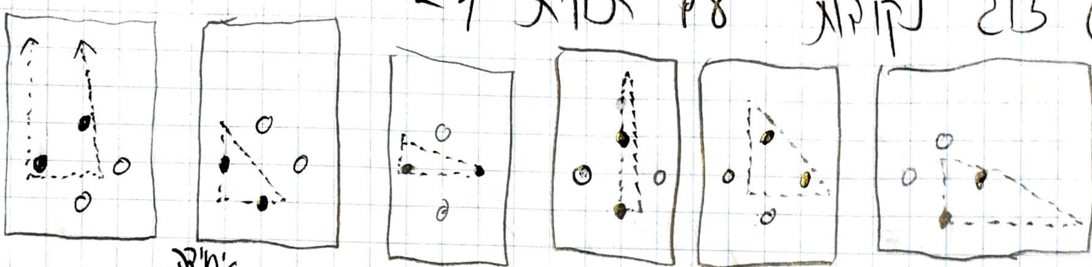
1) כל התוויות 1/0 - כמובן שקל לצייר משולש שיעטוף את כל הנק' או אף אחת מהן

2) נקודה אחת עם תווית 1 - נעטוף רק אותה במשולש קטן

3) 3 נקודות עם תווית 1 -



4) זוג נקודות עם תווית 1 -

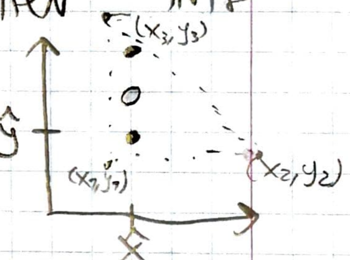


בנוסף יהיו 5 נקודות. בה"כ יש נקודה יחידה הכי שמאלית נמוכה (x_0,y_0) ונקודה יחידה הכי תחתונה, (x_1,y_1). לכן כל משולש שיכיל את שני הנקודות האלה מוכרח קיים שפינתו השמאלית תהיה ב-$(\hat{x},\hat{y})$ כך ש- $\hat{x}\leq x_0$ ו- $\hat{y}\leq y_1$. בה"כ $\hat{x}=x_0, \hat{y}=y_1$ (עבור צי יהיה משולש שמשש עבור כל הנקודות הנ"ל. בה"כ גם ש $\hat{x}=\hat{y}=0$, עבור הפינה השמאלית של המשולש היא $(0,0)$ אם זה לא המקרה תמיד אפשר להזיז את כל הנקודות מהתחום. כעת נשארנו עם הבעיה - האם ניתן לנפץ את 3 הנקודות שנותרו עם קו ישר אחד שחותך את שני הצירים. אם 3 הנק' על אותו ישר - סיימנו (כי אפשר לתייג 2 את האמצעית 0). אחרת תמיד יהיו שתי נקודות שכדאי להוציא אחת מהן מחוץ למשולש הקו יצטרך להשאיר את השנייה בתוך המשולש, לכן לא ניתן לנפץ את 3 הנק' כנדרש.

4) המימד הוא 2. הוכחה: $\{(x_1,y_1),(x_2,y_2)\}$, בה"כ שונות

- קל לראות שעבור כל שתי נקודות תמיד ניתן להגדיר משולש לכל אחת מהפונקציות ה-Label האפשריות:
  - ש"ח: $f(x_1,y_1)=f(x_2,y_2)=0$ - נגדיר משולש ריק בנקודה כלשהי $\hat{x}\notin\{x_1,x_2\}$
  - ש"ח: $f(x_1,y_1)=f(x_2,y_2)=1$ - נגדיר משולש שיקיף את שתי הנקודות. למשל שים אותן לצלע: $\{(x_1,y_1),(x_1,y_2),(x_2,y_1),(x_2,y_2)\}$ ונמתוח אותו
  - ש"ח $f(x_1,y_1)\neq f(x_2,y_2)$. בה"כ $f(x_1,y_1)=1$
    אם $x_1>x_2$, נעביר את האנך ב- $\frac{x_1+x_2}{2}$ ונסיים
    אחרת נעביר את הישר כך שיחצה בין הנקודות ונסיים

- נוכיח שיש מקבץ של 3 נקודות שלא ניתן לנפץ" ע"י משולש כזה. נגדיר $\hat{x},\hat{y}\in\mathbb{R}$ ושלוש נק': $\{(\hat{x},\hat{y}),(\hat{x},\hat{y}+1),(\hat{x},\hat{y}+2)\}$
  עם פונקציה - $f(\hat{x},\hat{y})=f(\hat{x},\hat{y}+2)=1, f(\hat{x},\hat{y}+1)=0$
  נניח משולש כזה - נקודה שחורה = Label 1.



