מבוא ללמידה חישובית תרגיל בית 6

1) א. נשים לב שניתן לבטא כל פונקציה ב-F כשרשור (concatenation) של d פונקציות ב-H:

$f\in F \Rightarrow f=f_1\times f_2\times\dots\times f_d$, $\forall_{1\le i\le d}: f_i\in H$

הוכחנו בנוסף ש- $\Pi_F(n)\le \Pi_{F_1}(n)\cdot\Pi_{F_2}(n)$ כאשר

$F=\{(f_1\times f_2) \mid f_1\in F_1, f_2\in F_2\}$

לכן:

$$\Pi_F(n)\le \Pi_H(n)\cdot\ldots\cdot\Pi_H(n)=\Pi_H(n)^d$$

נניח $n\ge d+1$ לכן מהלמה של Sauer-Shelah:

$$\Pi_F(n)\le \Pi_H(n)^d\le\left(\left(\frac{en}{d+1}\right)^{d+1}\right)^d=\left(\frac{en}{d+1}\right)^{d(d+1)}$$

כנדרש

ב. נשים לב שהמחלקה C מגדירה הרכבה של L-1 שכבות של Hyperplanes ממימד d ועוד שכבה אחת ממימד 1. בשכבה, כל מסווג בה הוא שרשור של פונקציות ב-H. כלומר:

$$C=\{f^{(0)}\circ f^{(1)}\circ f^{(2)}\circ\dots\circ f^{(L-1)} \mid f^{(0)}\in H\wedge\forall i\ge 1: f^{(i)}=f_1^{(i)}\times\dots\times f_d^{(i)}\wedge\forall_{1\le j\le d}: f_j^{(i)}\in H\}$$

הוכחנו בתרגול לכל הרכבה מתקיים א'-שוויון בצורה ל-$\Pi_C(n)$: $\Pi_C(n)\le\Pi_{C_1}(n)\cdot\Pi_{C_2}(n)$ כאשר $C=\{f_1\circ f_2 \mid f_1\in C_1\wedge f_2\in C_2\}$.

לכן:

$$\Pi_C(n)\le\Pi_H(n)\cdot\left[\prod_{i=1}^{L-1}\Pi_{F^{(i)}}(n)\right]\le\left(\frac{en}{d+1}\right)^{d+1}\left[\left(\frac{en}{d+1}\right)^{d(d+1)}\right]^{L-1}=$$

$$=\left(\frac{en}{d+1}\right)^{d+1+d(d+1)(L-1)}$$

1) ה) נבחר $N=d+1+d(d+1)(L-1)$ ונקבל מטענה ג'

$$\Pi_C(n) \le \left(\frac{en}{d+1}\right)^N \le (en)^N$$

נסמן $n=\mathrm{VCdim}(C)$. נקבל:

$$2^n = \Pi_C(n) \le (en)^N \overset{\text{מטענה ג'}}{\Longrightarrow} n \le 2N\log_2(eN)$$

ולכן $\boxed{\mathrm{VCdim}(C) \le 2N\log_2(eN)}$ כנדרש

---

1) ב) [סליחה על ההערבוב בסדר, הלחץ מוציא ממני טיפשות ושכחתי את סעיף ג' :)]

הפרמטרים ברשת הם המשקולות W וה-bias-ים b.
המשקולות $W^{(t)}$ בין השכבה ה-t לשכבה ה-$t+1$ הן מטריצה, שמגדירות פונקציה $\mathbb{R}^d \to \mathbb{R}^d$ כלומר בגודל d^2. יש $L-1$ כאלו ועוד מטריצה שמייצגת פונקציה $\mathbb{R}^d \to \mathbb{R}$ בשכבה האחרונה, כלומר בגודל d. בנוסף לכל נוירון יש bias (b) בגודל 1.
כלומר יש $d^2(L-1)+d$ פרמטרים מה-W-ים
ועוד $d(L-1)+1$ מה-b-ים. סה"כ:

$$N = d+1+d(d+1)(L-1)$$

ג) נעקוב אחר ריצת האלגוריתם:

• חישוב $j \leftarrow \underset{i\in\{1,2,3\}}{\arg\max}\, G(S,i)$:

$$G(S,1) = L(Y) - P(X_1=1)\cdot L(Y|X_1=1) - P(X_1=0)\cdot L(Y|X_1=0) =$$

נזכיר -

- $L(Y) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$
- $P(X_1=1)=\frac{3}{4} \quad P(X_1=0)=\frac{1}{4}$
- $L(Y|X_1=1) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918$
- $L(Y|X_1=0) = -0\log 0 - 1\log 1 = 0$

