מבוא ללמידה עמוקה - תרגיל 1

עידן רפאלי ואנאל בן סימון

30 בנובמבר 2020

חלק מעשי

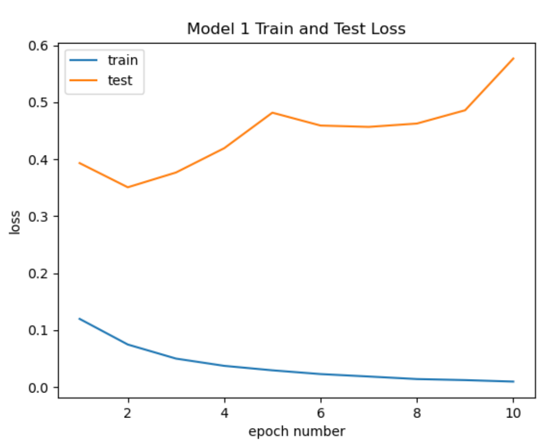
1. בסעיף 3 מתואר אופן ייצוג הדאטא שבחרנו בו.

2. ניסינו להריץ את האימון על 6 ארכיטקטורות שונות של רשתות, אשר נבדלות ביניהן במספר השכבות, מספר הנוירונים בכל שכבה, וסוגי האקטיבציה (Relu/Sigmoid). 4 רשתות הן בעלות מספר נוירונים גבוה יחסית, ו-2 רשתות בעלות מספר נמוך. הקלט בכל הארכטקטורות הוא וקטור בגודל 180 (כפי שתואר בסעיף 3), והפלט הוא נוירון בודד, שמציין את התיוג שהרשת מביאה עבור הדגימה - מספר בין 0 ל-1 (לאחר הפעלת Softmax). ניתן לראות את ארכיטקטורות הרשתות בקוד שצירפנו (שמות המחלקות הן $\text{Model1}, \ldots, \text{Model6}$). בסופו של דבר, הארכיטקטורה שהביאה לאחוזי הדיוק הטובים ביותר על הטסט היא הארכיטקטורה של מודל 1 (המחלקה Model1), בעלת אחוזי דיוק של כ-92.6%. האכיטקטורה "המנצחת" היא:
   - 3 שכבות נסתרות
   - שכבה נסתרת ראשונה: 512 נוירונים
   - שכבה נסתרת שניה: 512 נוירונים
   - שכבה נסתרת שלישית: 256 נוירונים
   - כל האקטיבציות הן Relu

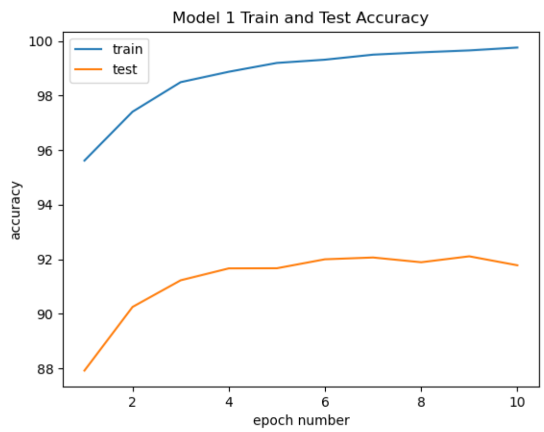
3. כל דוגמה בדאטא ייצגנו בעזרת וקטור בינארי באורך 180: ישנם 20 סוגים שונים של חומצות אמינו, וישנו רצף של 9 חומצות בכל דגימה. לכל סוג חומצה הגדרנו אינדקס (מספר בין 0 ל-19). כל 20 ערכים רצופים בוקטור מתאימים לחומצת אמינו בדגימה, כאשר ערכי כולם אפסים מלבד באינדקס המתאים לחומצת האמינו, שם הערך הוא 1. סה"כ ישנם $20 * 9 = 180$ ערכים שונים בוקטור.
   יש לציין כי זיהינו חוסר איזון משמעותי בין מספר הדגימות התויגיות בתיוג חיובי לעומת תיוג שלילי - כ-89% מהדגימות תויגו באופן שלילי (תיוג 0), ורק כ-11% תויגו באופן חיובי (תיוג 1). כדי לנסות להתגבר על חוסר האיזון, החלטנו לשכפל את הדגימות החיוביות **בשלב האימון בלבד** כך שמספרן סה"כ יהיה קרוב יחסית למספר הדגימות השליליות.

