מבוא ללמידה עמוקה תרגיל 1

עמית קינן 2088296426 אלדן חודורוב 201335965

3 בנובמבר 2021

1 בהרכבה של פונקציות

1.1 הרכבה של פונקציות ליניארית

הראו כי הרכבה של פונקציות ליניאריות הינה לינארית, נסתכל במקרה הוקטורי כאשר פונקציה לינארית מוגדרת בתור

$$f(x) = Ax \wedge g(y) = By$$
$$x \in \mathbb{R}^n, y \in \mathbb{R}^m A \in \mathbb{R}^{m \times n} B \in \mathbb{R}^{k \times m}$$
$$g(f(x)) = g(Ax) = B(Ax) = (BA)x$$

על כן אם נסמן את $BA = C \in \mathbb{R}^{k \times n}$ נקבל כי $g(f(x)) \in \mathbb{R}^{k \times n}$ הינה לינארית

1.2 הרכבה של פונקציות אפיניות הינה אפינית

באופן דיי ישיר מהסעיף הקודם נגדיר את $a \in \mathbb{R}^m$ וכן את $b \in \mathbb{R}^k$ ונגדיר מחדש את $f \wedge b$ באופן הבא

$$f(x) = Ax + a, g(y) = By + b$$

ועל ידי הרכבה נקבל

$$g(f(x)) = g(Ax + a) = B(Ax + a) + b = (\underbrace{BA}_{=C})x + \underbrace{\underbrace{B \cdot a}_{\in \mathbb{R}^k}}_{=c} + b$$
$$= Cx + c$$

אזי פונקציה אפינית גם כן

2 calculus behind the Gradient Descent method

2.1 מהו תנאי העצירה לסכימה האינטרטיבית $\theta^{n+1} = \theta^n - \alpha \nabla f_{\theta^n}(x)$

באופן דיי ישיר תנאי העצירה הינו ש $\nabla f_{\theta^n}(x) = 0$ ואז נקבל $\theta^{n+1} = \theta^n$

2.2 הראו עם התור טיילור מסדר שני של הפונקציה מה התנאי הנדרש להסיק את הסיווג של נקודה סטציונרית כמינימום\קסימום

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

$$H_{i,j}(x) = \frac{d^2 f}{dx_i dx_j}(x)$$

נראה כי התנאי הנדרש לסיווג הנקודה כמינימום\מקסימום לוקאלים הוא ש $H(x)$ היא פונקציה חיובית או שלילית בהחלט בהתאמה וכן כי $\nabla f(x) = 0$
כלומר נניח כעת כי־ $H(x)$ חיובית בהחלט והנגזרת מתאפסת ו־ נקבל כי

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

כעת נרצה להראות ש $f(x)$ היא נקודת מקסימום לוקאלי כלומר לכל dx קטן מספיק מתקיים:

$$f(x + dx) \geq f(x)$$

מההנחה כי $H(x) > 0$נקבל כי

$$dx^T \cdot H(x) \cdot dx \geq 0$$

על כן נרצה להראות כי

$$dx^T \cdot H(x) \cdot dx \geq O(||dx||^3)$$

וכן ידוע כי עבור dx קטן מספיק

$$O(||dx||^3) \leq O(||dx||^2) \approx dx^T \cdot H(x) \cdot dx$$

על כן

$$dx^T \cdot H(x) \cdot dx + O(||dx||^3) \geq 0$$

ומכך נסיק

$$f(x + dx) = f(x) + \underbrace{\nabla f(x)}_{=0} \cdot dx + \underbrace{dx^T \cdot H(x) \cdot dx + O(||dx||^3)}_{\geq 0} \geq f(x)$$

כלומר $f(x)$ הינה נקודת מינימום לוקאלי, ובאופן סימטרי ניתן להוכיח את הדבר על נקודת מקסימום לוקאלי.

3 תיצור פונקציית Loss של רשת שמנבאת זווית בין 0־360 ושיודעת לקחת בחשבון מרחק זוויתי ־ כלומר המרחק בין 2 ל־360 הינו 2 ולא 358

```
def degrees_loss(y_pred,y_true):
        abs_diff = abs(y_true-y_pred)
        return min(abs_diff, 360-abs_diff)
```

ניתן לראות כי המערכת מחשבת את המרחק בין הפרדיקציה לGT ומחזירה מרחק הזוויתי בין התוצאה לפרדיקציה.
אציין כי יש עוד אפשרויות, למשל העברה לרדיאנים וחישוב המרחק קוסינוס מבין התוצאות, משהו קצת יותר דומה לcosine similarity

4 גזירת פונקציות עם כלל השרשרת

4.1 $\frac{d}{dx}f(x+y,2x,z)$

בשביל לגזור את הביטוי נסמן את התתי פונקציות של f באופן הבא

$$\begin{aligned} g(x) &= x+y \\ h(x) &= 2x \\ s(x) &= z \end{aligned}$$

על כן אם נפעיל את כלל השרשרת נקבל

$$\begin{aligned} \frac{df(x+y,2x,z)}{dx} &= \frac{df(g(x),h(x),s(x))}{dx} = \\ & \frac{dg(x)}{dx}\cdot\frac{df(g(x),h(x),s(x))}{dg(x)} + \frac{dh(x)}{dx}\cdot\frac{df(g(x),h(x),s(x))}{dh(x)} + \frac{ds(x)}{dx}\cdot\frac{df(g(x),h(x),s(x))}{ds(x)} \\ & = 1\cdot\frac{df(g(x),h(x),s(x))}{dg(x)} + 2\cdot\frac{df(g(x),h(x),s(x))}{dh(x)} + 0\cdot\frac{df(g(x),h(x),s(x))}{ds(x)} \\ & = 1\cdot\frac{df(g(x),h(x),s(x))}{dg(x)} + 2\cdot\frac{df(g(x),h(x),s(x))}{dh(x)} \end{aligned}$$

4.2 $f(x) = f_1(f_2(...f_n(x)))$

$$\frac{df}{dx} = \frac{df_1}{df_2}\cdot\frac{df_2}{df_3}\cdots\frac{df_n}{dx}$$

כלומר נקודת הגזירה של $\frac{df_i}{df_{1+1}}$תגזר לפי הנקודה $f_{i+1}(f_{i+2}(\cdots f_n(x))$ לכל iולמעט f_n שתגזר לפי הנקודה x

4.3 $f_1(x,f_2(x,f_3(...f_{n-1}(x,f_n)))$

בשביל לגזור את הביטוי הנ"ל נתחיל קודם להסתכל על ההנגזרת של $f_{n-1}(x,f_n(x))$

$$\frac{df_{n-1}(x,f_n(x))}{dx} = \frac{df_{n-1}(x)}{dx(1,0)} + \frac{df_{n-1}(f_n(x))}{df_n(x)(0,1)}\cdot\frac{df_n(x)}{dx}$$

אם נסתכל עוד צעד אחורה על $f_{n-2}(x,f_{n-1}(x,f_n)))$ נקבל כי

$$\frac{df_{n-2}(x,f_{n-1}(x,f_n)))}{dx} = \frac{df_{n-2}(x)}{dx(1,0)} + \frac{df_{n-2}(f_{n-1}(x))}{df_{n-1}(x)(0,1)}\cdot\frac{df_{n-1}(x,f_n(x))}{dx}$$

כאשר הביטוי האחרון היא הנגזרת של f_{n-1} שהראינו. כלומר אם נמשיך בצורה רקורסיבית נגיע חזרה ל f_1 כלומר

$$\frac{df_1(x,f_2(x,f_3(...f_{n-1}(x,f_n))}{dx} = \frac{df_1(x)}{dx(1,0)} + \frac{df_1(f_2(x,f_3\cdots))}{df_2(x)(0,1)}\cdot\frac{df_2(x,f_3(x,f_4(\cdots f_n(x))))}{dx}$$

4.4 $f(x+g(x+h(x)))$

נגדיר את

נגזור את הפונקציה לפי הנגזרות הפנימיות שלה ונקבל

$$\frac{df^*}{dx} = \frac{df(x+g(x+h(x)))}{d(x+g(x+h(x)))}\cdot(1+\frac{dg(x+h(x))}{d(x+h(x)})\cdot(1+\frac{dh(x)}{dx})$$

5 הוכח כי $D_{kl}(P||Q)$ היא לא שלילית כלומר $D_{kl}(P||Q) \geq 0$

להוכיח את הטענה נשתמש בכך שאנו יודעים כי $\sum_i^n p_i log(\frac{p_i}{q_i}) \geq (\sum_i p_i) \cdot log\frac{\sum p_i}{\sum q_i}$
אם כך, ניתן לראות כי:

$$D_{kl}(P||Q) = \sum_i^n p_i log(\frac{p_i}{q_i}) \geq (\sum_i p_i) \cdot log\frac{\sum p_i}{\sum q_i}$$
$$1 \cdot log\frac{1}{1} = 1 \cdot 0 = 0$$

כאשר $p = q$ יש שיוויון במעבר כלומר $D_{kl}(P||Q) = 0$

6 הוכח כי $D_{kl}(P||Q)$ היא קמורה

נדרש להוכיח כי

$$D_{kl}(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2) \leq \lambda D_{kl}(p_1||q_1) + (1-\lambda)D_{kl}(p_2||q_2)$$

כאשר

$$(p_1, p_2) \wedge (q_1, q_2)$$

הם זוגות של הסתברויות לא שליליות. וכן $\lambda \in [0, 1]$
נשתמש באותה טענה מהסעיף הקודם

$$\sum_i^n p_i log(\frac{p_i}{q_i}) \geq (\sum_i p_i) \cdot log\frac{\sum p_i}{\sum q_i}$$

אם כן,

$$\begin{aligned}
&D_{kl}(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2) \\
&= \sum (\lambda p_1 + (1-\lambda)p_2) \cdot log\left(\frac{\lambda p_1 + (1-\lambda)p_2}{\lambda q_1 + (1-\lambda)q_2}\right) \\
&\leq \sum \left(\lambda p_1 log\left(\frac{\lambda p_1}{\lambda q_1}\right) + (1-\lambda)p_2 log\left(\frac{(1-\lambda)p_2}{(1-\lambda)q_2}\right)\right) \\
&= \sum \lambda p_1 log\left(\frac{\lambda p_1}{\lambda q_1}\right) + \sum (1-\lambda)p_2 log\left(\frac{(1-\lambda)p_2}{(1-\lambda)q_2}\right) \\
&\lambda D_{kl}(p_1||q_1) + (1-\lambda)D_{kl}(p_2||q_2)
\end{aligned}$$

כנדרש

7 הראה כי הטענות של Cybenko and Hornik נתנות להחלה על פונקציה Relu

נראה כי ניתן לייצר פונקציה σ שהינה סכום של פונקציות $Relu$ עם הזזה וניפוח שינה רציפה ומונוטונית המקיימת כי $lim_{x\to\infty}\sigma(x) = 1$ וכן $lim_{x\to-\infty}\sigma(x) = 0$
כלומר הפונקציה תקיים את התנאים הנדרשים ועל כן חלים עליה הטענות.
נגדיר את הפונקציה

$$\sigma(x) = Relu(x + 0.5) - Relu(x - 0.5)$$

ניתן לראות כי הפונקציה מקיימת

$$\sigma(x)=\begin{cases}0 & x<-0.5\\ x+0.5 & x\in[-0.5,0.5]\\ 1 & x>0.5\end{cases}$$

על כן ניתן לבטא באמצעות סכום של פונקציות σ

$$f=\sum_i \alpha_i \sigma^{'}(\omega_i x+b_i)$$

כלומר היא הינה צפופה ב־ $C[0,1]$ תחת הנורמת סופרימום

8 הרחב את הבניה שהראינו בכיתה לביטוי רשת רדודה עם רשת עמוקה עם $O(n)$ נוירונים

8.1 תבטא רשת עם שכבה 1 על ידי רשת עמוקה עם $O(n)$ נוירונים לפונקציה ה־ $Sigend$

כלומר נקבל כי

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

כך ש α_i יכול להיות שלילי או חיובי. ולא רק חיובי כמו שהראינו בשיעור.
בשיעור השתמשנו בסימון של h_i לסמן את הזרמי חיבור של הניורונים ברשת. נמשיך את הסימון הזה.
אז במטרה להשיג את הרשת המתבקשת, נשנה את הזרוע של h_1 שעקבה וסכמה את הערכים, כך שכעת היא רק תסכום את הנויירונים עם $\alpha_i>0$
ונוסיף לרשת זרוע נוספת שתקרא h_4 והיא תסכום את כל הנורונים עם $\alpha_i<0$ כאשר $h_{3,}h_2$ ישארו ללא שינוי.
h_4 תפעל על ידי הכפלה ב 1־ של ה־ σ המתאימה ושל האינפוט מהשכבה הקודמת $h_4=-\sigma_i(-1\cdot h_2)$ ־ כאשר אינפוט בסימן הפוך יתאפסו.
נזכור שבסוף הרשת נכפול שוב נסכום את התוצאות שקיבלנו ב h_1 ו־h_4 ונקבל את הנדרש. כי בפועל עבורh_1 נקבל מ $\alpha\geq 0$ מ h_2 את הערך 0 בעוד שב h_4 נקבל את ההפך.
כלומר קיבלנו יכולת לבטא את הרשת השטוחה על ידי רשת עם n שכבות כאשר בכל שכבה יש $O(1)$ נוירונים ־ (בפועל עד 4 נוירונים).