נקבל -

$$= 1 - \frac{3}{4}\cdot 0.918 \approx 0.311$$

במקרה של $i=2$ ו- $i=3$ יש שקילות מושלמת לכן נחשב רק את $i=2$:

$$G(S,2) = L(Y) - P(X_2=1)\cdot L(Y|X_2=1) - P(X_2=0)\cdot L(Y|X_2=0) =$$

נזכיר -

- $L(Y)=1$
- $P(X_2=1)=\frac{1}{2},\ P(X_2=0)=\frac{1}{2}$
- $L(Y|X_2=1) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$
- $L(Y|X_2=0) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$

נקבל -

$$= 1 - \frac{1}{2} - \frac{1}{2} = 0$$

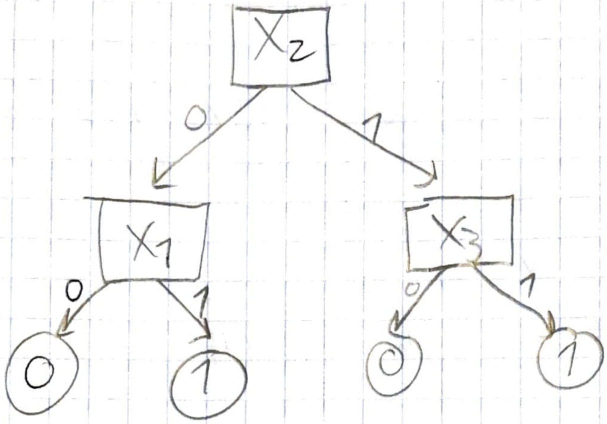
לכן $j=1$. $S_0=\{(0,0,1)\}$, $S_1=\{(1,1,1),(1,0,0),(1,1,0)\}$

עבור צאצא הזה של S_0 – כל הערכים, (כלומר הערך היחיד ב- S_0), שווים ל-0 לכן נחזיר עלה עם ערך 0.

בעת הבנייה של S_1 ניתן לראות בקלות שאין הבדל בין בחירת $j=2$ ל- $j=3$ בשניהם כל בחירה של $x_j=0$ או $x_j=1$ מקבלים קבוצות עם שתי התוויות כלומר בכל מקרה נטעה בדוגמה אחת.

סה"כ נטעה בדוגמה אחת מתוך הארבע כנדרש.

2) ב. עץ החלטה עם training error=0:



---

5) ב. נכתוב את בעיית המינימיזציה-

$$\min\|w\|^2 \quad \text{s.t. } x_1w=y_1,\dots,x_nw=y_n$$

ניתן לפתור עם לגרנז'יאן:

$$L(w;\lambda)=\|w\|^2+\sum_{i=1}^{n}\lambda_i(x_iw-y_i)$$

נגזור ונשווה ל-0: $\frac{\partial L}{\partial w}=2w+\sum_{i=1}^{n}\lambda_i x_i=0 \Longrightarrow w=-\frac{1}{2}\sum_{i=1}^{n}\lambda_i x_i$

לכן הווקטור מקדמים הוא $\alpha=-\frac{1}{2}\begin{pmatrix}\lambda_1\\ \vdots\\ \lambda_n\end{pmatrix}$

ג. יהי וקטור $\hat{x}\in\mathbb{R}^d$. נרצה לחשב את $\hat{x}^t w^*$:

$$\hat{x}^t w^*=\hat{x}^t(X^t\alpha)=$$

$$=(\hat{x}^t X^t)\alpha=$$

$$=[\hat{x}\cdot x_1^t,\dots,\hat{x}\cdot x_n^t]\begin{bmatrix}\alpha_1\\ \vdots\\ \alpha_n\end{bmatrix}$$

(תזכורת: $\hat{x}\in\mathbb{R}^d, X\in\mathbb{R}^{nd}$)

כאשר כל מכפלה בגוף היא dot product של וקטורים ב-$\mathbb{R}^d$. לכן אם $\hat{x}=x_j$ (השורה ה-j ב-X) אז:

$$x_j w^*=[x_j\cdot x_1,\dots,x_j\cdot x_n]\cdot\begin{bmatrix}\alpha_1\\ \vdots\\ \alpha_n\end{bmatrix}=K_{s_j}\cdot\alpha$$

כנדרש.