4. כפי שציינו בסעיף 2, מודל 1 נבחר בתור המודל הטוב ביותר. הפרמטרים שלו הן מטריצות בגודל 180×512, 512×512, 512×256 וכן 256×1, עם $biases$ בגדלים 512, 512, 256, 1 בהתאמה. בסעיף 5 מתוארים הגרפים של ה-Loss וה-Accuracy עבור דאטא האימון והטסט (הגרפים של כל המודלים מצורפים בקובץ הזיפ).

5. להלן הגרפים המתארים את ה-Loss וה-Accuracy על דאטא האימון והטסט, עבור מודל 1:



אפשר לראות בגרף הנ"ל כי ה־loss יורד כל הזמן, כצפוי על דאטא האימון, אך על הטסט ה־loss במגמת עליה לאורך האפוקים (עקב ה־overfitting)



אפשר לראות בגרף הנ"ל שה־accuracy עולה בשני הגרפים בהתחלה מהר. בדאטא האימון, הדיוק עולה עד לאחוזים גבוהים מאוד (קרוב ל־100%) וזה צפוי כמובן, אך על הטסט הדיוק מתייצב באזור ה־92% ולא עולה עוד, בגלל ה־overfitting.

6. את הדגימות מהדאטא של Spike SARS-CoV-2 יצרנו על־ידי הפרדת כל 9 אותיות (חומצות אמינו) לדגימה נפרדת, כך שהאות הראשונה של כל דגימה חופפת לאות האחרונה של הדגימה שלפניה. ביצענו פרדיקציה עם מודל 1 המאומן עבור דגימות הקורונה. קיבלנו שהדגימות עבורן המודל נתן להן את הציון הגבוה ביותר הן:

$$VIRGDEVRQ$$
$$FPREGVFVS$$
$$KCVNFNFNG$$
$$CLIGAEHVN$$
$$KTQSLLIVN$$

שאלות תאורטיות

1. יהיו $f:\mathbb{R}^n\to\mathbb{R}^m$, $g:\mathbb{R}^k\to\mathbb{R}^n$ פונקציות ליניאריות המוגדרות על־ידי $f(x)=Ax$ ו־$g(x)=Bx$ עבור $A\in\mathbb{R}^{m\times n}$, $B\in\mathbb{R}^{n\times k}$. מתקיים:

$$f\left(g\left(x\right)\right)=f\left(Bx\right)=ABx$$

וזו פונקציה ליניארית, כאשר $AB\in\mathbb{R}^{k\times m}$.

יהיו $f:\mathbb{R}^n\to\mathbb{R}^m$, $g:\mathbb{R}^k\to\mathbb{R}^n$ פונקציות אפיניות המוגדרות על־ידי $f(x)=Ax+b$ ו־$g(x)=Cx+d$ עבור $A\in\mathbb{R}^{m\times n}$, $B\in\mathbb{R}^{n\times k}$, $b\in\mathbb{R}^m$, $d\in\mathbb{R}^n$. מתקיים:

$$f\left(g\left(x\right)\right)=f\left(Cx+d\right)=A(Cx+d)+b=ACx+Ad+b$$

וזו פונקציה אפינית, כאשר $AC\in\mathbb{R}^{k\times m}$ וכן $Ad+b\in\mathbb{R}^m$.

2.

(א) תנאי העצירה עבור התהליך האיטרטיבי הוא:

$$\left|\nabla f\left(x^n\right)\right|\leq\epsilon$$

התנאי הזה נובע מהעובדה שכאשר x^n הוא נקודת קיצון, הגרדיאנט בנקודה זו מתאפס, אך משיקולים נומריים של המחשב, חייבים להרשות איזשהו tolerance על־ידי הדרישה ש־$\left|\nabla f\left(x^n\right)\right|\leq\epsilon$ עבור $\epsilon>0$ קטן כרצוננו.

(ב) ראשית, נקודה x תהיה נקודת מינימום או מקסימום לוקאלית של הפונקציה f רק כאשר הגרדיינט של הפונקציה באותה נקודה הוא 0, כלומר $\nabla f(x)=0$.

נסתכל קירוב טיילור מסדר שני בנקודה x:

$$f(x+dx)=f(x)+\nabla f(x)\cdot dx+dx^T\cdot H(x)\cdot dx+O\left(\|dx\|^3\right)$$

כמו שאמרנו, $\nabla f(x)=0$. כמו כן, לאחר שנעביר את $f(x)$ לאגף שמאל, נקבל:

$$f(x+dx)-f(x)=dx^T\cdot H(x)\cdot dx+O\left(\|dx\|^3\right)$$

כדי ש־x תהיה נקודת מינימום לוקאלי, נרצה שלכל dx יתקיים $f(x+dx)-f(x)\geq 0$, כלומר שלכל dx יתקיים ש־$dx^T\cdot H(x)\cdot dx\geq 0$ וזה יקרה אמ"ם מטריצת ההסייאן $H(x)$ היא מטריצה PSD וזה קורה אמ"ם לכל ערך עצמי λ של $H(x)$ מתקיים $\lambda\geq 0$ (נזכור ש־H סימטרית כי אנחנו מניחים ש־f גזירה פעמיים, כלומר H לכסינה אורתוגונלית). באופן דומה dx היא נקודת מקסימום לוקאלי אמ"ם $f(x+dx)-f(x)\leq 0$ אמ"ם לכל ערך עצמי λ של $H(x)$ מתקיים $\lambda\leq 0$.
אם קיימים ע"ע חיוביים ושליליים, זוהי נקודת אוכף.

3. נשתמש בפונקציית ההפסד הבאה:

$$\ell(y,y')=\sin\left(\frac{1}{2}(y-y')\right)^2$$

כאשר y מייצג את הלייבל האמיתי, ו־y' הפרדיקציה של הרשת (זוויות). הטווח של $\ell(y,y')$ הוא $[0,1]$. בחרנו בפונקצית ההפסד הנ"ל מכיוון שפונקציית הסינוס היא פונקציה מונוטונית עולה בתחום $[-90°,90°]$, וזה תואם לכך שככל שההפרש $y-y'$ קטן יותר, גם פונקציית ההפסד קטנה יותר, ולהפך. למשל, במקרי הקצה, כאשר עבור ההפרש הכי גדול האפשרי בין y ו־y' (בערך מוחלט, ועד כדי ציקליות) הוא $180°$, פונקציית ההפסד נותנת את הערך 1 ($\sin(90°)^2=1$) ועבור ההפרש הכי קטן שהוא $0°$ נקבל ש־$\ell(y,y')=\sin\left(\frac{1}{2}\cdot 0°\right)^2=0$.

קוד tensorflow שממש את פונקציית ההפסד הנ"ל:

```
from tensorflow.keras.losses import Loss

class AnglesLoss(Loss):
        def call(self, y_true, y_pred):
                return tf.square(tf.sin(0.5 * (y_true- y_pred)))
```

4. נזכור כי משפט Cybenko אומר כי אם σ פונקציה מונוטונית ורציפה עם $\sigma(-\infty)=0$ וכן $\sigma(\infty)=1$, אז משפחת הפונקציות:

$$f(x)=\sum_i \alpha_i\sigma(w_ix+b_i)$$

היא צפופה ב־$C([0,1])$ ביחס לנורמת הסופרימום: $d(f,g)=\sup|f(x)-g(x)|$. נזכור כי משפט Hornik מרחיב את משפט Cybenko לכל פונקציה σ כזו שהיא חסומה.

