Theory Questions

1. **(10 points) Suboptimality of ID3.** Solve exercise 2 in chapter 18 in the course book: Understanding Machine Learning: From Theory to Algorithms.

2. **(Suboptimality of ID3)**
Consider the following training set, where $\mathcal{X} = \{0,1\}^3$ and $\mathcal{Y} = \{0,1\}$:

$$1.\ ((1,1,1),1)$$
$$2.\ ((1,0,0),1)$$
$$3.\ ((1,1,0),0)$$
$$4.\ ((0,0,1),0)$$

Suppose we wish to use this training set in order to build a decision tree of depth 2 (i.e., for each input we are allowed to ask two questions of the form $(x_i = 0?)$ before deciding on the label).

1. Suppose we run the ID3 algorithm up to depth 2 (namely, we pick the root node and its children according to the algorithm, but instead of keeping on with the recursion, we stop and pick leaves according to the majority label in each subtree). Assume that the subroutine used to measure the quality of each feature is based on the entropy function (so we measure the *information gain*), and that if two features get the same score, one of them is picked arbitrarily. Show that the training error of the resulting decision tree is at least 1/4.

כאשר נחפש את הפרדיקט שממקסם את ה gain כדי לבחור שורש לעץ:

$$Gain(S,i) = \underbrace{C(\mathbb{P}_S[Y=1])}_{\text{pre-split error}} - \underbrace{(\mathbb{P}_S[X_i=1]\, C(\mathbb{P}_S[Y=1|X_i=1]) + \mathbb{P}_S[X_i=0] C(\mathbb{P}_S[Y=1|X_i=0]))}_{\text{post-split error}}$$

$$\text{Entropy, } C(\alpha) = -\tfrac{1}{2}\left(\alpha \log \alpha + (1-\alpha)\log(1-\alpha)\right)$$

$$gain(S,1) = C\left(\tfrac{1}{2}\right) - \left[\tfrac{3}{4}\cdot C(2/3) + \tfrac{1}{4}\cdot C(0)\right] \quad (\tfrac{1}{4}\cdot C(0) \to 0)$$

$$= -\tfrac{1}{2}\cdot\underbrace{\left[\tfrac{1}{2}\cdot\log(1/2) + \tfrac{1}{2}\cdot\log(1/2)\right]}_{-1} - \tfrac{3}{4}\cdot\left(-\tfrac{1}{2}\right)\cdot\underbrace{\left[\tfrac{2}{3}\log(2/3) + \tfrac{1}{3}\cdot\log(1/3)\right]}_{-0.78} \approx 0.21$$

$$= -\frac{1}{2} \cdot \underbrace{\left[\frac{1}{2} \cdot \log(1/2) + \frac{1}{2} \cdot \log(1/2)\right]}_{-1} - \frac{3}{4} \cdot \left(-\frac{1}{2}\right) \cdot \underbrace{\left[\frac{2}{3} \log(2/3) + \frac{1}{3} \cdot \log(1/3)\right]}_{-0.78} \approx 0.21$$

$$gain(S,2) = \underbrace{C(1/2)}_{1} - \left[\frac{1}{2} \cdot \underbrace{C(1/2)}_{1} + \frac{1}{2} \cdot \underbrace{C(1/2)}_{1}\right] = 0$$

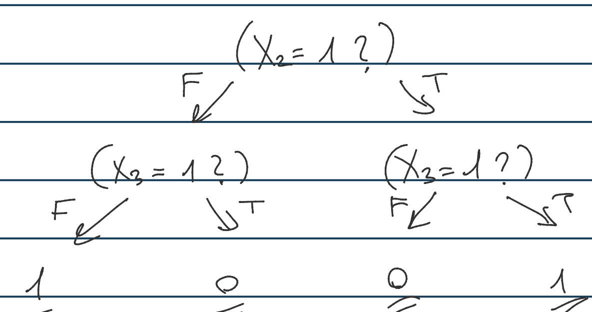
$$gain(S,3) = C(1/2) - \left[\frac{1}{2} \cdot C(1/2) + \frac{1}{2} \cdot C(1/2)\right] = 0$$

לכן הפרדיקט $(X_1=1?)$ ייבחר להיות השורש.
כתוצאה מכך, 3 הדוגמאות הראשונות ילכו לתת העץ הימני, והדוגמה האחרונה לתת העץ השמאלי.