3) ג

$$Z_t=\sum_{j=1}^{n} D_t[j]\cdot e^{-y_j w_t h_t(x_j)}$$

נציב $W_t=\frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t}$

נגזור את Z_t כפונקציה של W_t ונשווה

$$\frac{\partial Z_t}{\partial w_t}=\sum_{j=1}^{n} D_t[j]\cdot(-y_j)\cdot h_t(x_j)\cdot e^{-y_j w_t h_t(x_j)}$$

$$=-\sum_{j=1}^{n} D_t[j]\, y_j h_t(x_j)\cdot e^{-y_j\cdot h_t(x_j)\cdot\frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)}=$$

$$=-\sum_{j=1}^{n} D_t[j]\, y_j h_t(x_j)\left[\frac{(1-\varepsilon_t)}{\varepsilon_t}\right]^{\frac{-y_j h_t(x_j)}{2}}=$$

$$=-\underbrace{\sum_{\substack{j=1\\ h_t(x_j)=y_j}}^{n} D_t[j]}_{\varepsilon_t}\sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}}+\underbrace{\sum_{\substack{j=1\\ (h_t(x_j)=-y_j)}}^{n} D_t[j]}_{1-\varepsilon_t}\sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}=$$

$$=-\varepsilon_t\sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}}+(1-\varepsilon_t)\sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}=0$$

כנדרש.

3) ג.

נגדיר פונקציה -

$$g(i) := \begin{cases} 1, & h_t(x) \neq y_i \\ 0, & \text{else} \end{cases}$$

נסמן: $\underset{x\sim D_{t+1}}{P}[h_t(x) \neq y] = \sum_{i=1}^{n} D_{t+1}[i]\, g(i)$

כעת נציב את ההגדרה של D_{t+1}:

$$D_{t+1}[i] = \frac{D_t[i] \cdot e^{-y_i w_t h_t(x_i)}}{\sum_{j=1}^{n} D_t[j] e^{-y_i w_t h_t(x_j)}} \underset{\text{מסעיף א'}}{=} \frac{D_t[i] \cdot e^{-y_i w_t h_t(x_i)}}{2\sqrt{\varepsilon_t(1-\varepsilon_t)}}$$

נציב חזרה ונקבל:

$$\underset{x\sim D_{t+1}}{P}[h_t(x) \neq y] = \sum_{i=1}^{n} \frac{D_t[i] \cdot e^{-y_i w_t h_t(x_i)}}{2\sqrt{\varepsilon_t(1-\varepsilon_t)}} \cdot g(i) =$$

$$\overset{*}{=} \sum_{\substack{i=1 \\ g(i)=1}}^{n} \frac{D_t[i] e^{-y_i \cdot w_t \cdot (-y_i)}}{2\sqrt{\varepsilon_t(1-\varepsilon_t)}} = \sum_{\substack{i=1 \\ g(i)=1}}^{n} \frac{D_t[i]\, e^{w_t}}{2\sqrt{\varepsilon_t(1-\varepsilon_t)}} =$$

$$= \frac{e^{w_t}}{2\sqrt{\varepsilon_t(1-\varepsilon_t)}} \cdot \underbrace{\sum_{\substack{i=1 \\ g(i)=1}}^{n} D_t[i]}_{=\varepsilon_t} = \frac{e^{w_t} \cdot \varepsilon_t}{2\sqrt{\varepsilon_t(1-\varepsilon_t)}} \underset{\text{מסעיף א'}}{=} \frac{\sqrt{\varepsilon_t(1-\varepsilon_t)}}{2\sqrt{\varepsilon_t(1-\varepsilon_t)}} = \frac{1}{2}$$

כנדרש.

$*$ — עבור מקרים כך ש $g(i)=1$ אז $h_t(x) = -y_i$

ב. נובע (כמעט) ישירות מ-ג' - השגיאה של h_t תחת
ההתפלגות ה-t היא $\frac{1}{2}$. אבל גם אלגוריתם שמעריך
לבחור ל-h_{t+1} מסווג עם שגיאה $\gamma - \frac{1}{2}$ עבור
$0<\gamma\le\frac{1}{2}$. כלומר השגיאה של h_{t+1} צריכה להיות קטנה
ממש מחצי לכן שונה $h_{t+1} \neq h_t$

ט) ניקח התפלגות על הא״ש הנתון $(W:=\alpha)$:

$$\mathop{E}_{i\sim D}\left[y_i\sum_{j=1}^{k} w_j h_j(x_i)\right] \ge \mathop{E}_{i\sim D}[\gamma]$$

כמובן ש-γ קבוע לכן $E[\gamma]=\gamma$.