שהישר ביניהן מקביל

נבחר שתי הנקודות לציר ה-x, נסמן $(x_1,y_1),(x_2,y_2)$ חייבות לקיים $y_1=y_2$, $x_1<\hat{x}<x_2$ כדי שהמשולש יכיל את $(\hat{x},\hat{y})$. אם נרצה שים לב שבכדי ש- $(\hat{x},\hat{y}+2)$ תהיה בתוך המשולש ח"ב להתקיים:
$\hat{y}+2<g(\hat{x})$ כאשר g הוא משוואת הישר $\{(x_3,y_3),(x_2,y_2)\}$ (כאשר הישר עובר בין הנקודות האלה). אבל אז גם $\hat{y}+1<g(\hat{x})$ ולכן גם $(\hat{x},\hat{y}+1)$ בתוך המשולש בניגוד להנחה.

3) עבור k אינטרוולים נקבל $VCdim(H_k)=2k$

הוכחה:

- יהי מדגם של $2k+1$ נקודות, $S=\{x_1,\dots,x_{2k+1}\}$

כך ש- $x_1<x_2<\dots<x_{2k+1}$

נראה שלא ניתן לנפץ בעזרת k אינטרוולים של חלוקה. נגדיר $f(x_i)=\begin{cases}1, & i\in N_{odd}\\ 0, & i\in N_{even}\end{cases}$. נניח בשלילה שקיים $h\in H_k$ כך ש- $h(x_i)=f(x_i)$ לכל $1\le i\le 2k+1$.

אז לכל $1\le j\le k$ $h(x_{2j+1})=1$ $h(x_{2j})=0$

כלומר קיים אינטרוול $I_j=(a_j,b_j)$ כך ש- $x_{2j}\le a_j<x_{2j+1}\le b_j$. קיבלנו $k+1$ אינטרוולים, סתירה.

- נראה שכל מדגם בגודל $2k$ ומטה ניתן "לנפץ" בעזרת k אינטרוולים. ראשית, קל לראות שאם המדגם הוא קטן מ- $2k$ נוכל להוסיף $2k-|S|$ נקודות זהות לאחת הנקודות במדגם ועדיין לנפץ את המדגם החדש בגודל $2k$ נקודות. לכן מספיק לנפץ את המקרה של $2k$.

תהי f פונקציית Labeling כלשהי, $f:[0,1]\to\{0,1\}$. נראה שקיימת $h\in H_k$ כך ש- $h(x_i)=f(x_i)$

שים לב שב- $2k$ נקודות יש לכל היותר $2k$ "חילופים", כלומר נקודה i כך ש- $f(x_i)\neq f(x_{i+1})$ או $i=2k$, כלומר הנקודה האחרונה.

נסמן את הנקודות האלה, בהן מתבצע "חילוף" ב- $\{\hat{x}_1,\dots,\hat{x}_m\}$ כאשר $m\le 2k$. בה"כ הנקודות האלה ממוינות, כלומר $\hat{x}_j<\hat{x}_{j+1}$ לכל $1\le j<m$. נגדיר את האינטרוול ה- j להיות $[\hat{x}_j,\hat{x}_{j+1})$. כלומר $h(x_j)=1$, $h(x_{j+1})=0$ ולכל נקודה אחרת i - $\hat{x}_j<x_i<x_{j+1}$ מתקיים כנדרש $h(x_i)=h(\hat{x}_j)$ בסוף גם כבר נקבל או שמתקיים לכל x $h(x)=f(x)$ או $h(x)\neq f(x)$ אם והתחלנו בנקודה "לא נכונה" (אם $f(\hat{x}_1)=1$) במקרה כזה נחליף את האינטרוולים המשלימים - $h'(S)=[0,1]\setminus h(S)$ ונקבל את התוצאה הנדרשת.

