1.1 kernels

א) יהי: $k(x,x')$ PSD kernel חוקי. ספקו PSD kernel חוקי

$\tilde{k}(x,x')$ שנבנה מ k שמובטח שיהיה מנורמל.

כלומר לכל x מתקיים $\tilde{k}(x,x)=1$

הוכחה אינטואיטיבית:

נגדיר ~~ר~~ $\tilde{k}:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ כך:

$$\tilde{k}(x,x')=\frac{k(x,x')}{\sqrt{k(x,x)}\cdot\sqrt{k(x',x')}}$$

ראשית נראה כי $\tilde{k}(x,x')$ הוא גרעין PSD תקף

$\tilde{k}(x,\bar{x})$ שנית לכן כל שנשאר הוא להראות ש הוא PSD מנורמל.

ממשפט מרסר נקבל שאם $k(x,x')$ הוא גרעין תקף

אז ניתן לכתוב אותו כך $k(x,x')=\Phi(x)^T\Phi(x')$

ולכן נקבל

$$\tilde{k}(x,x')=\frac{k(x,x')}{\sqrt{k(x,x)}\cdot\sqrt{k(x',x')}}=\frac{\Phi(x)^T\Phi(x')}{\sqrt{\Phi(x)^T\Phi(x)}\sqrt{\Phi(x')^T\Phi(x')}}=$$

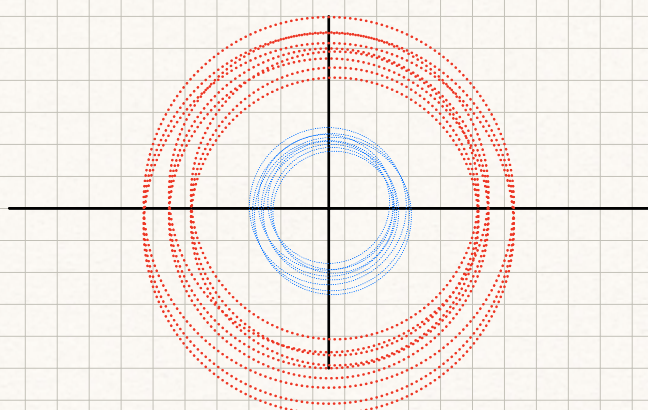
$$=\frac{\Phi(x)^T\Phi(x')}{\|\Phi(x)\|\,\|\Phi(x')\|}=\tilde{\Phi}(x)^T\cdot\tilde{\Phi}(x')$$

כאשר $\tilde{\Phi}(x)=\frac{\Phi(x)}{\|\Phi(x)\|}$ ולכן $\tilde{k}$ הוא PSD מנורמל.

$\blacksquare$

2. Consider a data set $S=\{(x_i,y_i)\}_{i=1}^m$ where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1,1\}$, and a feature map $\psi: \mathbb{R}^d \to \mathcal{F}$ where $\mathcal{F}$ is some feature space. Give an example of a data set S and a feature map ψ such that S is not linearly separable in $\mathbb{R}^d$ $(d \geq 2)$ but that the transformed dataset: $S_\psi=\{(\psi(x_i),y_i)\}_{i=1}^m$ is linearly separable in $\mathcal{F}$.

נתבונן בסט מידע שראינו בתרגול שנראה כך:



אם נשתמש בפונקציה (map) הבאה נקבל דאטה שנפריד

לינארית ב $\mathbb{R}^d$ $\psi(x_1,x_2)=(x_1,x_2,x_1^2+x_2^2)$

הנתון השלישי הוא מחושב במרחק מהראשית אוקלידית ולכן

כאשר הדאטה מחולק למחלקות לפי צבע, אם ננסה את הנתון

של מרחק מהראשית נוכל לקבל מידע שניתן להפריד לינארית

ב) $k_1(x,y)$ ו- $k_2(x,y)$ הם גרעינים ולידים אז:

$$k_\times(x,y) = k_1(x,y)\cdot k_2(x,y)$$

היא גם גרעין ולידי. כדי להוכיח זאת נשתמש בעובדה ש

גרעין ולידי הוא PSD.