כלומר:

$$\sum_{i=1}^{n} D(i)\cdot y_i \sum_{j=1}^{k} w_j h_j(x_i) \ge \gamma \Rightarrow \sum_{j=1}^{k} w_j \sum_{i=1}^{n} D(i)\, y_i h_j(x_i) \ge \gamma$$

מכיוון ש-$\sum_{j=1}^{k} w_j = 1$ מתקיים שקיים לפחות $\hat{j}$ כלשהו

$\hat{j}$ עבורו $\sum_{i=1}^{n} D(i)\, y_i h_{\hat{j}}(x_i) \ge \gamma$. לכן:

$$\gamma \le \sum_{i=1}^{n} D(i)\, y_i h_{\hat{j}}(x_i) =$$

$$= \sum_{\substack{i=1 \\ h_{\hat{j}}(x_i)=y_i}}^{n} D(i)\, y_i^2 + \sum_{\substack{i=1 \\ h_{\hat{j}}(x_i)=-y_i}}^{n} D(i)(-y_i^2) = \sum_{\substack{i=1 \\ h_{\hat{j}}(x_i)=y_i}}^{n} D(i) - \sum_{\substack{i=1 \\ h_{\hat{j}}(x_i)=-y_i}}^{n} D(i) =$$

$$= P_{i\sim D}[h_{\hat{j}}(x_i)=y_i] - P_{i\sim D}[h_{\hat{j}}(x_i)\neq y_i] =$$

$$= 1 - 2P_{i\sim D}[h_{\hat{j}}(x_i)\neq y_i]$$

נעביר אגפים ונקבל את הנדרש - $P_{i\sim D}[h_{\hat{j}}(x_i)\neq y_i] \le \frac{1}{2}-\frac{\gamma}{2}$

ד) נבנה את H לפי הרמז - נגדיר $2d$

היפותזות על פי המקטע $[a_j,b_j]$ באופן הבא -

$$\forall 1\le j\le 2d.\ j\in N_{odd}\to h_j(x)=\begin{cases}1, & x\ge a_j\\ -1, & x<a_j\end{cases},\quad j\in N_{even}\to h_j(x)=\begin{cases}1, & x\le b_j\\ -1, & x>b_j\end{cases}$$

כלומר יש לנו בעצם d זוגות של היפותזות, כך כל זוג מגדיר ביחד את ההיפותזה ש- $h(x)=1 \iff x\in[a_j,b_j]$ עבור כל d המקטעים $[a_j,b_j]$. נשארנו עם $2d-1$ היפותזות קבועות - $\forall j>2d\ h_j(x)=-1$. ונסמן $K=4d-1$ לפי הרמז.

כעת נגדיר $\forall j.\ \alpha_j=\frac{1}{4d-1}$ ו- $\gamma=\frac{1}{4d-1}$. כמובן ש- $\sum_{j=1}^{k}\alpha_j=1$ ו- $\gamma>0$ כנדרש.

כעת כדי להוכיח שהא"ש הנדרש מתקיים עבור עבור כל $(\hat{x},\hat{y})$ נפצל למקרים - $\hat{y}=1$ אזי, כלומר $\hat{x}$ בתוך המקטע, אזי $\forall 1\le j\le 2d:\ h_j(x)=1$ ו- $\forall j>2d: h_j(x)=-1$ לכן:

$$\hat{y}\sum_{j=1}^{k}\alpha_j h_j(\hat{x})=\sum_{j=1}^{2d}\alpha_j+\sum_{j=2d+1}^{k}-\alpha_j=\frac{2d}{4d-1}-\frac{2d+1}{4d-1}=\frac{1}{4d-1}=\gamma\ge\gamma$$

כנדרש.

כעת לגבי המקרה $\hat{y}=-1$, כלומר $\hat{x}$ לא במקטע, נסמן ב- k^* את האינדקס שעבורו -

$$\forall 1\le j\le k^*: h_j(\hat{x})=1$$
$$\forall k^*<j\le 2d: h_j(\hat{x})=-1$$

ונשאר $\forall 2d<j\le k: h_j(\hat{x})=-1$ ונקבל:

$$\hat{y}\sum_{j=1}^{k}\alpha_j h_j(\hat{x})=-\sum_{j=1}^{k^*}\alpha_j-\sum_{j=k^*+1}^{2d}-\alpha_j-\sum_{j=2d+1}^{4d-1}-\alpha_j=$$

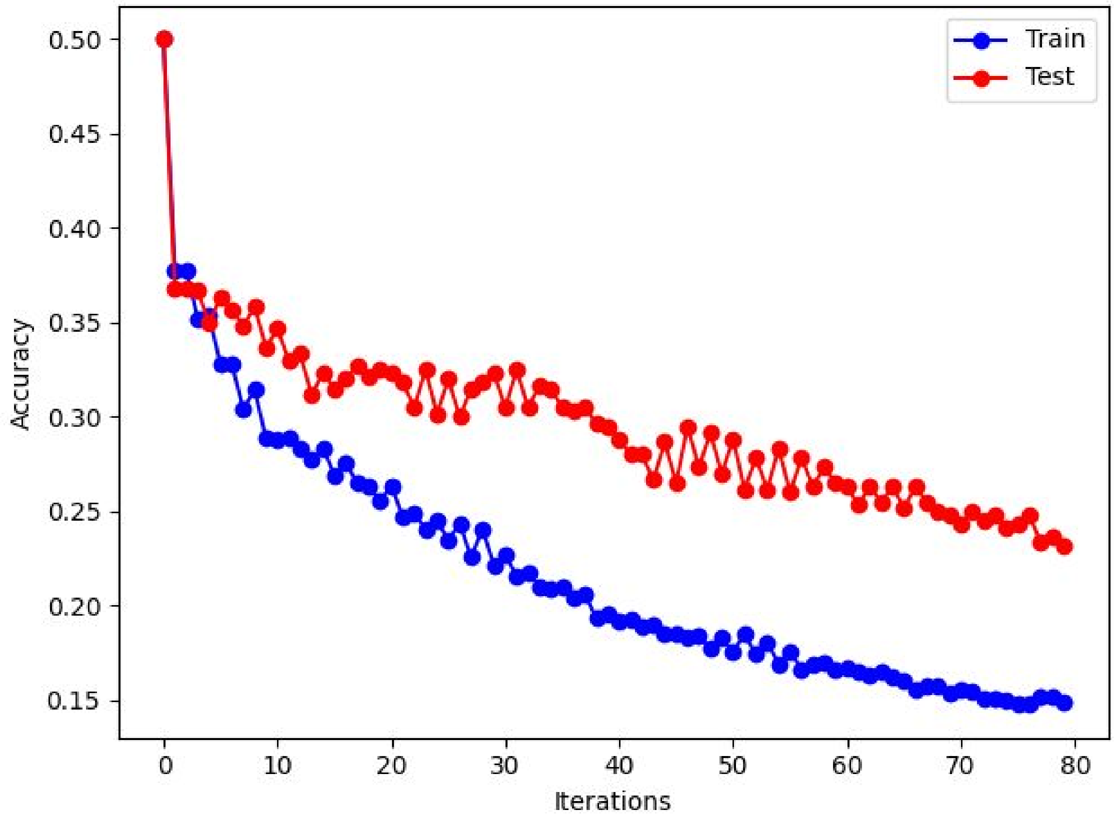
$$=\frac{-k^*}{4d-1}+\frac{2d-k^*}{4d-1}+\frac{4d-1-2d}{4d-1}=$$

$$=\frac{1}{4d-1}\left[4d-1-2d+2d-k^*-k^*\right]=\frac{1}{4d-1}\left[4d-1-2k^*\right]\ge\frac{1}{4d-1}=\gamma$$

והמעבר האחרון נובע מכך ש- $k^*<2d$ לכן $4d-1-2k^*\ge 1$. קיבלנו שבשני המקרים - $\hat{x}$ בתוך המקטע או $\hat{x}$ מחוצה לו, מתקיים הא"ש הנדרש.

Programming Assignment

1) There training and test errors per iteration are:



2) Here are the top 10 classifiers chosen by the algorithm:
(Direction = 1 iff word implies negative review)

Iteration: #0, Dimension: 26, Theta: 1.5, Direction: 1, Word: "bad"
Iteration: #1, Dimension: 195, Theta: 1.5, Direction: -1, Word: "performances"
Iteration: #2, Dimension: 372, Theta: 1.5, Direction: 1, Word: "boring"
Iteration: #3, Dimension: 37, Theta: 2.5, Direction: -1, Word: "great"
Iteration: #4, Dimension: 88, Theta: 1.5, Direction: 1, Word: "script"
Iteration: #5, Dimension: 22, Theta: 1.5, Direction: -1, Word: "life"
Iteration: #6, Dimension: 31, Theta: 1.5, Direction: -1, Word: "many"
Iteration: #7, Dimension: 311, Theta: 1.5, Direction: 1, Word: "worst"
Iteration: #8, Dimension: 120, Theta: 1.5, Direction: -1, Word: "family"
Iteration: #9, Dimension: 76, Theta: 1.5, Direction: 1, Word: "nothing"

Three classifiers I expected are in blue (very straightforward words indicating good/bad movies).
Three less intuitive classifiers are in red, they do not directly imply a negative/positive emotion.

3) The average exponential loss per errors per iteration is: