פתרון מטלה 3 — חישוביות וקוגניציה, 6119

20 בנובמבר 2025

שאלת הכנה

סעיף א׳

תהי $f:\mathbb{R}^m\to\mathbb{R}$ אשר מקבל משתנה מקרי $\bar{x}$. נניח שפרספטרון לינארי מנסה ללמוד את f עם x באלגוריתם Batch יחד עם פונקציית שגיאה כלשהי. נסמן ב־$\bar{w}$ את תוצאת הלמידה אחרי n צעדי עדכון, ונבדוק מה משפיע על שגיאת ההכללה $\varepsilon_g(\bar{w})$.

פתרון נניח ש־$\varepsilon:\mathbb{R}^{m+1}\to\mathbb{R}_{\geq 0}$ פונקציית השגיאה, אז בהגדרה,

$$\varepsilon_g(\bar{w})=\mathbb{E}(\bar{w}\bar{x},f(\bar{x}))$$

ולכן ε_g תלוי ב־f (תשובה א׳), בפונקציית השגיאה שנבחרה (תשובה ב׳) התפלגות $\bar{x}$ (תשובה ג׳). נשים לב שידוע כי $\varepsilon_g \xrightarrow{n\to\infty} L$ ולכן יש תלות גם ב־$\bar{w}^0$ (תשובה ד׳) וב־n (תשובה ז׳).

נבחין כי כתלות במימד ובבחירת דוגמות אופטימלית נוכל גם להשפיע על השגיאה להיות מינימלית (על־ידי כופלי לגרנז׳ או גזירה) ולכן גם מספר הדוגמות יכול להשפיע וכן הדוגמות הספציפיות שנבחרו (כלומר תשובות ה׳ וו׳), אבל זוהי הנחה נוספת שאין לנו סיבה להניח בתנאי השאלה.

סעיף ב׳

נתונות שתי פונקציות $f,g:\mathbb{R}^3\to\mathbb{R}$ המוגדרות על־ידי,

$$f(x_1,x_2,x_3)=x_1^2+2x_2-x_3,\quad g(x_1,x_2,x_3)=3x_1^2+2x_2+1$$

נבדוק מה ניתן לומר על וקטורי הגרדיאנט של f,g ב־$x_0\in\mathbb{R}^3$ כך ש־$\forall i\leq 3,\ x_i\neq 0$.

פתרון מתקיים,

$$\nabla f(x)=(2x_1,2,-1),\quad \nabla g(x)=(6x_1,2,0)$$

ולכן מההנחה מתקיים $2x_1\neq 6x_1$ וכן $-1=0$, לכן הם שונים באיבר השלישי והראשון (תשובה ב׳).

אם נניח רק ש־$x\neq 0$ אז הם שונים ברכיב אחד במקרה שבו $x_1=0$.

סעיף ג׳

תהי רשת נוירונים לינאריים מארכיטקטורה לא ידועה שמנסה ללמוד פונקציה לא לינארית בעזרת אלגוריתם online, נבדוק מה נכון במקרה זה.

פתרון אם קצב הלימוד η גדול ביחס לגרדיאנט ולמרחק מהמינימום המקומי אז לא מובטחת התכנסות, כלומר הגעה לשגיאה אופטימלית תדרוש הרבה זמן (תשובה ב׳).

נניח שאנחנו מאמנים כמות נתונה של פעמים על דוגמות ספציפיות אז הקטנת η תייעל את תהליך האימון (תשובה ג׳).

שאלה 1

תהי $f:\mathbb{R}^4\to\mathbb{R}$ המוגדרת על־ידי $f(x)=x^t\cdot x+a^t\cdot x$ כאשר $a=(2,4,8,16)^t$.

סעיף א'

נסמן $x_0=1$ ונחשב את $f(x_0)$.

פתרון מתקיים $f(x_0)=f(1)=1\cdot 1+a\cdot 1=2+4+8+16=30$ כאשר $1\in\mathbb{R}^4$ וקטור סקלרי.

סעיף ב'

נחשב את הגרדיאנט של f וכן את ערכו ב־x_0.

פתרון מתקיים,

$$\nabla f(x)=D(x^tx+a^tx)=2x+a$$

בכתיבה וקטורית, קרי $\nabla f(x_1,x_2,x_3,x_4)=(2x_1+a_1,2x_2+a_2,2x_3+a_3,2x_4+a_4)$. בהתאם גם $\nabla f(1)=2\cdot 1+a=a+2=(4,6,10,18)^t$.

סעיף ג'

נבדוק את $f(x_0+\varepsilon)-f(x_0)$ עבור,

$$\varepsilon^1=(0.1,0.2,0.1,0.2)^t,\quad \varepsilon^2=(0.1,0.2,0.2,0.1)^t,\quad \varepsilon^3=(0.2,0.2,0.1,0.1)^t$$

פתרון נבחין כי $\|x_0+\varepsilon^i\|^2=5.3$ וכן $f(x_0)=30$ ולכן,

$$f(x_0+\varepsilon^1)-f(x_0)=5.3+a^t(x_0+\varepsilon^1)-30=5.3-a^t\varepsilon^1=0.3$$

ולכן באופן דומה נקבל שגם $f(x_0+\varepsilon^2)-f(x_0)=1.1$ וגם $f(x_0+\varepsilon^3)-f(x_0)=1.7$.

סעיף ד'

נחשב את הזווית בין ε^i ל־$\nabla f(x_0)$.

פתרון לכל $1\le i\le 3$,

$$\theta(\nabla f(x_0),\varepsilon^i)=\arccos\left(\frac{\nabla f(x_0)\cdot\varepsilon^i}{\|\nabla f(x_0)\|\cdot\|\varepsilon^i\|}\right)=\arccos\left(\frac{(4,6,10,18)\cdot\varepsilon^i}{6.89927532426\ldots}\right)$$

ולכן מהצבה במחשבון נקבל,

$$\theta(\nabla f(x_0),\varepsilon^1)=0.454125247397\ldots,\quad \theta(\nabla f(x_0),\varepsilon^2)=0.67181883632\ldots,\quad \theta(\nabla f(x_0),\varepsilon^3)=0.801367264691\ldots$$

סעיף ה'

נסביר את המגמה שהתקבלה בין הפרשי f מסעיף ג' לבין הזוויות מסעיף ד'.

פתרון קיבלנו שהזווית היא הקטנה ביותר כשההפרש הוא הקטן ביותר, זה לא מפתיע שכן חישבנו קירוב של נגזרת כיוונית ל־ε^1 בסעיף ב' ובסעיף ד' חישבנו קירוב כזה שוב (שכן זווית מוגדרת על־ידי נרמול).

סעיף ו'

נמצא ε כך ש־$f(x_0+\varepsilon)-f(x_0)$ יהיה מקסימלי, כאשר $\|\varepsilon\|=\|\varepsilon^1\|$.

פתרון מצאנו ש־$\nabla f(x_0) = (4,6,10,18)^t$, לכן כהעתקה לינארית נקבל את העלייה הגבוהה ביותר בווקטור הכיוון $\frac{\nabla f(x_0)}{\|\nabla f(x_0)\|}$, התבקשנו לחשב עבור ε עם נורמה $\|\varepsilon^1\|$, ולכן נקבל,

$$\varepsilon = \frac{\nabla f(x_0)}{\|\nabla f(x_0)\|} \cdot \|\varepsilon^1\| \approx (0.0579, 0.086, 0.144, 0.2608)^t$$

סעיף ז׳

נכתוב את הקשר שבין הסעיפים הקודמים לבין האלגוריתם של למידת גרדיאנט.

פתרון בלמידת גרדיאנט אנו למעשה מחשבים את הגרדיאנט במטרה למצוא וקטור כיוון מקסימלי במטרה לקחת את הווקטור המנוגד לו, הגודל $\|\varepsilon^1\|$ הוא למעשה η, קבוע הלמידה.

נעבור לחלק ב׳ של השאלה.

סעיף א׳

עבור $x, \varepsilon_0 \in \mathbb{R}^n$ נפתח את פולינום טיילור שלהם מסדר ראשון לערך $f(x+\varepsilon)$.

פתרון מהגדרה נקבל שפיתוח סביב x_0 הוא $P_1(x) = f(x_0) + Df|_{x_0}(x - x_0)$, לכן בפרט גם $P_1(x_0+\varepsilon) = f(x_0) + Df|_{x_0}\varepsilon$.

סעיף ב׳

נמצא קירוב ל־$f(x_0+\varepsilon^1)$ עבור $x_0 = 1\varepsilon^1 = (0.1, 0.2, 0.1, 0.2)^t$.

פתרון במקרה שלנו מתקיים,

$$P_1(x_0+\varepsilon) = 34 + (4,6,10,18) \cdot \varepsilon$$

ולכן $P(x_0+\varepsilon^1) = 34+6.2 = 40.2$. מהצד השני חישוב ישיר מניב $f(x_0+\varepsilon^1) = 40.3$. כלומר נקבל ש־$|f(x_0+\varepsilon) - P_1(x_0+\varepsilon)| = 0.1$.

סעיף ג׳

נגדיר $\varepsilon^* = 2\varepsilon^1$ ונבדוק את P_1 ו־f שוב.

פתרון הפעם $f(x_0+\varepsilon^*) = 46.8$ וכן $P_1(x_0+\varepsilon^*) = 46.4$, כלומר הפעם $\Delta = 0.4$, הקפיצה גדלה בקצב כפול, זה כמובן הגיוני שכן f פולינום מסדר 2.

סעיף ד׳

נסביר את הקשר שבין שני הסעיפים הקודמים לבין אלגוריתם למידת גרדיאנט.

פתרון באלגוריתם למידת גרדיאנט אנו מסתמכים על לינאריות הגרדיאנט בנקודה, כלומר אנו מבצעים סדרה של קירובים לינאריים לפונקציית הרשת, אנו רואים שבחלק הקודם המסקנה היא שהכיוונים המתקבלים בדרך זו הם יעילים, כלומר מצביעים לכיוון הלמידה האופטימלי, אבל הפעם אנו רואים שההזזה של המשקולות לא בהכרח תהיה כזו, ותלויה בעיקר ב־η.

שאלה 2

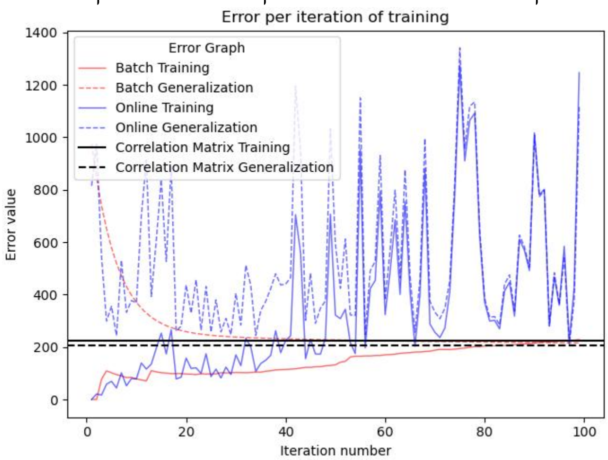
בשאלה זו נדון בלמידה מפוקחת באלגוריתמים שונים לפונקציה $y(x) = 1+x+x^2+x^3$ כאשר $x \sim U([-5,5])$. בכל סעיף נעסוק בפרספטרון לינארי עם סף, ולכן נניח בלי הגבלת הכלליות שהקלט יהיה מהצורה $x = (x,1)^t$.

סעיף א׳

נבנה מבחן ממוחשב שבו נוצרות 100 דוגמות וערכיהן וננסה לאמן פרספטרון לינארי בהתאם בקצב לימוד $\eta = 0.01$. את האימון נעשה בשיטת גרדיאנט חיי, בקבוצות, ולבסוף נשתמש גם במנגנון אימון עם מטריצת קורלציה.

סעיף ב׳

נריץ את האימון בהתאם לסוגי המבחנים ונבחן את שגיאת האימון וההכללה שלהם עבור כל שלב במהלך האימון עבור 100 שלבי אימון.



פתרון

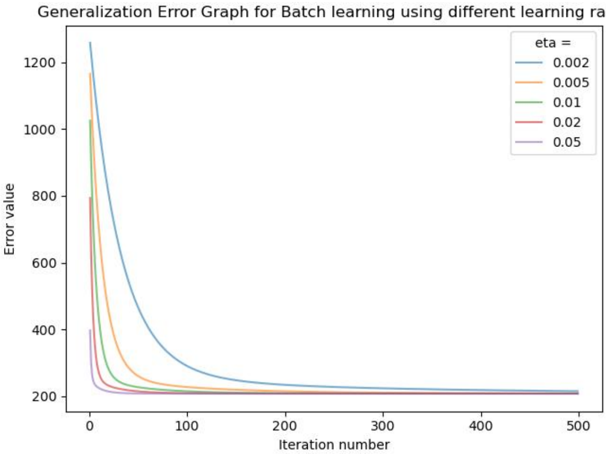
נבחין כי בזמן שהאימון לפי קורלציה הוא עבור כל הקלטים במהלך חישובי יחיד, ולכן מתקבל המינימום של הליך האימון באופן מיידי, הליך הלמידה בגרדיאנט מגוון יותר. עבור הלמידה בקבוצות נבחין כי עקומת הלמידה חלקה במובן המתמטי, אנחנו רואים פונקציה מונוטונית בשני סוגי האימון, הסיבה היא הריצה על כמות גדולה של דוגמות במקביל. לעומת זאת, באימון המיידי הריצה היא על כל קלט בנפרד, ולכן נוכל לראות קפיצות קשות יחסית עבור קלטים שההבדל ביניהם הוא מהותי, ובהתאם גם ההתכנסות היא פחות מהירה ופחות נראית לעין.

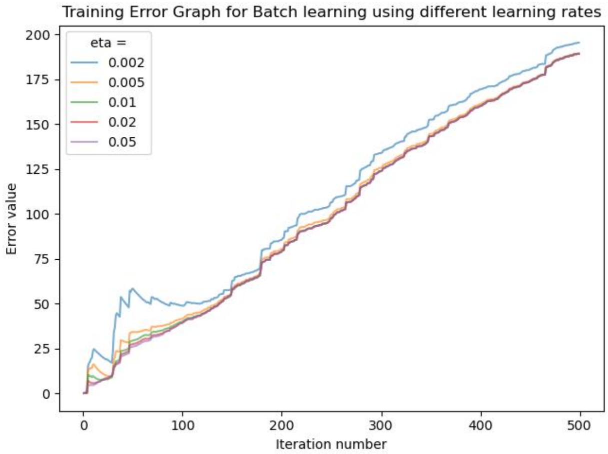
סעיף ג׳

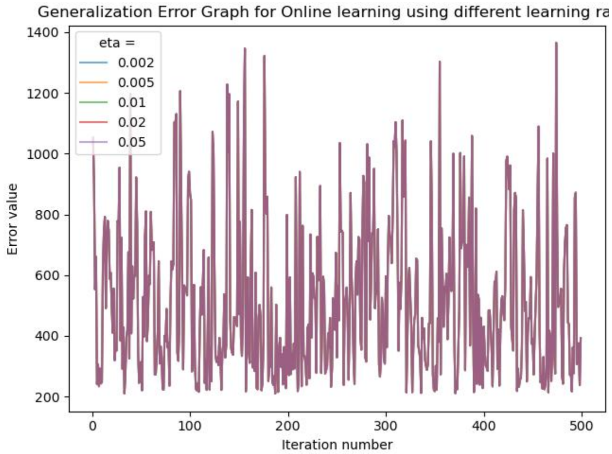
עתה ניצור 500 דוגמות כפי שתואר בתחילת השאלה, ונבדוק את הליך האימון עם קצבי הלימוד $\eta \in \{0.002, 0.005, 0.01, 0.02, 0.05\}$ ונבדוק את הליך הלמידה. נריץ את שני אלגוריתמי למידת הגרדיאנט עם 500 צעדי הרצה, עבור אלגוריתם מיידי נבדוק את שגיאת ההכללה ועבור קבוצות נחשב את שני סוגי השגיאה.

סעיף ד׳

פתרון נריץ את הניסוי הממוחשב ונציג שלושה גרפים של השגיאות כפי שתואר.







סעיף ה׳

ננתח את השפעת קצב הלימוד על השגיאות באלגוריתמים.

פתרון באלגוריתם הקבוצות אפשר לראות שקבוע קטן יותר מאיץ את ההתכנסות של שגיאת האימון (במחיר של זמן הרצה) וכי באופן מנוגד הקטנה זו מקטינה את קצב ההתכנסות של שגיאת ההכללה. עבור שגיאת הריצה התוצאה הגיונית שכן קבוע קטן גורר התכנסות מדויקת יותר על דוגמות ראשונות, ומהצד השני נוצר מצב שבו האימון מתאים רק לדוגמות אלה, לכן אפשר לראות ירידה איטית יותר של שגיאת ההכללה, ואף ירידה של שגיאת האימון בהתחלה.

עבור האלגוריתם המיידי, התוצאות הן לא מספיק טובות ויש לערוך ניסוי מפורט יותר, אבל אפשר לראות שבאופן כללי השגיאה משתנה מאוד כתלות בדוגמות, ולמעשה לא הגענו להתכנסות בפרמטרים שהזמנו, אין זה מפתיע, שכן האלגוריתם מתאים את עצמו מחדש לקלטים על מחיר קלטים אחרים, ולמעשה גורם לתנודות קשות.