5) נגדיר $\varepsilon = \sqrt{\frac{1}{2n}\ln\left(\frac{2}{\delta\cdot w(h)}\right)}$

א) מהיון שמתקיים $\forall h\in E:\ P[e_p(h)-e_s(h)>\varepsilon]\le 2e^{-2n\varepsilon^2}$

נציב ונקבל:

$$P[e_p(h)-e_s(h)>\varepsilon]\le 2e^{-2n\cdot\frac{1}{2n}\ln\left(\frac{2}{\delta\cdot w(h)}\right)}=$$

$$=2e^{-\ln\left(\frac{2}{\delta\cdot w(h)}\right)}=\frac{2}{\frac{2}{\delta\cdot w(h)}}=\delta\cdot w(h)$$

נקבל עבור H כולה: Union Bound

$$P\left[\bigcup_{i\in N} e_p(h_i)-e_s(h_i)>\varepsilon\right]\le\sum_{i\in N}P[e_p(h_i)-e_s(h_i)>\varepsilon]\le$$

$$\le\sum_{i\in N}\delta\cdot w(h_i)=\delta\cdot\sum_{i\in N}w(h_i)\le\boxed{\delta}$$ כנדרש.

המעבר האחרון נובע מהנתון: $\sum_{h\in H}w(h)\le 1$

(הערה: שמתי לב שעכשיו צריך להוכיח בהסתברות $\delta-1$ ולא δ, נגדיר $\delta'=1-\delta$ ונשתמש באותה הוכחה עבור δ' וסיימנו.)

ב. לפי סעיף א' מתקיים לכל h בסיכוי $\delta-1$:

נסמן $T_h=\sqrt{\frac{1}{2n}\ln\left(\frac{2}{\delta w(h)}\right)}$

$$e_p(h)\le e_s(h)+T_h$$

מהגדרה -

$$|e_p(h)-e_s(h)|\le T_h$$

נפתח את הערך המוחלט ונקבל:

$$(***)\quad e_p(h)\le e_s(h)+T_h,\quad e_s(h)\le e_p(h)+T_h\quad(*)$$

לפי ההגדרה של h_{srm} מתקיים:

$$(**)\quad e_s(h_{srm})+T_{h_{srm}}\le e_s(h)+T_h$$

לכן מחיבור $(*)$ ו-$(**)$ נקבל:

$$e_s(h_{srm})+e_s(h)+T_{h_{srm}}\le e_p(h)+e_s(h)+2T_h$$

ומכיוון שלפי $(***)$ מתקיים גם עבור h_{srm}:

$$e_p(h_{srm})\le e_s(h_{srm})+T_{h_{srm}}\Longrightarrow e_p(h_{srm})+e_s(h)\le e_p(h)+e_s(h)+2T_h$$

ולכן $e_p(h_{srm})-e_p(h)\le 2T_h$ כנדרש (בפרט $h=h_{min}$)

(6

$$E[\Delta_b(Y,h(X))] = \sum_{x,y} P[X=x, Y=y]\cdot \Delta_b(y,h(x)) =$$

$$= \sum_{x} P[X=x, Y=0]\cdot \Delta_b(0,h(x)) + P[X=x, Y=1]\cdot \Delta_b(1,h(x)) =$$

$$= \sum_{x} P[X=x]\left(\underbrace{P[Y=0|X=x]\cdot \Delta_b(0,h(x)) + P[Y=1|X=x]\cdot \Delta_b(1,h(x))}_{\circledast}\right) =$$

נרצה למזער את הביטוי הנ"ל עבור x נתון, $\hat{x}$.

ע"י ~~[illegible]~~ למזער את (ביטוי) $\circledast$:

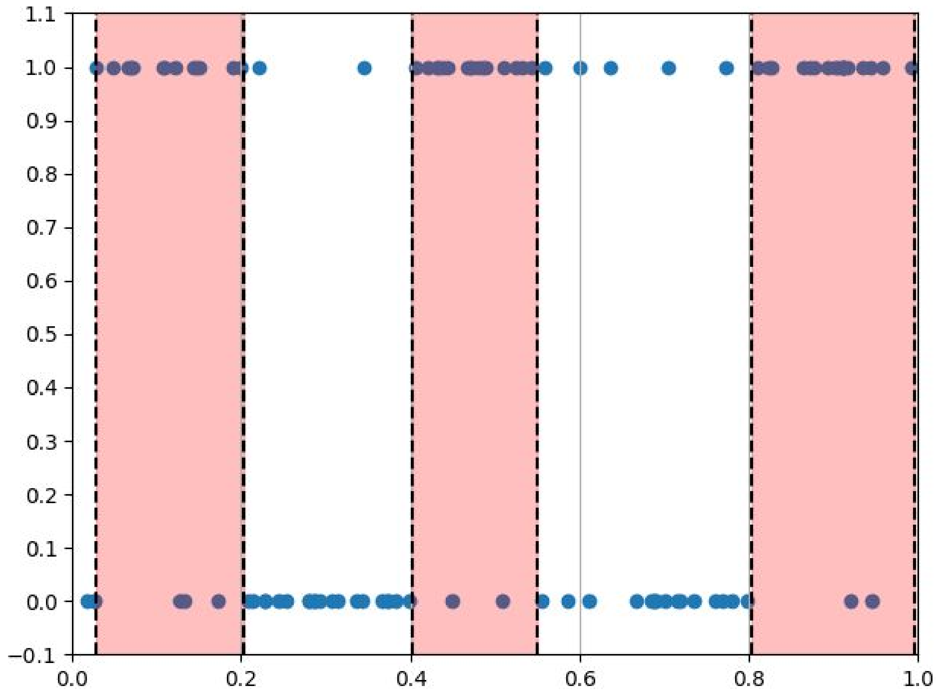
$$\circledast = \begin{cases} P[Y=0|X=\hat{x}], & h(\hat{x})=1 \\ P[Y=1|X=\hat{x}]\cdot \frac{1}{2}, & h(\hat{x})=0 \end{cases} = \begin{cases} 1-P[Y=h(\hat{x})|X=\hat{x}], & h(\hat{x})=1 \\ \frac{1}{2}\left(1-P[Y=h(\hat{x})|X=\hat{x}]\right) & h(\hat{x})=0 \end{cases}$$

לכן נרצה למצוא את h:

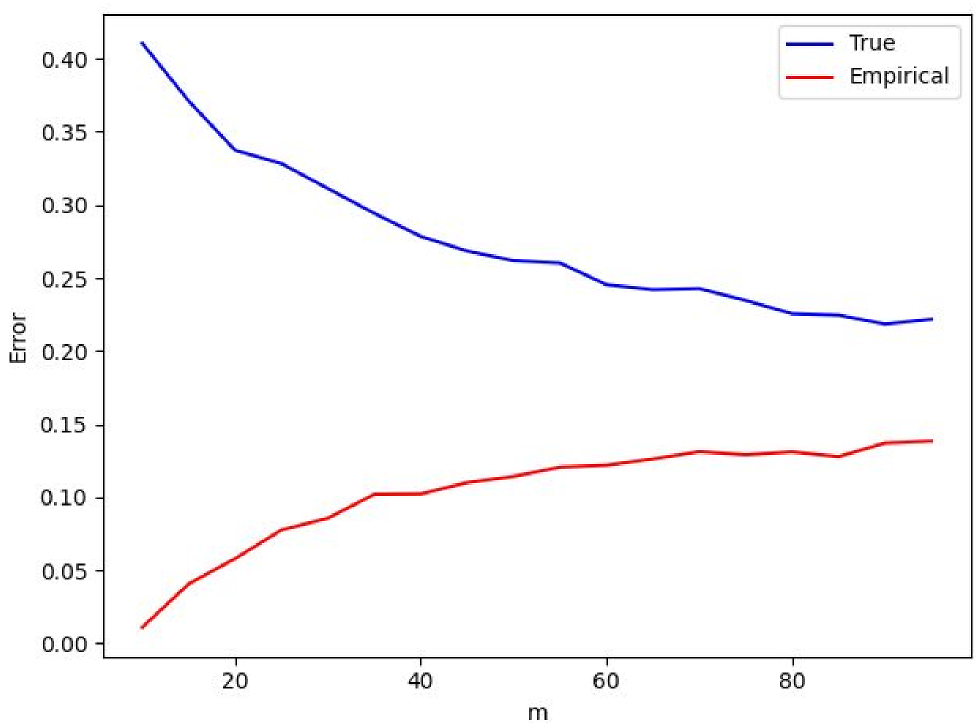
$$h(\hat{x}) = \underset{y\in\{0,1\}}{\arg\max}\left[P[Y=y|X=\hat{x}]\cdot b(y)\right]$$

כאשר $b(y)=\begin{cases} \frac{1}{2}, & y=1 \\ 1, & y=0 \end{cases}$

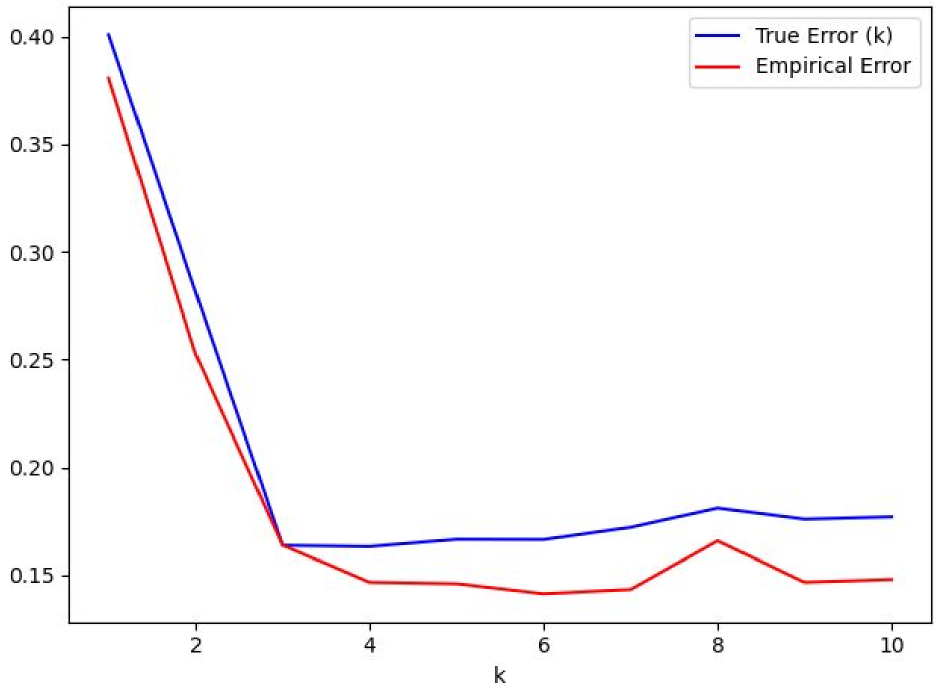
1) The red vertical stripes are the intervals provided by the find_best_interval:



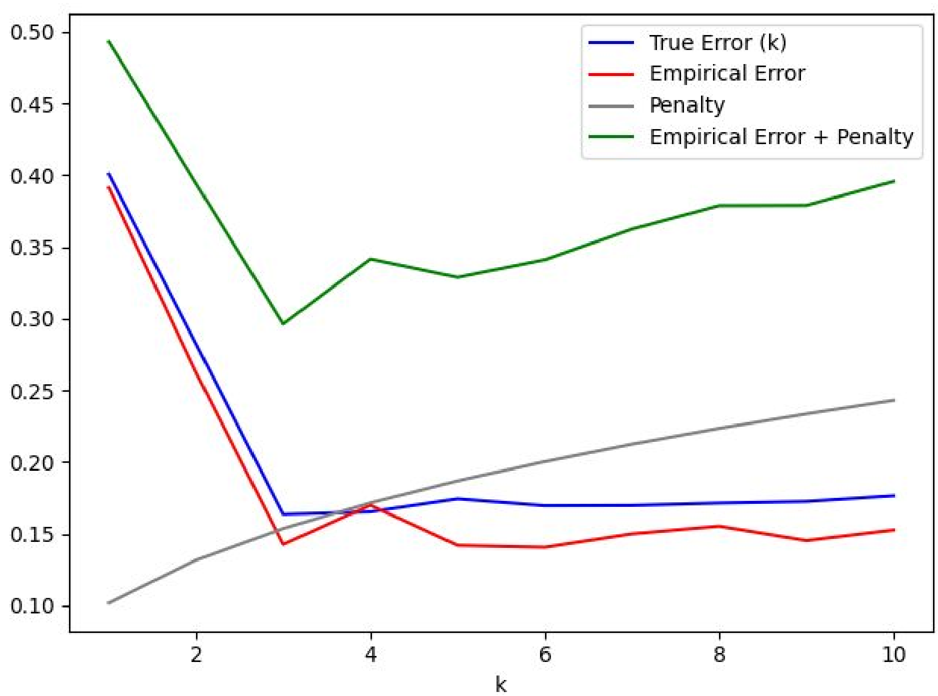
2) The best hypothesis is: h(x) = 1 if x is in one of the intervals {(0, 0.2), (0.4, 0.6), (0.8, 1)}, 0 else
3) As seen in the following chart the true error decreases as m increases which makes sense for two reasons:
    a) The true error should converge to its minimum as m goes to infinity, while the empirical error should converge to the true error
    b) Empirical error is smaller for few examples since it’s easy to train a model based on very little examples (for example if m≈k)



4) k* intervals doesn't necessarily constitute a good hypothesis, it may cause overfitting. This explains why k* is 6 but the best k in terms of true error is 3



5) Here we can clearly see the expected result - adding the penalty function causes the best k to be 3. Without the penalty function it's still 6



6) In all 3 tests the best hypothesis was based on k=3. So to find the optimal hypothesis itself all we needed is to run find_best_intervals with k=3 and a big enough sample