אפשר להיעזר בכללים הבאים

(1) $$x^T A y = Tr[x^T A y] = Tr[y x^T A]$$

(2) $$Tr[AB] = \sum_i [AB]_{ii} = \sum_i \sum_j A_{ij} B_{ji}$$

כאשר x ו- y הם וקטורים A,B מטריצות

(a) יהי $k(x,y)$ valid kernel ונתייחס ל- k היא הגרעין של מטריצת הגרם מעל סט סופי של נקודות $\{X_i\}_{i=1}^N$ כך ש- $k_{ij}=k(x_i,x_j)$ הראה של כל סט סופי של N נקודות קיים פונקציה $f: \mathcal{X}\to\mathbb{R}^N$ כך ש

$$k(x_i,x_j) = f^T(x_i) f(x_j) \quad (3)$$

כאשר $\mathcal{X}$ היא מרחב של נקודות x_i. בהשתמש בנתון הנ"ל הראה ש:

$$k_1(x,y)\cdot k_2(x,y) = \sum_i \sum_j g_i(x) f_j(x) f_j(y) g_i(y)$$

כאשר $k_1(\cdot,\cdot)$ ו- $k_2(\cdot,\cdot)$ הם גרעינים ולידים והפונקציות $f,g: \mathcal{X}\to\mathbb{R}^N$ כאשר $f_i(x)$ נותן את האינדקס ה-i של הפונקציה $f(x)$.

(a) נראה ש-k הוא קרנל אם לכל N נקודות $\{x_i\}_{i=1}^N$ קיימת פונקציה:

$f: \mathcal{X} \to \mathbb{R}^m$ כך ש- $k(x_i, x_j) = f(x_i)^T f(x_j)$ לכל i,j.

יהי K מטריצת גרם של k $k(x,y)$ עם לכל אוסף $\{x_i\}$

נגדיר $K_{ij} = k(x_i, x_j)$

מאחר ו- $k(x,y)$ הינו valid kernel, המטריצה K הינה

PSD, כלומר לכל וקטור $z = [z_1, z_2, \dots, z_N]^T$

נקבל $z^T K z \geq 0$

נתבונן במטריצה: $A = [f(x_1), f(x_2), \dots, f(x_N)]$

מטריצת הגרם k $k(x,y)$ בהינתן כל $\{x_i\}_{i=1}^N$ והפונקציה

f נתונה ע"י $A^T A$.

כעת נפתח את $z^T(A^TA)z$ בעזרת הביטויים (1) ו-(2):

$$z^T(A^TA)z = Tr\left(z^T(A^TA)z\right) \overset{(1)}{=} Tr\left((Az)^T(Az)\right)$$

$$\overset{(2)}{=} \sum_i (Az)^T_i (Az)_i = \sum_i\sum_j (Az)_i (Az)_j = \sum_i\sum_j (z_i A_i)^T (A_j z_j)$$

כעת נגדיר וקטור את $f(x) = (f_1(x), \dots, f_m(x))^T$

כאשר $f_i(x)$ היא האיבר ה-i ב-$f(x)$.

נקבל

$$\sum_i\sum_j (z_i A_i)^T (A_j z_j) = \sum_i\sum_j z_i \left(f_i(x)^T f_j(x)\right) z_j =$$

$$= \sum_i\sum_j z_i \left(f_j(x)^T f_i(x)\right) z_j = \sum_i\sum_j g_i(x) f_j(x) f_j(y) g_i(y)$$

כאשר $g_i(x) = z_i$ ו- $g_j(y) = z_j$

כן הראינו ש $k(x_i,x_j) = f(x_i)^T f(x_j)$ לכל i,j כאשר

$$f(x) = [f_1(x), \dots, f_n(x)]^T \qquad g(x) = (g_1(x), \dots, g_m(x))^T$$

☒

(b) בהינתן הביטוי:

$$k_\times(x,y) = k_1(x,y)\,k_2(x,y) =$$

$$= \sum_i \sum_j g_i(x) f_j(x) f_j(y) g_i(y)$$

נוכל לכתוב

$$k_\times(x,y) = h(x)^T h(y)$$

כאשר

$$h(x) = [g_1(x) f_1(x), g_1(x) f_2(x), \dots, g_n(x) f_n(x)]^T$$

לכן הראינו ש $k_\times(x,y) = h(x)^T h(y)$

עבור פונקציה $h(\cdot)$ מה שמוכיח ש $k_\times(\cdot,\cdot)$ הוא valid kernel. ☒

PCA

4) יהי $X:\Omega\to\mathbb{R}^d$ מ"מ עם ממוצע 0 ו Cov $\Sigma\in\mathbb{R}^{d\times d}$

הראה ש לכל $v\in\mathbb{R}^d$ כך ש $\|v\|_2=1$ השונות

של $\langle v,X\rangle$ היא לא גדולה יותר מהשונות שמושגת

ע"י PCA הממפה את X בתת מרחב ממימד אחד

(הנחה ש הPCA מחושב ע"פ Σ)

יהי $v\in\mathbb{R}^d$ כך ש $\|v\|_2=1$ ונסמן $\tilde{X}=\langle v,X\rangle$

ראשית התוחלת של $\tilde{X}$ היא

$$E(\tilde{X}) = \langle v, E(X)\rangle = 0$$

השונות של $\tilde{X}$ היא:

$$\mathbb{E}\left(\left(\tilde{X}-\mathbb{E}(\tilde{X})\right)^2\right)=\mathbb{E}(\tilde{X}^2)=\mathbb{E}\left[\langle v,x\rangle\cdot\langle v,x\rangle\right]=$$

$$=v^T\mathbb{E}\left[X\cdot X^T\right]v=v^T\Sigma v\le u_1^T\Sigma u_1$$

מכיוון של וקטורים עצמיים
כפי שראינו בתרגול

PCA עושה שיקוף על המידע בכיוון u_1.

($u_1\in\mathbb{R}^d$ היא הוקטור העצמי המקסימלי של Σ)

Convex optimization 1.3

1) יהיו $f_1,\dots,f_m: C\to R$ פונקציות קמורות ! $\gamma_1,\dots,\gamma_m\in\mathbb{R}$

הוכח מוגדר? g

$$g(u)=\sum_{i=1}^{m}\gamma_i f_i(u)$$

היא פונקציה קמורה? .

מוגדרת: C היא קבוצה קמורה, וכל $f_1,\dots,f_m$ היא פונקציה קמורה, כלומר לכל $i\in[m]$ $u,v\in C$ ולכל $\alpha\in[0,1)$ מתקיים:

$$f_i(\alpha v+(1-\alpha)u)\le\alpha f_i(v)+(1-\alpha)f_i(u)$$

(נכפיל משני הצדדים ב γ_i (סכ״ג) לכן אין שינוי בסימן)

$$\gamma_i f_i(\alpha v+(1-\alpha)u)\le\gamma_i(\alpha f_i(v)+(1-\alpha)f_i(u))$$

נתבונן בגדלו

$$\gamma_1 f_1(\alpha v+(1-\alpha)u)+\dots+\gamma_m f_m(\alpha v+(1-\alpha)u)\le$$ (לכל $i\in[m]$ מתקיים אי־השוויון לבדו)

$$\gamma_1(\alpha f_1(v)+(1-\alpha)f_1(u))+\dots+\gamma_m(\alpha f_m(v)+(1-\alpha)f_m(u))$$

כלומר

$$\sum_{i=1}^{m}\gamma_i f_i(\alpha v+(1-\alpha)u)\le\sum_{i=1}^{m}\gamma_i(\alpha f_i(v)+(1-\alpha)f_i(u))$$

מכאן ופונקציה $g(u)=\sum_{i=1}^{m}\gamma_i f_i(u)$

קמורה מוגדרת. $\boxed{//}$

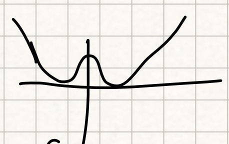
2) תן דוגמא נגדית לטענה הבאה: תהיינה $f,g:\mathbb{R}\to\mathbb{R}$ נגדיר פונקציה:

$h=f\circ g$. אם f,g קמורות אז h קמורה.

פונקציה קמורה: $f(x)=(x-1)^2$

פונקציה קמורה: $g(x)=x^2$

לא קמורה: $h(x)=(x^2-1)^2$



3) יהי: $f:C\to\mathbb{R}$ פונקציה המוגדרת על C מרחב קמור

הוכח ש f קמורה אמ"מ האפיגרף היא קמור

$$epi(f)=\{(u,t): f(u)\le t\}$$

אם f קמורה אז: נניח בשלילה ש $epi(f)$ לא קמורה:

כלומר קיימים שתי נקודות $(u_1,t_1),(u_2,t_2)$ כך ש

קיים $\alpha\in[1,0]$ כך ש

$$f(\alpha u_1+(1-\alpha)u_2) > \alpha t_1+(1-\alpha)t_2$$

אבל זו בסתירה לכך ש f קמורה. $\boxed{\ }$

4) יהי: $f_i: V\to\mathbb{R}$, $i\in I$, $f:V\to\mathbb{R}$

$$f(u)=\sup_{i\in I} f_i(u)$$

אם f_i קמורות לכל $i\in I$ אז f קמורה.

נניח בשלילה ש f לא קמורה כלומר
קיימים $x,y\in V$ ו $\alpha\in[0,1]$ כך ש

$$f(\alpha x+(1-\alpha)y) > \alpha f(x)+(1-\alpha)f(y)$$

מהגדרת f, קיים $i\in I$ כך ש $f(\alpha x+(1-\alpha)y)=f_i(\alpha x+(1-\alpha)y)$.
מהגדרת f מתקיים ש $f(x)\ge f_i(x)$ וגם $f(y)\ge f_i(y)$
כלומר

$$f_i(\alpha x+(1-\alpha)y) > \alpha f(x)+(1-\alpha)f(y) \ge \alpha f_i(x)+(1-\alpha)f_i(y)$$

כלומר קיים $i\in I$ המקיים את אי השוויון

$$f_i(\alpha x+(1-\alpha)y) > \alpha f_i(x)+(1-\alpha)f_i(y)$$

בסתירה לנתון שלכל $i\in I$ f_i קמורה
ולכן $f(u)=\sup_{i\in I} f_i(u)$ קמורה

1.4 Sub gradient for Soft SVM Objective

5. בהינתן $x\in\mathbb{R}^d$ ו $y\in\{\pm1\}$ הראה ש hinge loss היא קמורה ב w,b
כלומר הגדר

$$f(w,b):=\ell^{hinge}_{x,y}(w,b)=\max(0,1-y(x^Tw+b))$$

נראה ש f קמורה ב w,b.
מכיוון ש $1-y(x^Tw+b)$
פונקציה לינארית ב w,b היא קמורה (מקיימת את אי השוויון)
אפינית, הפונקציה 0 גם פונקציה קמורה
והמקסימום בין שתי פונקציות קמורות הינו פונקציה קמורה

6. הסק את הגרדיינט של ה hinge loss $g\in\partial\ell^{hinge}_{xy}(w,b)$.

הפונקציה גזירה

$$g_{xy}(w,b) = \begin{cases} -y & \text{if } y\cdot(X^T W+b) < 1 \\ 0 & \text{if } y(X^T W+b) \geq 1 \end{cases}$$

7. יהיו $f_1,\dots,f_m:\mathbb{R}^d\to\mathbb{R}$ ס של פונקציות קמורות ! $g_k\in\partial f_k(x)$

לכל $k\in[m]$ יהי ה סאב גרדינט k בנקודה

הגדר $f:\mathbb{R}^d\to\mathbb{R}$ ע״י $f(x)=\sum_{i=1}^{m} f_i(x)$

הוכיחו

$$\sum_k g_k \in \partial \sum_k f_k(x)$$

$$\partial f(x) = \partial \sum_{k=1}^{m} f_k(x) = \sum_{k=1}^{m} \partial f_k(x) = \{ g_1+g_2+\dots+g_m \mid g_i\in\partial f_i \}$$

ומשיג כן

$$\forall\, i\in[m] \qquad g_k\in\partial f_k(x)$$

אכן:

$$\sum_k g_k \in \partial \sum_k f_k(x)$$

$\blacksquare$

8)

$$\partial f(w,b) = \partial\left(\frac{1}{m}\sum_{i=1}^{m} \ell^{hinge}_{x_i y_i}(w,b) + \frac{\lambda}{2}\|w\|^2\right)$$

$$= \frac{1}{m}\sum_{i=1}^{m} \partial \ell^{hinge}_{x_i y_i}(w,b) + \partial \frac{\lambda}{2}\|w\|^2 =$$

נגדיר ; (כמו בסעיף (6)):

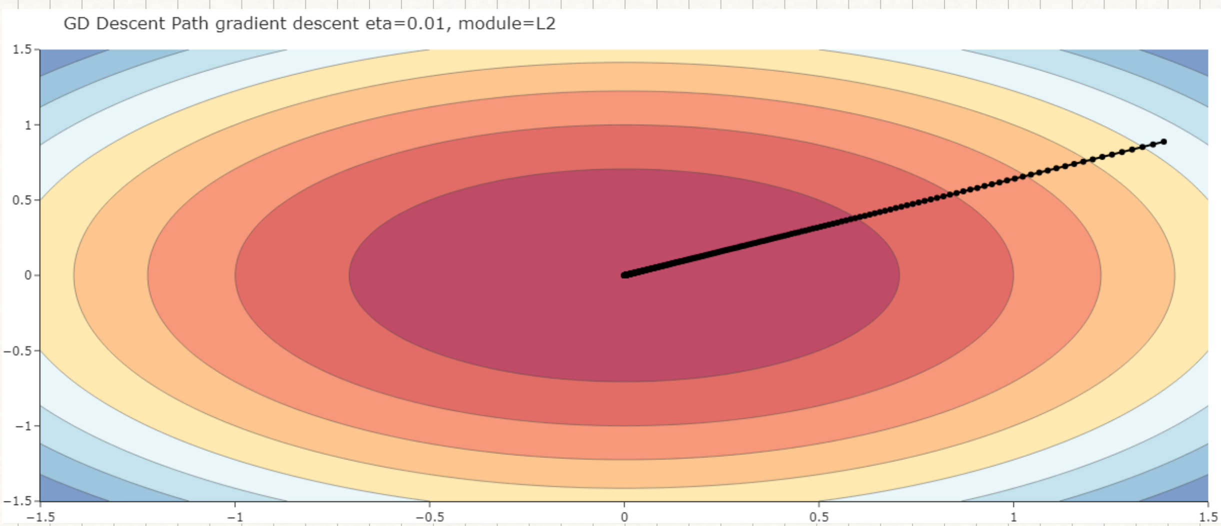
$$\left(g_{xy_i}(w,b) = \begin{cases} -y_i & \text{if } y_i(X_i^T W+b) < 1 \\ 0 & \text{if } y_i(X_i^T W+b) \geq 1 \end{cases} \right)$$

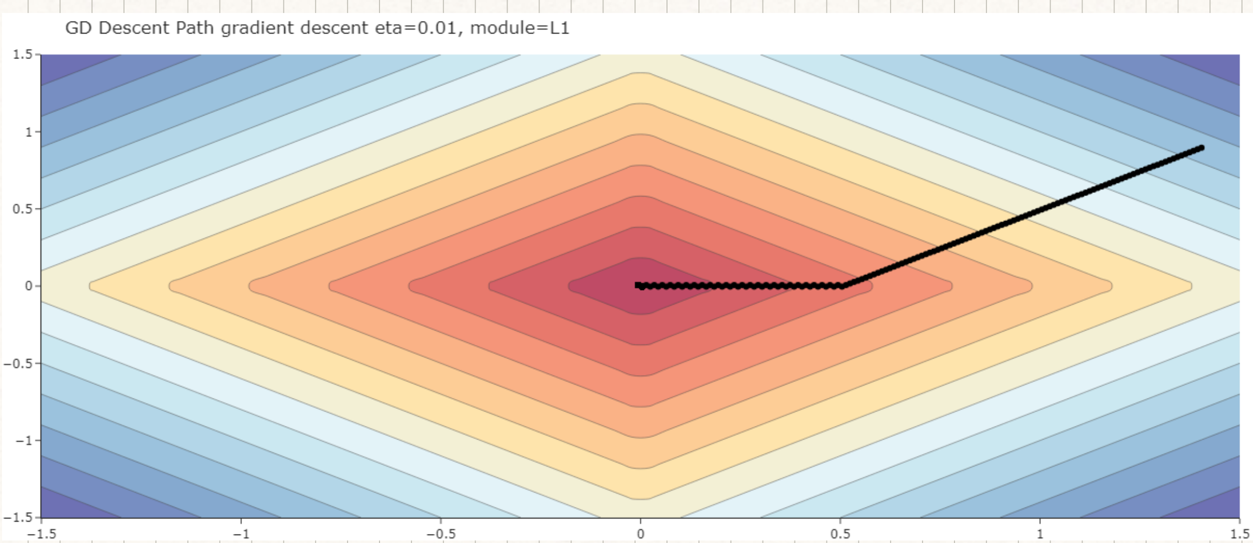
$$= \boxed{\frac{1}{m}\sum_{i=1}^{m} g_{x_i,y_i}(w,b) + \lambda\|w\|}$$

לכן הבעיה שקולה
ל f עבור w

2.1.1

1)





בשני המודלים הכיוון נקבע ע"י הנקודה הכי "נמוכה" ביחס לנקודה הנוכחית (בסביבה של 0.01 סביב הנקודה הנוכחית).
ב L2, לאורך כל התהליך הנקודה הכי נמוכה כלומר הכיוון ההפוך לכיוון הגרדיאנט היה לכיוון מרכז העיגול.

בעוד בגל יש קטע בהתחלה שבו כיוון הגרדיינט פונה לכיוון
אחר מכיוון המרכז. וגם כשהוא מגיע לנקודה (0,05) בערך
כיוון הירידה משתנה ופונה יותר לכיוון המרכז.
(ניתן לראות שגם זה לא מקוין אלא יצא זיגזג
ש באמת על קבר נגמר במרכז).

מתנצל לא הספקתי להגיש הכל. עשיתי כמיטב יכולתי.
אמיר קלמן.