מבוא למערכות לומדות - חורף 22/23

236756

HW4 dry

לוי הורביץ

313511602

שאלה 1

סעיף 1:

כן, f קמורה.

סעיף 2:

נגדיר:

$$g(x) = \begin{cases} 2x & x < 0 \\ 2 & x > 0 \\ 0 & x = 0 \end{cases}$$

נשים לב שלכל X > 0 ו- X < 0 הפונ' f גזירה ומתקיים $f\,' = g$ ובנוסף מתקיים שגם $f\,'$ גזירה (פנוקציות אלמנטריות).

נוכיח שהיא קמורה בתחום הנ"ל:

עבור $x < 0$ מתקיים $f\," = 2 \geq 0$ וכן עבור $x > 0$ מתקיים $f\," = 0$, לכן f קמורה לפי משפט.

ומכיוון ש- f גזירה וקמורה אז התת - גרדיאנט של f הוא הגרדיאנט שלה עצמה (*), ולכן מתקיים שלכל x בתחום הנ"ל אכן $g(x) \in \partial f(x)$.

נשאר להוכיח במקרה ש X = 0 שאכן $0 \in \partial f(0)$.

על מנת להוכיח שאכן $0 \in \partial f(0)$, צריך להוכיח על פי ההגדרה שמתקיים:

$$\forall v \in V: f(v) \geq f(0) + 0^T(v - 0)$$

$$\Leftrightarrow \forall v \in V: f(v) \geq 0 + 0 = 0$$

נשים לב שעל פי הגדרת f, אכן מתקיים:

$$\forall v \in V: f(v) \geq 0$$

והוכחנו ש-$0 \in \partial f(0)$. מש"ל.

(*) לפי ההגדרה:

$$\forall v \in V: f(v) \geq f(u) + q^T(v - u)$$

$$\Leftrightarrow \forall v \in V: \frac{f(v) - f(u)}{v - u} \geq q^T$$

וזה מתקיים באופן מיידי מתכונת קמירות

$$\forall v \in V: \frac{f(v) - f(x)}{v - x} \geq \frac{\partial f(x)}{\partial x} = g(x) = q^T$$

סעיף 3:

i	x_i	$f(x_i)$	$\frac{\partial}{\partial x}f(x_i)$
0	-1	*1*	*-2*
1	*-0.5*	*1/4*	*-1*
2	*-0.25*	*1/16*	*-0.5*
3	*-1/8*	*1/64*	*-0.25*
4	*-1/16*	*1/236*	*-1/8*

אפשר לראות לאחר 5 איטרציות שערכי *X* הולכים ומתקרבים ל- 0 (משמאל לציר *Y*) ובהתאמה ערכי הפונקציה הולכים וקרבים ל 0 עם מחזוריות שיטתית (ערכי הפונקציה נחתכים ברבע בכל איטרציה).

סעיף 4:

i	x_i	$f(x_i)$	$\frac{\partial}{\partial x}f(x_i)$
0	-1	*1*	*-2*
1	*1*	*2*	2
2	*-1*	*1*	*-2*
3	1	*2*	2
4	*-1*	*1*	*-2*

קל לראות לאחר 5 איטרציות שהפונקציה תקועה על הערכים 1 ו 2 ולכן הפונקציה לא תתכנס לערך המינימלי שלה שהוא ב 0.

<u>שאלה 2</u>

$$argmax_w \prod_{i=1}^{m} P(y_i, x_i; w) \underbrace{=}_{(1)} argmax_w \prod_{i=1}^{m} P(y_i|x_i; w) \cdot P(x_i|w) \underbrace{=}_{(2)}$$

$$= argmax_w \prod_{i=1}^{m} P(y_i|x_i; w) \cdot P(x_i) \underbrace{=}_{(3)} argmax_w \prod_{i=1}^{m} P(y_i|x_i; w) \underbrace{=}_{(4)}$$

$$= argmax_w \log \prod_{i=1}^{m} P(y_i|x_i; w) \underbrace{=}_{(5)} argmax_w \sum_{i=1}^{m} \log P(y_i|x_i; w) \underbrace{=}_{(6)}$$

$$= argmax_w \sum_{i=1}^{m} \log \frac{1}{2b} e^{-\frac{|y_i-\mu|}{b}} \underbrace{=}_{(7)} argmax_w \sum_{i=1}^{m} \log \frac{1}{2b} + \log e^{-\frac{|y_i-\mu|}{b}} \underbrace{=}_{(8)}$$

$$= argmax_w \sum_{i=1}^{m} \log \frac{1}{2b} - \frac{|y_i-\mu|}{b} \underbrace{=}_{(9)} argmax_w \sum_{i=1}^{m} - \frac{|y_i-\mu|}{b} \underbrace{=}_{(10)}$$

$$= argmin_w \frac{1}{b} \sum_{i=1}^{m} |y_i - \mu| \underbrace{=}_{(11)} argmin_w \sum_{i=1}^{m} |y_i - \mu| \underbrace{=}_{(12)} argmin_w \sum_{i=1}^{m} |\mu - y_i| \underbrace{=}_{(13)}$$

$$= argmin_w \sum_{i=1}^{m} |w^T x_i - y_i| \underbrace{=}_{(14)} argmin_w \frac{1}{m} \sum_{i=1}^{m} |w^T x_i - y_i|$$

הסברים למעברים:

1. על פי כלל ההסתברות השלמה.
2. X הוא בלתי תלוי ב-w, ולכן ניתן להשמיט את ההתניה ב-w.
3. $\prod_{i=1}^{m} P(x_i)$ היא סקלר שאינו תלוי בארגומנט אותו אנו מחפשים שמביא את הביטוי למקסימום, ולכן ניתן להשמיט אותו.
4. Log היא פונקציה מונוטונית עולה, ולכן הפעלתו על הביטוי אינו משנה את הארגומנט שמביא את הביטוי למקסימום.
5. על פי חוקי לוגריתמים.
6. על פי הנתון: $\varepsilon_i \sim Laplace(0, b)$, ומכיוון ש-$\varepsilon_i$ הוא הרעש של y בהינתן שאנו מסתכלים על x כלשהו, אז בעצם אנחנו מסתכלים רק על איך y יכול להתפלג כאשר x שלנו הוא קבוע. לכן בעצם אנחנו מקבלים: $y_i|x_i \sim Laplace(0, b)$ כי ההתפלגות של y נקבעת רק על פי ε_i שהוא מתפלג כפי שהסברנו לעיל.
7. על פי חוקי לוגריתמים.
8. על פי חוקי לוגריתמים.
9. $\sum_{i=1}^{m} \log \frac{1}{2b}$ הוא סקלר שאינו תלוי בארגומנט אותו אנו מחפשים שמביא את הביטוי למקסימום, ולכן ניתן להשמיט אותו.
10. הוצאת b החוצה מתוך הסכימה – b אינו תלוי במשתנה הסכימה. בנוסף, השמטת המינוס תוך חיפוש של הארגומנט שיביא את הביטוי למינימום ולא למקסימום – מעבר זה נכון מכיוון שהביטוי $\frac{|y_i-\mu|}{b} \geq 0$ (b גדול ממש מ-0 וערך מוחלט תמיד גדול שווה 0).
11. $\frac{1}{b}$ הוא סקלר שאינו תלוי בארגומנט אותו אנו מחפשים שמביא את הביטוי למינימום, ולכן ניתן להשמיט אותו.
12. על פי חוקי ערך מוחלט.

13. על פי הגדרת התפלגות לפלס: $\mathbb{E}[y_i] = \mu = w^T x_i$.
14. $\frac{1}{m}$ הוא סקלר שאינו תלוי בארגומנט אותו אנו מחפשים שמביא את הביטוי למינימום, ולכן ניתן להוסיף אותו.

<u>שאלה 3</u>

התשובה היא A.

B ו C לא, מכיוון שכל איטרציה מחלקת את המרחב בגרף לשניים ואין אפשרות לחלק את המרחב לשניים (בציר x ובציר y) בשאר הגרפים בלי שנקבל סתירה בנתונים כי לכל חלוקה שנעשה נקבל באותו צד פלוסים ומינוסים נכונים/שגויים וזה בלתי אפשרי.

D אפשר לחלק למשל ע"י כך שלכל $x \leq 0$ צריך להיות **(+)** ולכל $x > 0$ צריך להיות **(-)**. אבל אנחנו מניחים תמיד שמשתמשים ב-*ERM* בשביל לאמן את ה-*weak classifier*. זאת אומרת שהוא חייב להיות *decision stump* שמחזיר את השגיאה הכי נמוכה על ההתפלגות הנתונה. וב *D* יש 3 טעיות וב *A* יש 2 טעיות (באמצעות הסיווג החלש שנציע בהמשך) אז האלגוריתם מעדיף את האפשרות עם הטעות הכי קטנה שזאת *A*.

נציע *weak classifier* באופן הבא:

לכל $y \leq 3.5$ צריך להיות **(+)** ולכל $y > 3.5$ צריך להיות **(-)**, בהנחה שכל משבצת בגרף היא יחידה אחת.

ואכן אפשר לראות שהאלגוריתם בגרף *A* נותן משקל יתר ל- **(+)** מכיוון שהם בצד הלא נכון של הגרף בשונה מה- **(-)**, שהם לא מקבלים משקל יתר מכיוון שהם בצד הנכון של החלוקה בגרף, ואכן אפשר לראות שיש רק 2 טעיות במקום 3 (בשונה מ *D*).