כעת נשים לב, שלא משנה איזה פרדיקט נבחר לפיצול את העץ הימני, נבצע שגיאות חיזוי.
קיימות 2 אפשרויות לפיצול תת העץ:
1) אם נפצל לפי $(X_2=1?)$:
הדוגמאות הראשונה והשלישית ילכו לתת העץ הימני ומכיוון ש $y_1=1 \neq y_3=0$ נטעה בחיזוי של אחת מהן.
2) אם נפצל לפי $(X_3=1?)$:
הדוגמאות השנייה והשלישית ילכו לתת העץ השמאלי ומכיוון ש $y_2=1 \neq y_3=0$ נטעה בחיזוי של אחת מהן.

סה"כ קיבלנו שכך או כך בכל עץ שיתקבל נטעה בחיזוי של לפחות דוגמה אחת, ומכיוון שיש 4 דוגמאות נקבל שהשגיאה המינימלית היא לפחות $1/4$. ■

2. Find a decision tree of depth 2 that attains zero training error.



3. **(20 points) AdaBoost.** Let $x_1, \ldots, x_m \in \mathbb{R}^d$ and $y_1, \ldots, y_m \in \{-1, 1\}$ its labels. We run the AdaBoost algorithm as given in the lecture, and we are in iteration t. Assume that $\epsilon_t > 0$.

(a) Show that the error of the current hypothesis relative to the new distribution is exactly 1/2, that is:

$$\Pr_{x \sim D_{t+1}}[h_t(x) \neq y] = \frac{1}{2}.$$

נוכיח שמתקיים: $\sum_{i:h_t(x_i)\neq y_i} D_{t+1}(i) = \frac{1}{2}$ שכן מתקיים:

$$\Pr_{X\sim D_{t+1}}[h_t(X)\neq Y] = \mathbb{E}_{D_{t+1}}\left[\mathbb{I}\{h_t(X)\neq Y\}\right] = \sum_{i=1}^{m} D_{t+1}(i)\cdot \mathbb{I}\{h_t(x_i)\neq y_i\} = \sum_{i:h_t(x_i)\neq y_i} D_{t+1}(i)$$

[לפי הגדרת D_{t+1}] [לפי הגדרת Z_t ולפי $y_i h_t(x_i) = -1$ מתקיים $h_t(x_i)\neq y_i$]

נשים לב שמתקיים כי:

$$\sum_{i:h_t(x_i)\neq y_i} D_{t+1}(i) \overset{\downarrow}{=} \sum_{i:h_t(x_i)\neq y_i} \frac{D_t(i)\cdot \exp(-y_i w_t h_t(x_i))}{Z_t} = \sum_{i:h_t(x_i)\neq y_i} \frac{D_t(i)\cdot \exp(w_t)}{Z_t}$$

$$= \frac{e^{w_t}}{Z_t}\cdot \overbrace{\sum_{i:h_t(x_i)\neq y_i} D_t(i)}^{=\epsilon_t} = \frac{e^{w_t}\cdot \epsilon_t}{Z_t} = \frac{\left(\frac{1-\epsilon_t}{\epsilon_t}\right)^{0.5}\cdot \epsilon_t}{2\cdot\sqrt{\epsilon_t(1-\epsilon_t)}} = \frac{(1-\epsilon_t)^{0.5}\cdot \epsilon_t^{0.5}}{2\cdot \epsilon_t^{0.5}\cdot(1-\epsilon_t)^{0.5}} = \frac{1}{2}$$

[לפי הגדרת w_t ולפי הגדרת Z_t מהמשוואה]

■

(b) Show that AdaBoost will not pick the same hypothesis twice consecutively; that is $h_{t+1} \neq h_t$.

נניח בשלילה ש AdaBoost בוחר את אותה ההשערה פעמיים ברציפות, כלומר $h_{t+1} = h_t$.
כאמור שמתקיים $\epsilon_t < \frac{1}{2}$ ו $\epsilon_{t+1} < \frac{1}{2}$ כי מדובר בהשערות שמתקבלות ע"י לומד חלש.
מסעיף הקודם ראינו שהשגיאה של h_t מול ההתפלגות D_{t+1} היא $\frac{1}{2}$.
מכיוון שההשערות h_{t+1}, h_t זהות, גם השגיאה של h_{t+1} מול D_{t+1} היא $\frac{1}{2}$. סתירה. ■

4. **(20 points) Sufficient Condition for Weak Learnability.** Let $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a training set and let $\mathcal{H}$ be a hypothesis class. Assume that there exists $\gamma > 0$, hypotheses $h_1, \ldots, h_k \in \mathcal{H}$ and coefficients $a_1, \ldots, a_k \geq 0$, $\sum_{i=1}^{k} a_i = 1$ for which the following holds:

$$y_i \sum_{j=1}^{k} a_j h_j(x_i) \geq \gamma \tag{1}$$

for all $(x_i, y_i) \in S$.

(a) Show that for any distribution D over S there exists $1 \leq j \leq k$ such that

$$\Pr_{i \sim D}[h_j(x_i) \neq y_i] \leq \frac{1}{2} - \frac{\gamma}{2}.$$

(Hint: Take expectation of both sides of inequality (1) with respect to D.)
Remark: Note that the condition above is sufficient for *empirical* weak learnability, the condition defined in lecture #9 for the Adaboost analysis.

$\exists 1 \leq j \leq k$ מוכיח: $\Pr_{i \sim D}[h_j(x_i) \neq y_i] > \frac{1-\gamma}{2}$

נקח תוחלת על שני צדי אי השוויון (1):

$$\gamma = \mathbb{E}[\gamma] \leq \mathbb{E}\left[y_i \cdot \sum_{j=1}^{k} a_j h_j(x_i)\right] = \sum_{i=1}^{n} \Pr_{X,Y \sim D}[X = x_i, Y = y_i] \cdot y_i \sum_{j=1}^{k} a_j h_j(x_i)$$

$$\overset{D \text{ התפלגות}}{=} \sum_{i=1}^{n} D(i)\, y_i \sum_{j=1}^{k} a_j h_j(x_i) = \sum_{j=1}^{k} a_j \sum_{i=1}^{n} D(i)\, y_i h_j(x_i) = *$$

שוב שוויון: $h_j(x_i) = y_i$ מוכיח $y_i \cdot h_j(x_i) = 1$ ואחרת $h_j(x_i) \neq y_i$ מוכיח $y_i \cdot h_j(x_i) = -1$. לכן:

$$* = \sum_{j=1}^{k} a_j \left[\sum_{i : h_j(x_i) = y_i} D(i) - \sum_{i : h_j(x_i) \neq y_i} D(i)\right] = \overset{\text{משלים ל-1}}{\text{מהתפלגות}}$$

$$= \sum_{j=1}^{k} a_j \cdot \left(\Pr_{i \sim D}[h_j(x_i) = y_i] - \Pr_{i \sim D}[h_j(x_i) \neq y_i]\right) =$$

$$= \sum_{j=1}^{k} a_j \Pr_{i \sim D}[h_j(x_i) = y_i] - \sum_{j=1}^{k} a_j \Pr_{i \sim D}[h_j(x_i) \neq y_i] =$$

$$= \sum_{j=1}^{k} a_j \left(1 - \Pr_{i \sim D}[h_j(x_i) \neq y_i]\right) - \sum_{j=1}^{k} a_j \Pr_{i \sim D}[h_j(x_i) \neq y_i] =$$

$$= \sum_{j=1}^{k} a_j - 2 \sum_{j=1}^{k} a_j \cdot \Pr_{i \sim D}[h_j(x_i) \neq y_i] = 1 - 2 \sum_{j=1}^{k} a_j \cdot \Pr_{i \sim D}[h_j(x_i) \neq y_i]$$

$$\gamma \leq 1 - 2 \sum_{j=1}^{k} a_j \Pr_{i \sim D}[h_j(x_i) \neq y_i]$$ אז כך קיבלנו:

$$\iff \sum_{j=1}^{k} a_j \Pr_{i \sim D}[h_j(x_i) \neq y_i] \leq \frac{1-\gamma}{2}$$

$$\frac{1-\gamma}{2} = \sum_{i=1}^{k} a_j \cdot \left(\frac{1-\gamma}{2}\right) < \sum_{i=1}^{k} a_j \Pr_{i \sim D}[h_j(x_i) \neq y_i]$$ אבל מהנחה שלנו

אבל מההנחה בשלב $\frac{1-\gamma}{2} = \sum_{j=1}^{k} a_j \cdot \left(\frac{1-\gamma}{2}\right) < \sum_{j=1}^{k} a_j Pr_{i\sim D}[h_j(x_i)\neq y_i]$

כלומר קיבלנו $\frac{1-\gamma}{2} < \sum_{j=1}^{k} a_j \cdot Pr_{i\sim D}[h_j(x_i)\neq y_i] \le \frac{1-\gamma}{2}$. סתירה.

■

(b) Let $S = \{(x_1,y_1),\ldots,(x_n,y_n)\} \subseteq \mathbb{R}^d \times \{-1,1\}$ be a training set that is realized by a d-dimensional hyper-rectangle classifier, i.e., there exists a d dimensional hyper-rectangle $[b_1,c_1]\times\cdots\times[b_d,c_d]$ which contains all of the positive points in S and doesn't contain the negative points in S. Let $\mathcal{H}$ be the class of decision stumps of the form

$$h(x) = \begin{cases} 1 & x_j \le \theta \\ -1 & x_j > \theta \end{cases}, \quad h(x) = \begin{cases} 1 & x_j \ge \theta \\ -1 & x_j < \theta \end{cases},$$

for $1 \le j \le d$ and $\theta \in \mathbb{R}\cup\{\infty,-\infty\}$ (for $\theta\in\{\infty,-\infty\}$ we get constant hypotheses which predict always 1 or always -1). Show that there exist $\gamma>0$, $k>0$, hypotheses $h_1,\ldots,h_k\in\mathcal{H}$ and $a_1,\ldots,a_k\ge 0$ with $\sum_{i=1}^{k} a_i = 1$, such that the condition in inequality (1) holds for the training set S and hypothesis class $\mathcal{H}$. This implies that $\mathcal{H}$ is empirically weak learnable w.r.t. data realizable by a d-dimensional hyper-rectangle.
(Hint: Set $k=4d-1$, $a_i=\frac{1}{4d-1}$ and let $2d-1$ of the hypotheses be constant.)

ניקח: $k=4d-1$, $\forall i\in[k]: a_i = \frac{1}{4d-1}$

כעת נגדיר את $k(=4d-1)$ ההשערות במחלקה:

$$h_{b_j}(x) = \begin{cases} 1 & x_j \ge b_j \\ -1 & \text{else} \end{cases}, \quad h_{c_j}(x) = \begin{cases} 1 & x_j \le c_j \\ -1 & \text{else} \end{cases}$$

d היפו' ($j\in[d]$)   d היפו'

ונגדיר גם $2d-1$ היפותזות קבועות שבפשטות יחזירו לייבל -1 לכל אינפוט שיקבלו.

תהי דוגמה (x,y) כלשהי.
נחלק למקרים לפי הלייבל y:
• אם $y=1$:
אז הנקודה בתוך ההיפר מלבן, כלומר $\forall j\in[d]: b_j\le x_j\le c_j$.
מתקיים:

$$y\cdot\sum_{j=1}^{k} a_j h_j(x) = \frac{1}{4d-1}\cdot\sum_{j=1}^{k} h_j(x) = \frac{1}{4d-1}\cdot\left(2d-(2d-1)\right) = \frac{1}{4d-1}$$

[$2d$ ההיפו' שאינן קבועות יחזירו 1 כי הנקודה בתוך ההיפר מלבן.]   [ההיפו' $2d-1$ הקבועות יחזירו -1]

(נקודה בתוך ההיפר מלבן.)

• אם $y=-1$:

אז הנקודה לא בתוך ההיפר מלבן, כלומר $\exists j\in[d]: x_j<b_j \vee x_j>c_j$

מתקיים:

$$y\cdot\sum_{j=1}^{k} a_j h_j(x) = -\frac{1}{4d-1}\cdot\sum_{j=1}^{k} h_j(x) = -\frac{1}{4d-1}\left(-(2d-1)+\sum_{j=1}^{d} h_{b_j}(x)+\sum_{j=1}^{d} h_{c_j}(x)\right)$$

[קבוע 2d-1 (ההיפו' הקבועה)]

$$= \frac{2d-1}{4d-1} - \frac{1}{4d-1}\cdot\left(\overbrace{\sum_{j=1}^{d} h_{b_j}(x)+\sum_{j=1}^{d} h_{c_j}(x)}^{*}\right)$$

כעת נבחין כי בהינתן שהנקודה לא בתוך ההיפר מלבן, לפחות אחת מ $2d$ ההיפותזות: $h_{b_1},\dots,h_{b_d},h_{c_1},\dots,h_{c_d}$ תחזיר -1. לכן הסכום * יהיה מקסימלי כאשר יתר $2d-1$ ההיפו' יחזירו $+1$.

כלומר, $* \le 2d-1-1 = 2d-2$

חזרה לחישוב:

$$\frac{2d-1}{4d-1}-\frac{1}{4d-1}\left(\overbrace{\sum_{j=1}^{d} h_{b_j}(x)+\sum_{j=1}^{d} h_{c_j}(x)}^{\le 2d-2}\right) \ge \frac{2d-1}{4d-1}-\frac{1}{4d-1}\cdot(2d-2) = \frac{1}{4d-1}$$

לסיכום, קיבלנו במקרה ש $y=1$: $y\cdot\sum_{j=1}^{k} a_j h_j(x) = \frac{1}{4d-1}$

ובמקרה ש $y=-1$: $y\sum_{j=1}^{k} a_j h_j(x) \ge \frac{1}{4d-1}$

לכן אם נבחר $\gamma = \frac{1}{4d-1}$ נקבל את הדרוש. ■

2. **(20 points) Properties of KL divergence**. Recall the definition of the KL-divergence from slide 15 in recitation 8.

   (a) Show that the KL-divergence is always non-negative.
   (Hint: Consider the convex function $f(y) = y \log y$, and use Jensen's inequality which states that for any distribution q, $\mathbb{E}_{z\sim q}[f(z)] \geq f(\mathbb{E}_{z\sim q}[z])$).

   (b) Let p_1, p_2, q_1, q_2 be distributions over $\mathcal{X}$ such that p_1 is independent of p_2 and q_1 is independent of q_2. Denote the product distributions $p = p_1 \times p_2$ and $q = q_1 \times q_2$ over $\mathcal{X}^2$ (i.e. $p(x_1, x_2) = p_1(x_1)p_2(x_2)$ for any $x_1, x_2 \in \mathcal{X}$ and similarly for q). Prove the following:

$$D_{KL}(p, q) = D_{KL}(p_1, q_1) + D_{KL}(p_2, q_2).$$

(a) נגדיר $f(z) = -\log(z)$ קמורה שכן $\log$ הוא פונ'

$$D_{KL}(p,q) := \sum_{x\in X} p(x)\cdot \log\left(\frac{p(x)}{q(x)}\right) = \sum_{x\in X} p(x)\cdot\left[-\log\left(\frac{q(x)}{p(x)}\right)\right]$$

$$= \mathbb{E}_p\left[-\log\left(\frac{q(x)}{p(x)}\right)\right] \overset{\text{jensen}}{\geq} -\log\left(\mathbb{E}_p\left[\frac{q(x)}{p(x)}\right]\right) = -\log \sum_{x\in X} p(x)\cdot\frac{q(x)}{p(x)}$$

$\left[z = \frac{q(x)}{p(x)}\right]$

$$= -\log \sum_{x\in X} q(x) = -\log 1 = 0$$

$\sum_{x\in X} q(x) = 1$ סכום כל ההסתברויות

■

(b)

$$D_{KL}(p,q) = \sum_{x_1\in X}\sum_{x_2\in X} p(x_1,x_2)\cdot\log\left(\frac{p(x_1,x_2)}{q(x_1,x_2)}\right) = \sum_{x_1\in X}\sum_{x_2\in X} p_1(x_1)\cdot p_2(x_2)\cdot\log\left(\frac{p_1(x_1)\cdot p_2(x_2)}{q_1(x_1)\cdot q_2(x_2)}\right)$$

$$= \sum_{x_1\in X}\sum_{x_2\in X} p_1(x_1)\cdot p_2(x_2)\cdot\left[\log\left(\frac{p_1(x_1)}{q_1(x_1)}\right) + \log\left(\frac{p_2(x_2)}{q_2(x_2)}\right)\right]$$

$$= \sum_{x_1\in X}\sum_{x_2\in X} p_1(x_1)\cdot p_2(x_2)\cdot\log\left(\frac{p_1(x_1)}{q_1(x_1)}\right) + \sum_{x_1\in X}\sum_{x_2\in X} p_1(x_1)\cdot p_2(x_2)\cdot\log\left(\frac{p_2(x_2)}{q_2(x_2)}\right)$$

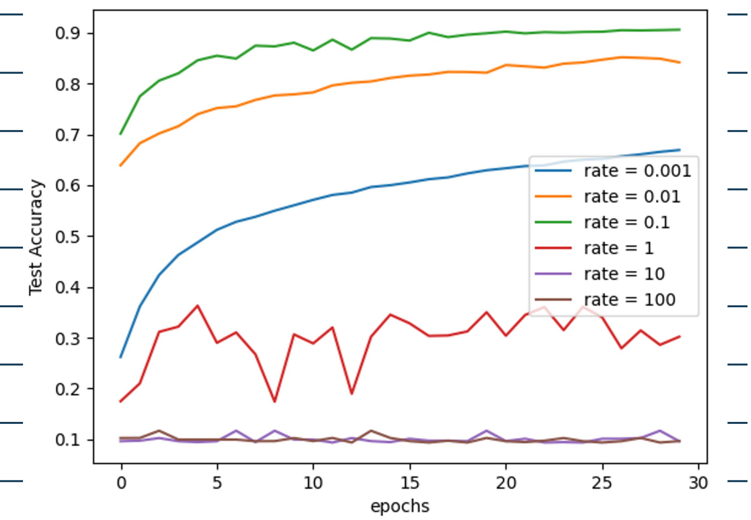
$$= \sum_{x_1\in X} p_1(x_1)\cdot\log\left(\frac{p_1(x_1)}{q_1(x_1)}\right)\cdot\underbrace{\sum_{x_2\in X} p_2(x_2)}_{\text{סכום כל ההסתברויות שווה 1}} + \sum_{x_2\in X} p_2(x_2)\cdot\log\left(\frac{p_2(x_2)}{q_2(x_2)}\right)\cdot\underbrace{\sum_{x_1\in X} p_1(x_1)}_{=1}$$

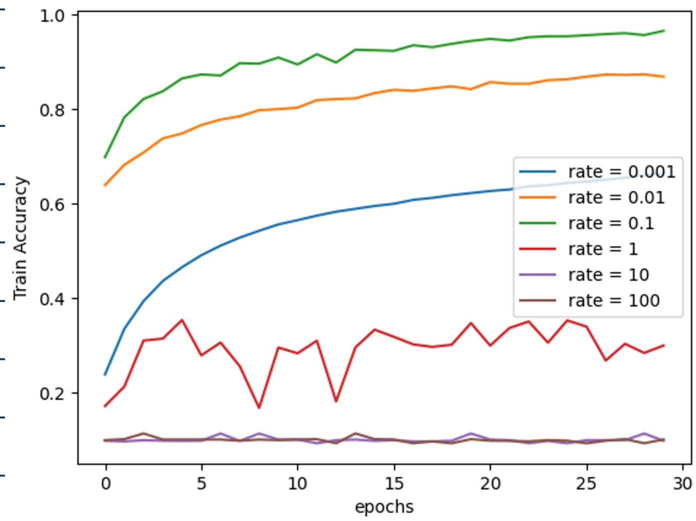
$$= \sum_{x_1\in X} p_1(x_1)\cdot\log\left(\frac{p_1(x_1)}{q_1(x_1)}\right) + \sum_{x_2\in X} p_2(x_2)\cdot\log\left(\frac{p_2(x_2)}{q_2(x_2)}\right) = D_{KL}(p_1,q_1) + D_{KL}(p_2,q_2)$$

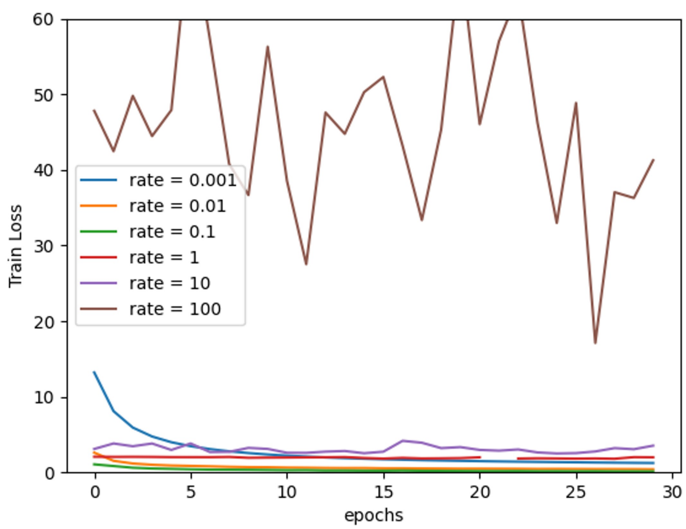
■

שאלה תכנותית

סעיף ב'








```
finished SGD
Epoch 29: test accuracy: 0.906
```

ניתוח הגרפים:

ניתן לראות שהדיוק הטוב ביותר מתקבל עבור קצב למידה 0.1 .
עבור קצבי למידה גבוהים יותר, התשובה משיגה דיוק מאוד גרוע (פחות מ 0.4), זאת משום שצעד ה SGD גדול מדי וככל הנראה האלגוריתם מפספס את נקודת המינימום. עקב עובדה זו ניתן לקבל מהדיוק עבור קצב למידה = 1 כאשר הדיוק מתנודד בטווח (0.2, 0.4) ללא מגמה ברורה. עובדה נוספת יכולה להיות ערך פונ' ההפסד עבור קצב למידה = 100 כאשר [illegible] ירידה בערך פונ' ההפסד.

קצב למידה = 1 כושר הדיוק מתנודד בטווח (0.2 ,0.4) ללא מגמה ברורה. עבור נוסף יכולה להיות ערך פונ' ההפסד
עבור קצב למידה = 100 כושר רואים בבירור שאין ירידה בערך פונ' ההפסד.
עבור קצבי למידה נמוכים מ 0.1, המצב יותר טוב מבחינת הדיוק לעומת קצבי למידה גבוהים מ 0.1 אך עדיין לא מושג דיוק של 0.9. ככל הנראה הסיבה לכך היא שעבור קצבי למידה נמוכים מדי, הצעדים ב SGD קטנים מדי ולא מתקרבים מספיק לנקודת המינימום.

סעיף ג'

```
Epoch 27 test accuracy: 0.9463
Epoch 28 test accuracy: 0.9413
Epoch 29 test accuracy: 0.9433
Finished SGD - test accuracy: 0.9433
```