הפונקציה Relu אמנם אינה חסומה, אך נוכל להגדיר פונקציה אחרת, שהיא כן חסומה, ומתלכדת עם Relu בקטע $[-\infty,1]$:

$$\sigma(x)=\begin{cases}0 & x\leq 0\\ x & 0\leq x\leq 1\\ 1 & x\geq 1\end{cases}$$

נשים לב כי למעשה מתקיים $\sigma(x)=\mathrm{Relu}(x)-\mathrm{Relu}(x-1)$. כמו כן הפונקציה σ היא פונקציה מונוטונית ורציפה, וכן היא חסומה, ולכן ממשפט Hornik, משפחת הפונקציות $f(x)=\sum_i\alpha_i\sigma(w_ix+b_i)$ צפופה ב־$C([0,1])$. ומכיוון שזה נכון עבור σ שהגדרנו, זה נכון גם עבור Relu, שמתלכדת איתה בקטע $[0,1]$.

5. נכליל את הבניה של הרשת העמוקה שמביעה רשת רדודה ב־$O(N)$ נוירונים, שראינו בכיתה, גם מבלי להניח ש־$\alpha_i>0$ בפונקציה המתארת את הרשת הרדודה:

$$f(x)=\sum_{i=1}^N\alpha_i\sigma(w_ix+b_i)$$

ברשת העמודה יהיו הפעם 6 נוירונים בכל שכבה (ולא 3 כמו ברשת שראינו בכיתה), כך שבסהכ יהיו $6N = O(N)$ נוירונים סה"כ. בכל שכבה נוסיף נוירונים (שנסמנם h_4, h_5) ששווים בערכם ל־$h_4 = \sigma(h_2)$ וכן $h_5 = \sigma(-h_2)$, כלומר הנוירון h_4 יהיה בעל קשת מ־h_2 עם משקולת 1, והנוירון h_5 יהיה בעל קשת מ־h_2 עם משקולת -1. נשנה את הרשת שהוצגה בכיתה כך שנחבר כעת את הנוירון h_4 (ולא h_2) לנוירון h_1 (עם משקולת 1). מה שנקבל הוא שהנוירון h_1 סוכם בצורה הדרגתית את כל הגורמים החיוביים בסכימה (אלו עבורם $\alpha_i \geq 0$). בנוסף נוסיף נוירון (שנסמנו h_6) שתפקידו יהיה דומה לזה של h_1 - לסכום את כל האיברים השליליים בסכום. זאת נעשה על־ידי חיבור הנוירון h_5 ו־h_6 מהשכבה הקודמת לנוירון h_6 של השכבה הבאה, עם משקולות 1. לבסוף, בשכבה האחרונה של הרשת, הפלט של הרשת יהיה נוירון בודד שערכו $h_1 - h_6$ (ללא אקטיבציית Relu).

נשים לב שאם לדוגמה $\alpha_i \geq 0$, בשכבה ה־i של הרשת נקבל ש־$h_2 = \alpha_i \sigma(w_i(h_3 - L) + b_i) \geq 0$, ולכן יתקיים $h_4 = \sigma(h_2) = h_2 = \alpha_i \sigma(w_i(h_3 - L) + b_i)$ ובשכבה שלאחר מכן יתקיים $h_1 = \sigma(h_1 + h_4) = \sigma(h_1 + \alpha_i \sigma(w_i(h_3 - L) + b_i))$. בנוסף, $h_5 = \sigma(-h_2) = 0$ ולכן $h_6 = \sigma(h_6 + h_5) = \sigma(h_6) = h_6$ (כלומר במצב זה h_6 לא ישתנה). באופן דומה, אם $\alpha_i \leq 0$, אז h_6 יסכום לתוכו את האיבר החדש שב־h_1 (לאחר שהפכנו אותו לחיובי דרך h_5), וערכו של h_1 לא ישתנה. בסופו של דבר h_1 יכיל את סכום כל האיברים החיוביים בסכימה, ו־h_6 יכיל את סכום כל האיברים השליליים בסכימה (או יותר נכון הערך המוחלט של הסכום), ולכן נוירון הפלט שערכו $h_1 - h_6$ יכיל בדיוק את הסכום המלא שמתאר את הפונקציה $f(x)$.

להלן ציור המתאר את הארכיטקטורה החדשה של הרשת בשכבה ה־i, כפי שתיארנו לעיל: