



הפצה לאחור של השגיאה ברשת נוירונים

חישוב הגרדיאנט של פונקצית השגיאה ברקורסיה, משכבת
הפלט אחורה

טיוטה

ד"ר ארז שיינר

מבוסס על המאמר:

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536. DOI: 10.1038/323533a0.

הסיפור הבסיסי

יש לנו רשת של n שכבות, ולכל קלט פלט רצוי. נרצה לבצע צעד בכיוון הירידה התלולה ביותר על פונקצית השגיאה על מנת לתקן את כל המשקלים. לשם כך צריך את הגרדיאנט של פונקצית השגיאה ביחס לכל מטריצת משקלים, ומיד נתאר כיצד מחשבים אותו ברקורסיה, שכבה אחר שכבה מהפלט אחורה; לכן השיטה נקראת הפצת השגיאה לאחור.

1 הרשת

הגדרה 1 השכבות

נניח שיש לנו n שכבות, וכל שכבה מכילה וקטור כניסה ווקטור יציאה בגדלים זהים. נסמן את גודל הוקטור של השכבה ה- k ב- i_k ; את וקטור הכניסה נסמן $x^k \in \mathbb{R}^{i_k}$, ואת הרכיב ה- i שבו נסמן x_i^k ; את וקטור היציאה נסמן $y^k \in \mathbb{R}^{i_k}$, ואת הרכיב ה- i שבו נסמן y_i^k .

הגדרה 2 פונקצית הסיגמואיד; הפלט של שכבה

לכל $k \geq 2$ ולכל $1 \leq i \leq i_k$ הפלט מוגדר להיות פונקציה של הקלט:

$$y_i^k = \sigma(x_i^k)$$

כאשר $\sigma(x) = \frac{1}{1 + e^{-x}}$ היא פונקצית הסיגמואיד.

טענה 3

לכל $x \in \mathbb{R}$ מתקיים

$$\sigma'(x) = \frac{e^{-x}}{(1 + e^{-x})^2} = \sigma(x) \cdot (1 - \sigma(x)).$$

הגדרה 4 המשקלים וההטיות; המעבר בין השכבות

לכל $1 \leq k \leq n - 1$ קיימת מטריצת משקלים

$$W^k \in \mathbb{R}^{i_{k+1} \times i_k}$$

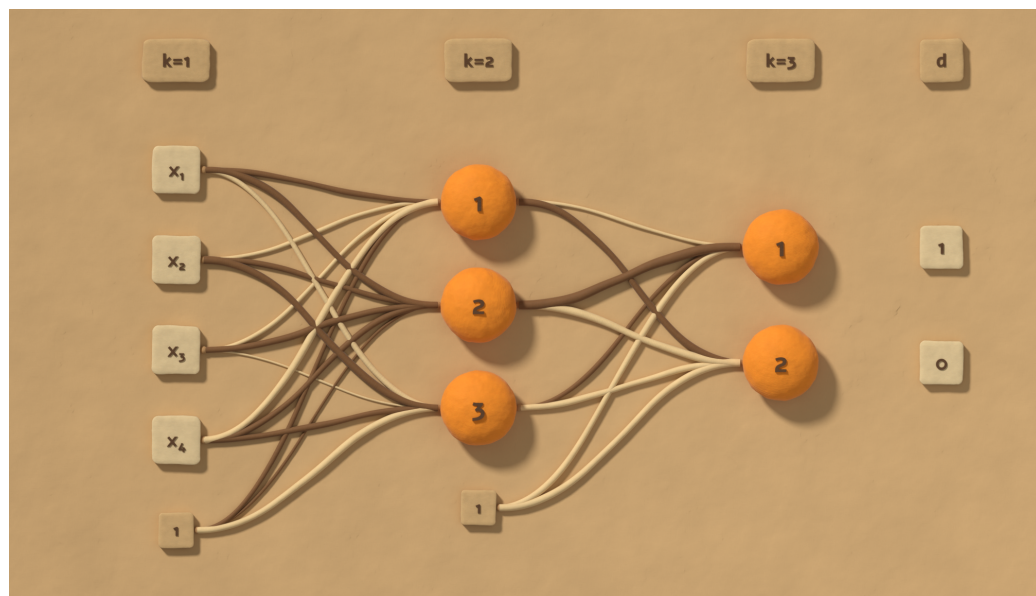
וקטור הטייה

$$b^k \in \mathbb{R}^{i_{k+1}},$$

והמעבר בין השכבות מתבצע ע"י החישוב

$$x^{k+1} = W^k y^k + b^k.$$

השכבה הראשונה יוצאת דופן: x^1 הוא וקטור הקלט של רשת הנוירונים כולה, והפלט שלו שווה לקלט, $y^1 = x^1$. הפלט של השכבה האחרונה, y^n , הוא הפלט של הרשת כולה.



רשת עם $n = 3$ שכבות, $i_3 = 2$, $i_2 = 3$, $i_1 = 4$. כל חבל הוא משקל w_{ij}^k , מהרכיב y_j^k אל הרכיב x_i^{k+1} . החבלים שיוצאים מהאריח 1 שבתחתית השכבה k הם ההטיות b_i^k : הטייה היא משקל של קלט שערכו תמיד 1. עובי החבל הוא $|w|$; חבל בהיר למשקל חיובי וכהה לשלילי. המשקלים נבחרו באקראי ועוגלו לספרה אחת: $b^2 = (0.7, 0.8)$, $W^2 = \begin{pmatrix} 0.3 & -1.2 & -0.5 \\ -0.8 & 0.9 & 0.7 \end{pmatrix}$, $b^1 = (-0.5, -0.2, 0.8)$, $W^1 = \begin{pmatrix} -0.6 & 0.4 & 0.4 & 0.8 \\ -0.8 & -0.6 & -0.8 & -0.7 \end{pmatrix}$.

2 השגיאה והעדכון

הגדרה 5 פונקציית השגיאה

לכל קלט יהיה לנו פלט רצוי $d \in \mathbb{R}^{i_n}$, ונגדיר את השגיאה עבור הקלט x^1 להיות

$$E(x^1, d, W^1, \dots, W^{n-1}, b^1, \dots, b^{n-1}) = \frac{1}{2} \|y^n - d\|^2 = \frac{1}{2} \sum_{p=1}^{i_n} (y_p^n - d_p)^2.$$

בהינתן x^1 ו- d קבועים, נרצה לבצע צעד בכיוון הירידה התלולה ביותר על פונקציית השגיאה על מנת לתקן את כל המשקלים. מיד נתאר כיצד מחשבים את הגרדיאנט של פונקציית השגיאה ביחס לכל מטריצת משקלים ברקורסיה; לכן השיטה נקראת הפצת השגיאה לאחור.

הגדרה 6 הגרדיאנטים

לכל $1 \leq k \leq n-1$ נגדיר את מטריצת הגרדיאנט $\nabla_{W^k} E \in \mathbb{R}^{i_{k+1} \times i_k}$ ע"י

$$(\nabla_{W^k} E)_{ij} = \frac{\partial E}{\partial w_{ij}^k}$$

ואת גרדיאנט ההטייה $\nabla_{b^k} E \in \mathbb{R}^{i_{k+1}}$ ע"י

$$(\nabla_{b^k} E)_i = \frac{\partial E}{\partial b_i^k}.$$

שיטה 7 העדכון

העדכון של מטריצות המשקלים יהיה

$$W_{\text{new}}^k = W_{\text{old}}^k - \eta \nabla_{W^k} E, \quad b_{\text{new}}^k = b_{\text{old}}^k - \eta \nabla_{b^k} E,$$

כאשר η הוא פרמטר המבטא את קצב השינוי.

3 הפצה לאחור

עבור $1 \leq k \leq n-1$ ובהנתן i, j כך ש- $1 \leq i \leq i_{k+1}$ ו- $1 \leq j \leq i_k$, אנחנו רוצים לחשב את $\frac{\partial E}{\partial w_{ij}^k}$ ואת $\frac{\partial E}{\partial b_i^k}$, כלומר את הנגזרת החלקית של פונקציית השגיאה לפי האיבר במקום ij במטריצת המשקלים W^k , או לפי הרכיב ה- i של וקטור ההטייה b^k . המשפט הבא נותן את כולן ברקורסיה.

