29/6/20 - ב׳ # תרגול 13

נדבר על Convex Optimization (אופטימיזציה קמורה)

1. אלגוריתמים מבוססי קמירות לאופטימיזציה.
2. התאוריה של הקמירות.

מוטיבציה

עד כה ראינו אלגוריתמים שפותרים בעיות רגרסיה וסיווג, ע״י בחירת היפותזה מתוך מחלקת היפותזות באופן שממזער את ה- loss. מעולם לא דיברנו על הפרטים בבחירת ההיפותזה (שאינה מוצגת כפתרון סגור). באופטימיזציה קמורה אנחנו מנסים לענות על **השאלה** - כיצד נבחר את ההיפותזה הטובה ביותר מתוך מחלקת היפותזות?

זו שיטה שנועדה לבעיות יעילות למצוא אותה כמו בבעיות QP.

היום ניתן גישה כיצד לחפש במרחבי חיפוש גדולים. (היוריסטיקות).

דוגמא: בבעיית SVM אנחנו מחפשים מישור מפריד שממזער את הפונקציה: $\left[\frac{1}{n}\sum \max(0, 1-y_i(w\cdot x_i-b))\right]+\lambda\|w\|^2$, כלומר אנחנו מחפשים $(w,b)\in\mathbb{R}^{d+1}$, ומחלקת ההיפותזות (מרחב חיפוש) אינסופית ⇐ בעיה קשה.

1. אלגוריתמים מבוססי קמירות לאופטימיזציה

I. Gradient Descent

אלגוריתם איטרטיבי לפתרון בעיות אופטימיזציה. הוא ממומש את עקרון ה- ERM בכך שהוא מתחיל ממקום רנדומלי, כשהוא "הולך" בכיוון **ההפוך** לגרדיאנט (שמצביע על כיוון **העלייה** המקסימלי). נבחין שזהו לא אלגוריתם שמטרתו ללמוד היפותזה טובה אלא למצוא אותה מתוך המחלקה. כמו כן, הוא פועל לוקלית, (עשוי להתקע במינימום לוקלי) ולכן נריץ אותו מס׳ פעמים. כאשר אנחנו פועלים על פונק׳ קמורה, מינימום לוקלי הוא גלובלי.

האלגוריתם:

Parameters: T - number of iterations, η - learning rate >0

Initialize: $w^{(1)}=0$ (או קרוב ל-0 בצורה רנדומלית)

for $t=1,\dots,T$ do:

$$w^{(t+1)}=w^{(t)}-\eta\,\partial f(w^{(t)})$$ (גרדיאנט ביחס ל- training data)

Return: $w^{(T)}$

נבחין שבתהליך חישוב הגרדיאנט הוא יקר, כי הוא נעשה ביחס לכל ה-training.
לכן חשבו על השיפור במימון הרעיון הבא:

II) Stochastic Gradient Descent

גישה אחרת לבחירת הכיוון הטובה ביותר במחלקה, שנאמנה לרעיון הגרדיאנט.
לפיה, נחשב את הגרדיאנט ביחס לדוגמה אקראית, ומהתחום לערך היחידי הזה
נפעל. נבחין שבתוחלת הכיוון הזה מסכים עם הכיוון של הגרדיאנט עבור ה
מדגם האימון.

האלגוריתם:

Parameters: T - number of iterations, η - learning rate > 0

Initialize: $w^{(1)} = 0$

for $t = 1, .., T$ do:

$z_t \sim D$

$v_t \in \partial l(w^{(t)}, z_t)$

$w^{(t+1)} = w^{(t)} - \eta v_t$

Return: $w^{(T)}$

★ על הרצף בין הגרסא המקורית לזו הסטוכסטית ישנם אלגוריתמים רבים
שמוגדרים ע"י ה- batch size שלהם: כמות הדגימות עליהם הם
מחשבים את הגרדיאנט (המקורית: $b.s = m$, הסטוכסטית: $b.s = 1$).

2) התאוריה של הקמירות

I) בעית קמירות: בעיה מהצורה - $\min_{x \in C} f(x)$ כאשר:
f הוא פונקציה קמורה, C הוא קבוצה קמורה.

דוגמא:

- linear least squares: $\{x_k\}_{k=1}^{n}$ דגימות ב- $\mathbb{R}^d$, $\{y_k\}_{k=1}^{n}$ תיוגים ב- $\mathbb{R}$,
ואנחנו מניחים כי - $y_i \approx \langle w, x_i \rangle + b$ עבור $w \in \mathbb{R}^d$, $b \in \mathbb{R}$.
אנחנו מנסים למזער את ה- MSE.
נניח כי: $x = (3, 6, 9)$, $y = (4, 6, 8)$. נמצא את פרמטרי המודל ע"י
שימוש ב- Gradient Descent:

נחשב את הגרדיאנט של: $MSE(w,b) = \frac{1}{N}\sum_{i=1}^{N}\underbrace{(y_i-(wx_i+b))^2}_{\text{גזירה ריבועית} \to \text{קמורה ולכן האלגוריתם יתכנס ☺}}$

ונקבל:

$$(\star)\quad \frac{\partial MSE}{\partial w} = \frac{2}{N}\sum_{i=1}^{N} -x_i(y_i-(wx_i+b)) = \frac{2}{N}\sum_{i=1}^{n} -x_iy_i + wx_i^2 + x_ib$$

$$(\star)\quad \frac{\partial MSE}{\partial b} = \frac{2}{N}\sum_{i=1}^{N} -(y_i-(wx_i+b)) = \frac{2}{N}\sum_{i=1}^{N} -y_i + wx_i + b$$

נציב את x, y ונקבל את הביטויים הבאים:

לא גמיש ניתן להציב ישירות ☺

$$(\star\star)\quad \frac{\partial MSE}{\partial w} = -80 + 84w + 12b$$

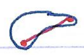
$$(\star\star)\quad \frac{\partial MSE}{\partial b} = 2b - 12 + 12$$

(★) נאתחל את הביטויים הללו (w,b), ונתקדם בכיוון ההפוך מהגרדיאנט, שוב ושוב עד להתכנסות של הערכים w, b (כלומר נפעל לפי אחד מ-2 האלגו' שהצגנו).

(★★) כאן אפשר פשוט לפתור את מע' המשוואות, כי X^TX הפיכה.

- פתרון הבעיה בעזרת דוגמת קוד, הרצאה 1 ב-15:00 עד לסוף –
- מהדוגמא אפשר לראות כיצד אלגוריתמי ה- gradient descent, Stochastic gradient descent לומדים עבור פרמטרי learning rate, מס' איטרציות, ו- batch size שונים.

(II) קבוצות קמורות: יהא V מ"ו. $C \subseteq V$ נקראת קמורה אם לכל 2 וקטורים $u,v \in C$ ולכל $\alpha \in [0,1]$ סקלר מתקיים: $\alpha u + (1-\alpha)v \in C$.

לא קמורה –  , קמורה – 

דוגמאות:
- כל תת מרחב לינארי הוא קבוצה קמורה: $U \subseteq V$ תת"מ לינארי מקיים שלכל $u,v \in U$, $\alpha \in [0,1]$: $\alpha u + (1-\alpha)v \in U$ כי מהומוגניות ואדטיביות של תת המרחב.
- כדורי היחידה: $B = \{v \in V : \|v\| \le 1\}$. יהיו $u,v \in B$, $\alpha \in [0,1]$ אזי: מא"ש המשולש מתקיים:

$$\|\alpha u + (1-\alpha)v\| \le \|\alpha u\| + \|(1-\alpha)v\| = \alpha\underbrace{\|u\|}_{\le 1} + (1-\alpha)\underbrace{\|v\|}_{\le 1} \le$$
$$\le \alpha + (1-\alpha) = 1$$

ומכיוון ש- $w = \alpha u + (1-\alpha)v \in V$ מקיים $\|w\| \le 1$, מתקיים: $w \in B$.

- חצאי מרחב סגורים: $W = \{v \mid \langle w,v\rangle \le b\}$ (עבור $0 \ne w \in V$, $b \in \mathbb{R}$) מתקיים ל- $u,v \in W$, $\alpha \in [0,1]$:

$\langle w, \alpha u + (1-\alpha)v\rangle = \alpha\langle w,u\rangle + (1-\alpha)\langle w,v\rangle \le \alpha b + (1-\alpha)b = b$

ומכיוון ש- $\alpha u+(1-\alpha)v \in V$ מקיים: $\langle w, \alpha u+(1-\alpha)v\rangle \le b$

מתקיים: $\alpha u + (1-\alpha)v \in W$.

טענה 1:

• 1. חיתוך קבוצות קמורות $C=\bigcap_{i\in I} C_i$ הוא קבוצה קמורה.
C_i is convex

• 2. הקב' $\lambda C = \{\lambda c \mid c\in C\}$ עבור C קמורה ו- $\alpha\in\mathbb{R}$ היא קמורה.

הוכחה: יהא $\alpha\in[0,1]$.

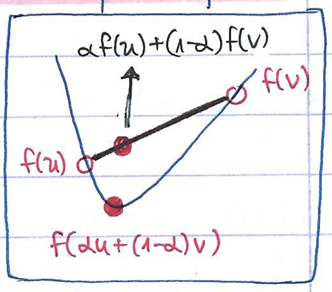
1. יהיו $u,v\in C=\bigcap_{i\in I}C_i$, צ.ל: ש: $\alpha u+(1-\alpha)v\in C$.
מהנתון מתקיים שלכל $i\in I$: $u,v\in C_i$ קבוצה קמורה, ולכן
לכל $i\in I$: $\alpha u+(1-\alpha)v\in C_i$. מכך נובע ש:
$\alpha u+(1-\alpha)v\in\bigcap_{i\in I}C_i = C$ כנדרש, ו- C קב' קמורה.

2. יהיו $c_1,c_2\in\lambda C$. צ.ל: $\alpha c_1+(1-\alpha)c_2\in\lambda C$. מכך ש: $c_1,c_2\in\lambda C$
נובע שקיימים $c_1',c_2'\in C$ שמקיימים: $c_1=\lambda c_1'$, $c_2=\lambda c_2'$.
מכך ש- C קמורה מתקיים: $\alpha c_1'+(1-\alpha)c_2'\in C$
ולכן: $\lambda C \ni \lambda(\alpha c_1'+(1-\alpha)c_2') = \alpha\lambda c_1'+(1-\alpha)\lambda c_2' = \alpha c_1+(1-\alpha)c_2$
ולכן λC קבוצה קמורה, כנדרש.

■

III פונקציות קמורות: יהא V מ"ו, ותהא C קבוצה קמורה. פונקציה $f: C\to\mathbb{R}$ נקראית פונק' קמורה אם לכל $u,v\in C$ ולכל $\lambda\in[0,1]$ מתקיים:

$$f(\lambda u+(1-\lambda)v)\le\lambda f(u)+(1-\lambda)f(v)$$



f נקראת strictly convex אם אי-השוויון חזק מתקיים רק עבור $u\neq v$ ו- $\lambda\in(0,1)$. (כלומר מתקיים רק עבור המצבים המנוונים).

משמעות גאומטרית:

★ כל מיתר שמחבר 2 נקודות על הגרף יהיה תמיד $\geq$ מהגרף.

★ בשיעור ראינו שמשמעות נוספת היא שה-Epigraph (קב' הנקודות שנמצאות מעל לגרף הפונק') היא קבוצה קמורה.

דוגמאות: • הנורמה היא פונק' קמורה: לכל $u,v\in V$, $\alpha\in[0,1]$ מא- $\Delta\neq$:

$\|\alpha u+(1-\alpha)v\| \le \|\alpha u\|+\|(1-\alpha)v\| = \alpha\|u\|+(1-\alpha)\|v\|$

(ואכן עבור $f(x)=\|x\|$ התנאי מתקיים)

• פונקציה אפינית ל-R: יהיו $w\in V$, $b\in\mathbb{R}$. פונק' אפינית היא $f:V\to\mathbb{R}$ המוגדרת להיות $f(u)=\langle w,u\rangle+b$. יהיו $u,v\in V$, $\alpha\in[0,1]$, אז:

$$f(\alpha u+(1-\alpha)v)=\langle w,\alpha u+(1-\alpha)v\rangle+b=$$
$$=\alpha(\langle w,u\rangle+(1-\alpha)\langle w,v\rangle)+b=$$
$$=\alpha f(u)+(1-\alpha)f(v)$$

אפשר לייצר פונק' קמורות ע"י הרכבה וסכום של פונק' קמורות:

טענה 1: • יהיו $f_i:V\to\mathbb{R}$ ל-$i\in[m]$ פונק', ויהיו $\gamma_1,...,\gamma_m\in\mathbb{R}>0$.
נתבונן בפונקציה: $g:V\to\mathbb{R}$ הנתונה ע"י: $g(u)=\sum_{i=1}^{m}\gamma_i f_i(u)$
אם $\{f_i\}_{i=1}^{m}$ הן פונקציות קמורות, אזי גם g.

• 2 יהיו: $f:\mathbb{R}^m\to\mathbb{R}$, $A\in\mathbb{R}^{m\times n}$, $b\in\mathbb{R}^m$. נגדיר:
$g:\mathbb{R}^n\to\mathbb{R}$ ע"י $g(u)=f(Au+b)$.
אם f קמורה, אזי גם g.

• 3 יהיו $f_i:V\to\mathbb{R}$ ל-$i\in I$, ותהיה $f:V\to\mathbb{R}$
הנתונה ע"י: $f(u)=\sup_{i\in I} f_i(u)$.
אם $\{f_i\}_{i\in I}$ הן פונקציות קמורות אזי גם f.

הוכחה: • 1 תרגיל 6.

• 2 יהיו $u,v\in\mathbb{R}^n$, $\alpha\in[0,1]$. אזי:

$$g(\alpha u+(1-\alpha)v)=f(A(\alpha u+(1-\alpha)v)+b)=$$
$$=f(\alpha(Au+b)+(1-\alpha)(Av+b))\leq$$
$$\leq\alpha f(Au+b)+(1-\alpha)(f(Av+b))=$$
$$=\alpha g(u)+(1-\alpha)g(v)$$

כלומר g פונק' קמורה, כנדרש.

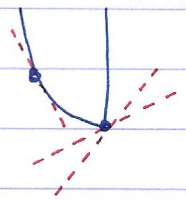
• 3 תרגיל 6: ע"י כך שנסמן f חיתוך האפיגרפים (קב' קמורה כחיתוך קב' קמורות) כאפיגרף של f.

$\blacksquare$

אבחנה: ניזכר בפונק' ה-squared loss: $f:\mathbb{R}^d\to\mathbb{R}$ המוגדרת ע"י $f(w)=(\langle w,x\rangle-y)^2$. היא פונק' אפינית, והפונק' $\alpha\mapsto\alpha^2$ היא פונק' סקלרית ולכן קמורה, ולכן מטענה 2 נקבל ש-f קמורה.

IV Subgradient: תהא $f:V\to\mathbb{R}$, אז $g\in V$ הוא ה-subgradient של f בנק' $u\in V$ אם לכל $v\in V$ מתקיים: $f(v)\geq f(u)+\underbrace{\langle g,v-u\rangle}_{\text{שעתוק לינארי}}$
קב' ה-subgradient של f ב-$u\in V$ מסומן ב-$\partial f(u)$.

★ רעיון ה-subgradient מאפשר לנו לדבר על "דיפרנציאביליות" גם בפונקציות קמורות שאינן דיפרנציאביליות.



subgradients $\Rightarrow$ [הרעיון: מבטיח שה-subgradient נמצא תמיד מתחת לגרף הפונקציה.]

טענה: אם f דיפרנציאבילית ב-x, אז: $\partial f(x) = \{\nabla f(x)\}$.

טענה 3: לכל $i \in [n]$, תהא $f_i : V \to \mathbb{R}$ פונקציה קמורה. נגדיר: $f: V \to \mathbb{R}$ ע"י $f(u) = \max_{i \in [n]} f_i(u)$. יהא $u \in V$. נגדיר $j = \operatorname{argmax}_{i\in[n]} f_i(u)$. אז: $\partial f_j(u) \subseteq \partial f(u)$.

הוכחה: תהא $g \in \partial f_j(u)$. אז מהגדרת f, j ומהגדרת subgradient מתקיים:

$$f(v) \underset{f \text{ מקסימום של } v}{\geq} \overbrace{f_j(v) \geq f_j(u) + \langle g, v-u \rangle}^{g \in \partial f_j(u) \text{ כי}} \underset{j \text{ הוא argmax}}{=} f(u) + \langle g, v-u \rangle$$

כלומר g הוא אכן subgradient של f ב-u. ∎

אבחנה: נתבונן ב-$f(x)=|x|$. נבחין שניתן לראות אותה ע"י $f(x) = \max\{x, -x\}$.
מטענה 3 מתקיים:

$$\partial f(x) = \begin{cases} 1 & : x>0 \\ -1 & : x<0 \end{cases}$$

ושי- $\{-1,1\} \in \partial f(0)$, וזה הגיוני כי לכל אחד מהתחומים הפונק' גזירה, וב-0 יש כמה subgradients, ובפרט $\{-1,1\}$.

טענה: תהא $f: V \to \mathbb{R}$ פונקציה קמורה. נניח שעבור $\bar{w} \in V$ כלשהו מתקיים: $0 \in \partial f(\bar{w})$. אז: $\bar{w}$ הוא מינימום גלובלי של f.

הוכחה: תהא $w \in V$. מתקיים: $f(w) \geq f(\bar{w}) + \langle 0, w-\bar{w} \rangle = f(\bar{w})$

הגדרת subgradient ב-$\bar{w} = 0$.

ולכן $\bar{w}$ אכן מינימום גלובלי. ∎

V: נגזרת מסדר גבוה, והקשר לקמירות:

טענה 4: תהא פונק' דיפרנציאבילית $f: \mathbb{R}^n \to \mathbb{R}$. f קמורה $\Leftrightarrow$ לכל $x,y \in Dom(f)$ מתקיים: $f(y) \geq f(x) + \nabla f(x)^T (y-x)$

דוגמא: יהא $c\in\mathbb{R}^n$. נגדיר: $f:\mathbb{R}^n\to\mathbb{R}$ ע"י $f(x)=\langle c,x\rangle$.

f קמורה מכיוון שלכל $x\in\mathbb{R}^n$: $\nabla f(x)=c$, ולכן:

$$f(y)=c^Ty=c^Tx+c^T(y-x)=f(x)+\nabla f(x)^T(y-x)$$

טענה 5: פונק' גזירה פעמיים $f:\mathbb{R}^n\to\mathbb{R}$ קמורה $\iff$

לכל נק' x ב- Dom(f) ההסיאן בנק' x הוא מטריצה

psd (positive semi-defined): $\underbrace{\nabla^2}_{\text{הסיאן}} f(x)\underbrace{\succeq}_{psd} 0$.

סימונים:

דוגמא: תהא $A\in\mathbb{R}^{n\times n}$, ונגדיר: $f:\mathbb{R}^n\to\mathbb{R}$ ע"י $f(x)=x^TAx$.

אזי f קמורה $\iff$ $A\succeq 0$:

לכל $x\in\mathbb{R}^n$ מתקיים: $\nabla f(x)=2x^TA$, ולכן: $\nabla^2 f(x)=2A$

אזי $A\succeq 0 \iff \nabla^2 f(x)\succeq 0$.

* נבחין שבמקרה ש-f פונק' סקלרית: $f:\mathbb{R}\to\mathbb{R}$, טענה 5 מצטמצמת

לתנאי שמצאנו בתיכון: $f''(x)\geq 0$.