

למידת מכונה – דו"ח מטלת בית 3

Neural Networks

מגיש: לירון חיים

ת.ז: 206234635

תיאור המודל הסופי

הרשת בנויה משכבה נסתרת אחת בין שכבת הקלט לשכבת הפלט. כאשר בין הקלט לשכבה הנסתרת פועלת פונקציית אקטיבציה Sigmoid, ובין השכבה הנסתרת פועלת פונקציית האקטיבציה SoftMax (עבור התאמה ל-Multi Class).

גודל השכבות:

- שכבת הקלט – 784 ניורונים, כמספר הפיצ'רים, מספר הפיקסלים לתמונה.
- שכבה נסתרת – 100 ניורונים.
- שכבת פלט – 10 ניורונים, כמספר התיוגים האפשריים, 10 ספרות.

תהליך האימון

אתחול הפרמטרים (מטריצות המשקולות, וקטורי bias) נעשה באופן אקראי עם ערכים בין 0 ל-1 (ניסיתי גם לאתחל עם ערכים שונים סביב ה-0 לפי פונקציות אתחול מוכרות). לאחר מכן, ריצה על סט האימון בלולאה של מספר אפוקים רצוי. בכל אפוק נערבב את סדר התמונות והתיוגים שלהן בהתאם כדי שהרשת לא תלמד את הסדר. כעת בלולאה שרצה על כל התמונות, נבצע F-Prop, לאחר מכן Back-Prop ונעדכן את הפרמטרים (לוגיקת SGD). לאחר שנסיים לעבור את כל האפוקים, נחזיר את הפרמטרים והם יהוו את המודל המאומן.

אציין שלפני שליחת הקלט, נרמלתי אותו ע"י חלוקה בגודל פיקסל (255) כדי לעבוד ברשת רק עם ערכים בסדרי גודל דומים בין השכבות. באופן דומה, נרמלתי גם את הערכים במעבר בין השכבות הבאות (לפי הכנסה לפונקציית אקטיבציה) ע"י חלוקה בגודל השכבה.

פונקציית ה-Loss שלי היא $Loss = -\log V_t^L = -\log h_2$. ולשם הנוחות והביצועים, גזרתי והשתמשתי בכלל השרשת:

$$\begin{aligned} \text{W1:} \quad \frac{dL}{dz_1} &= \frac{dL}{dz_2} \frac{dz_2}{dh_1} \frac{dh_1}{dz_1} = ((h_2 - y) \cdot W_2) * (sig(z_1) * (1 - sig(z_1))) \\ \frac{dL}{dW1} &= \frac{dL}{dz_1} \frac{dz_1}{dW1} = \frac{dL}{dz_1} \cdot x \\ \text{W2:} \quad \frac{dL}{dz_2} &= \frac{d(-\log(h_2))}{dz_2} = (h_2 - y) \\ \frac{dL}{dW2} &= (h_2 - y) \cdot h_1^T \\ \text{B2:} \quad \frac{dL}{dB2} &= (h_2 - y) = \frac{dL}{dz_2} \\ \text{B1:} \quad \frac{dL}{dB1} &= \frac{dL}{dz_1} \frac{dz_1}{dB1} = \frac{dL}{dz_1} \cdot 1 = \frac{dL}{dz_1} \end{aligned}$$

לקוח ממצגת ההרצאה

תהליך הבדיקה

נרוץ על סט הבדיקה שלנו בלולאה, כאשר בכל איטרציה נבצע F-Prop עם תמונה ונקבל פרדיקציה לתיוג (האינדקס של הערך המקסימלי בשכבת הפלט, שהרי מדובר בערכים הסתברותיים של SoftMax). כל תיוג נרשום בשורה לקובץ test_y.

תהליך כיול היפר-פרמטרים

ההיפר-פרמטרים שיש לבדוק ולכייל הם: מספר השכבות הנסתרות, גודל כל שכבה נסתרת, אפוקים וקצב הלמידה. על מנת לבדוק את כל אלו, חילקתי את סט האימון שקיבלנו לסט אימון חדש וסט וולידציה לפי K-Fold. אימנתי K (20) מודלים ומדדתי כל השגיאה הממוצעת על K הבדיקות המתאימות להם. בנוסף, כתבתי קוד קצר שרושם לקובץ ומתעד את ההיפר-פרמטרים והתוצאות בכל ריצה של התוכנית. בגלל אילוצי כוח עיבוד, בדקתי את הפרמטרים בצורה ידנית ובחרתי באלו הנותנים תוצאה שמספקת אותי. מבחינת פונקציות אקטיבציה, ניסיתי את tanh, sigmoid, relu, בצירופים שונים ולבד.

נראה כי עם שכבה נסתרת אחת התוצאות המתקבלות טובות ולכן בחרתי לקחת שכבה אחת.

```
PARAMETERS:
NUM_LAYERS = 3
ACTIVATION_FUNCTIONS = ['EMPTY' 'sigmoid' 'softmax']
LAYER_SIZES = [784 100 10]
EPOCHS = 10
RATE = 0.37
SEEDS = 5312, [4216 3293 8650 3716 5181 8304 3191 7086 9497 1526]
!! TRAIN ACC(%): 98.7981818181818 !!
```

לבסוף בחרתי בהיפר-פרמטרים הבאים:

(ה-SEEDS שמצינים כאן הוגרלו רנדומלית, ונועדו לקבע לי את המודל שיווצר בהרצה על שרתי האוניברסיטה.)