משפט 8 הפצה לאחור של השגיאהיהי $1 \leq k \leq n-1$. אזי:

$$(א) \text{ לכל } 1 \leq i \leq i_{k+1} \text{ ולכל } 1 \leq j \leq i_k \text{ מתקיים } \frac{\partial E}{\partial w_{ij}^k} = \frac{\partial E}{\partial x_i^{k+1}} \cdot y_j^k \text{ וכן } \frac{\partial E}{\partial b_i^k} = \frac{\partial E}{\partial x_i^{k+1}}$$

$$(ב) \text{ לכל } 1 \leq i \leq i_{k+1} \text{ מתקיים } \frac{\partial E}{\partial x_i^{k+1}} = \frac{\partial E}{\partial y_i^{k+1}} \cdot y_i^{k+1} (1 - y_i^{k+1})$$

$$(ג) \text{ לכל } 1 \leq i \leq i_k \text{ מתקיים } \frac{\partial E}{\partial y_i^k} = \sum_{q=1}^{i_{k+1}} \frac{\partial E}{\partial x_q^{k+1}} \cdot w_{qi}^k$$

$$(ד) \text{ תנאי ההתחלה של הרקורסיה: לכל } 1 \leq i \leq i_n \text{ מתקיים } \frac{\partial E}{\partial y_i^n} = y_i^n - d_i$$

במילים: בהינתן כל הנגזרות של פונקציית השגיאה ביחס לפלט של שכבה, אפשר לחשב את הנגזרות ביחס לפלט של השכבה הקודמת לה. כאמור, זו בדיוק הפצת השגיאה לאחור.

מסקנה 9יהי $1 \leq k \leq n-1$. אזי לכל $1 \leq i \leq i_{k+1}$ ולכל $1 \leq j \leq i_k$ מתקיים

$$\frac{\partial E}{\partial w_{ij}^k} = \frac{\partial E}{\partial y_i^{k+1}} \cdot y_i^{k+1} (1 - y_i^{k+1}) \cdot y_j^k,$$

ולכל $1 \leq i \leq i_k$ מתקיים

$$\frac{\partial E}{\partial y_i^k} = \sum_{q=1}^{i_{k+1}} \frac{\partial E}{\partial y_q^{k+1}} \cdot y_q^{k+1} (1 - y_q^{k+1}) \cdot w_{qi}^k.$$

4 סיכום האלגוריתםלכל קלט x^1 עם פלט רצוי d :

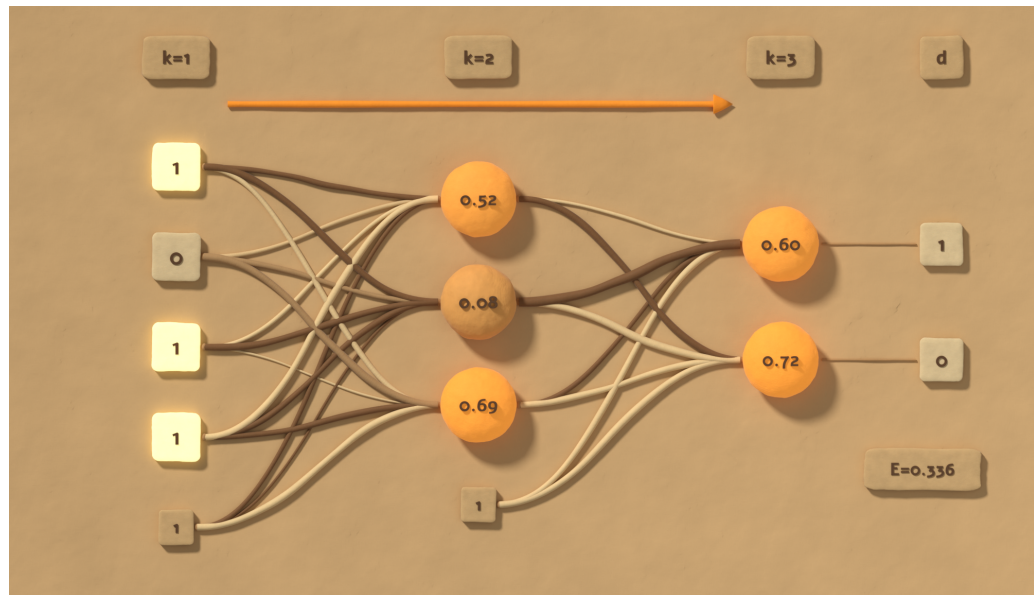
– **מעבר קדימה.** $y^1 = x^1$; לכל $k = 1, \dots, n-1$ מחשבים $x^{k+1} = W^k y^k + b^k$ ואת $y_i^{k+1} = \sigma(x_i^{k+1})$ (הגדרות 2 ו-4).

– **תנאי ההתחלה.** $\frac{\partial E}{\partial y_i^n} = y_i^n - d_i$ לכל i (משפט 8(ד)).

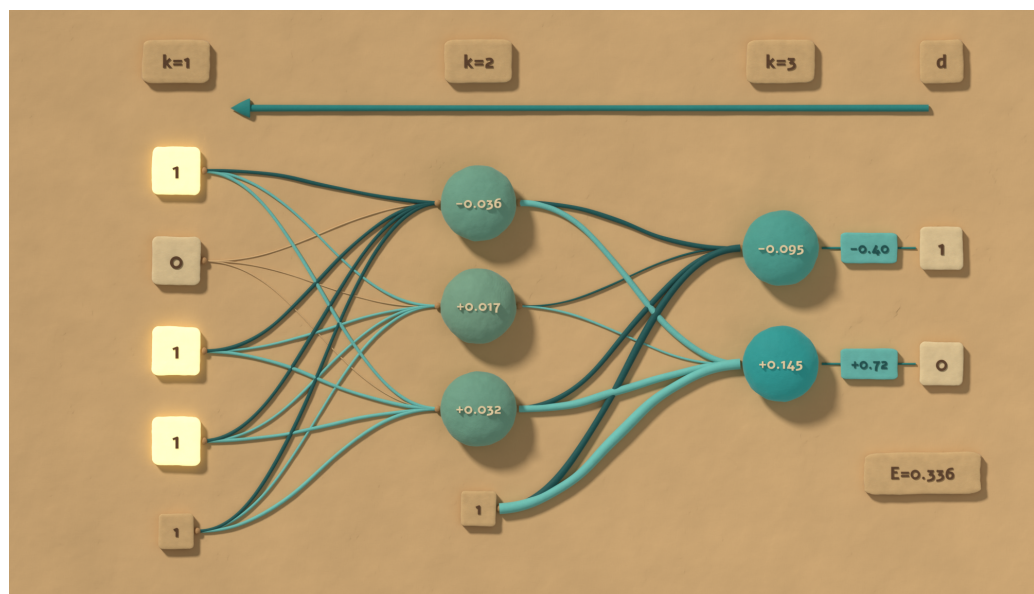
– **מעבר אחורה.** לכל $k = n-1, n-2, \dots, 1$: מחשבים $\frac{\partial E}{\partial x_i^{k+1}}$ לפי (ב), את $\nabla_{W^k} E$ ואת $\nabla_{b^k} E$ לפי (א), ואת $\frac{\partial E}{\partial y_i^k}$ לפי (ג), שמשמש בצעד הבא.

– **עדכון.** $W^k \leftarrow W^k - \eta \nabla_{W^k} E$ ו- $b^k \leftarrow b^k - \eta \nabla_{b^k} E$ (שיטה 7).

שני האיוורים הבאים מריצים את האלגוריתם על הרשת שבאיוור הראשון, עם הקלט $x^1 = (1, 0, 1, 1)$ והפלט הרצוי $d = (1, 0)$.



מעבר קדימה. בכל נורון כתוב הפלט שלו y_i^k , והוא זוהר לפיו: $y^2 = (0.52, 0.08, 0.69)$ ו- $y^3 = (0.60, 0.72)$. הרשת טועה בשני הרכיבים, והשגיאה היא $E = \frac{1}{2} \|y^3 - d\|^2 = 0.336$. החבליים שיוצאים מ- $y_2^1 = 0$ דהויים: הם אינם תורמים דבר ל- x^2 .



מעבר אחורה, בטורקיז. התגים שבין הפלט לפלט הרצוי הם תנאי ההתחלה $\partial E / \partial y_i^3 = y_i^3 - d_i$ (משפט 8(d)), כאן -0.40 ו- $+0.72$. בכל נורון כתוב $\partial E / \partial x_i^k$: בשכבה 3 לפי (ב), ובשכבה 2 לפי (ג) ואז (ב). למשל, לפי (ג), $\partial E / \partial y_1^2 = -0.095 \cdot 0.3 \approx -0.0285$, ולכן לפי (ב) $\partial E / \partial x_1^2 \approx -0.036$. עובי כל חבל הוא $|\partial E / \partial w_{ij}^k|$ לפי (א), בהיר לחיובי וכהה לשלילי. החבליים שיוצאים מ- $y_2^1 = 0$ דקים ודהויים, כי הגרדיאנט שלהם הוא אפס בדיוק: $\partial E / \partial w_{12}^1 = (\partial E / \partial x_1^2) \cdot y_2^1 = 0$. כל המספרים חושבו לפי המשפט ונבדקו מול הפרש סופי. $\partial E / \partial x_2^3 = 0.145$.

נספח: הוכחות

הוכחת טענה 3. לפי כלל השרשרת $\sigma'(x) = -\frac{e^{-x}}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})^2}$ וכן

$$\sigma(x)(1-\sigma(x)) = \frac{1}{1+e^{-x}} \cdot \frac{(1+e^{-x})-1}{1+e^{-x}} = \frac{e^{-x}}{(1+e^{-x})^2}.$$

הוכחת משפט 8. ובכן, היכן המשתנים b_i^k, w_{ij}^k מופיעים בנוסחת השגיאה? כיוון ש $x^{k+1} = W^k y^k + b^k$ אזי הרכיב i -י של וקטור הקלט של השכבה $k+1$ הוא

$$x_i^{k+1} = b_i^k + \sum_{p=1}^{i_k} w_{ip}^k y_p^k,$$

ולכן לפי כלל השרשרת

$$\frac{\partial E}{\partial w_{ij}^k} = \frac{\partial E}{\partial x_i^{k+1}} \cdot \frac{\partial x_i^{k+1}}{\partial w_{ij}^k} = \frac{\partial E}{\partial x_i^{k+1}} \cdot y_j^k, \quad \frac{\partial E}{\partial b_i^k} = \frac{\partial E}{\partial x_i^{k+1}} \cdot \frac{\partial x_i^{k+1}}{\partial b_i^k} = \frac{\partial E}{\partial x_i^{k+1}},$$

וזוהו (א).

בעת, כיוון ש $y_i^{k+1} = \sigma(x_i^{k+1})$, נובע לפי טענה 3 כי

$$\frac{\partial y_i^{k+1}}{\partial x_i^{k+1}} = \sigma'(x_i^{k+1}) = \sigma(x_i^{k+1})(1-\sigma(x_i^{k+1})) = y_i^{k+1}(1-y_i^{k+1}),$$

ולכן

$$\frac{\partial E}{\partial x_i^{k+1}} = \frac{\partial E}{\partial y_i^{k+1}} \cdot \frac{\partial y_i^{k+1}}{\partial x_i^{k+1}} = \frac{\partial E}{\partial y_i^{k+1}} \cdot y_i^{k+1}(1-y_i^{k+1}),$$

וזוהו (ב). עד כה גילינו כי

$$\frac{\partial E}{\partial w_{ij}^k} = \frac{\partial E}{\partial y_i^{k+1}} \cdot y_i^{k+1}(1-y_i^{k+1}) \cdot y_j^k.$$

בעת, על מנת ליצור נוסחת נסיגה, נשים לב כי לכל $1 \leq q \leq i_{k+1}$

$$x_q^{k+1} = b_q^k + \sum_{p=1}^{i_k} w_{qp}^k y_p^k,$$

ולכן, לפי כלל השרשרת, לכל $1 \leq i \leq i_k$

$$\frac{\partial E}{\partial y_i^k} = \sum_{q=1}^{i_{k+1}} \frac{\partial E}{\partial x_q^{k+1}} \cdot \frac{\partial x_q^{k+1}}{\partial y_i^k} = \sum_{q=1}^{i_{k+1}} \frac{\partial E}{\partial x_q^{k+1}} \cdot w_{qi}^k,$$

וזוהו (ג); ולפי (ב),

$$\frac{\partial E}{\partial y_i^k} = \sum_{q=1}^{i_{k+1}} \frac{\partial E}{\partial y_q^{k+1}} \cdot y_q^{k+1}(1-y_q^{k+1}) \cdot w_{qi}^k.$$

כלומר, בהינתן כל הנגזרות של פונקציית השגיאה ביחס לפלט של שכבה, אפשר לחשב את הנגזרות ביחס לפלט של השכבה הקודמת לה. כאמור, זו בדיוק הפצת השגיאה לאחור.

נותר רק לחשב את תנאי ההתחלה של הרקורסיה, כלומר הנגזרות

$$\frac{\partial E}{\partial y_i^n} = \frac{\partial}{\partial y_i^n} \left(\frac{1}{2} \|y^n - d\|^2 \right) = \frac{\partial}{\partial y_i^n} \left(\frac{1}{2} \sum_{p=1}^{i_n} (y_p^n - d_p)^2 \right) = y_i^n - d_i,$$

וזוהו (ד).

הוכחת מסקנה 9. הנוסחא הראשונה היא הצבת (ב) ב־(א), והשנייה היא הצבת (ב) ב־(ג) של משפט 8, כפי שנעשה בהוכחתו.