תרגול 4: פונקציות גרעין (Kernels)

מסווג לינארי כללי

בהינתן קלט $x\in\mathbb{R}^d$ ובעיית סיווג בינארי $y\in\{-1,1\}$, מסווג לינארי כללי הינו מהצורה $\hat{y}=f(x)=sign(w^Tx+b)$ (לדוגמה LDA, Logistic Regression, פרספטרון, SVM). נזכיר שאת הקבוע b אפשר לשרשר לווקטור w אם מוסיפים ל x קואורדינטה קבועה, כלומר הקלט הופך ל $\begin{bmatrix} x \\ 1 \end{bmatrix}$.

תזכורת: בהנתן מרחב ווקטורי V, נגדיר עבור תת-קבוצה $X\subset V$ את אוסף כל הצירופים הלינאריים של ווקטורים מ X:

$$\text{span}(X)=\left\{\sum_{i=1}^n\alpha_i x_i \mid x_i\in X,\alpha_i\in\mathbb{R}\right\}$$

טענה: במקרים רבים מתקיים $w\in\text{span}\left(\{x_i\}_{i=1}^n\right)$, כלומר אפשר לכתוב את w כצרוף לינארי של סט הדוגמאות:

$$w=\sum_{i=1}^n\alpha_i x_i ,$$

עבור $\alpha_i\in\mathbb{R}$. את המסווג ניתן לכן לכתוב ע"י $\hat{y}=f(x)=sign\left(\sum_{i=1}^n\alpha_i x_i^T x+b\right)$.

נשים לב כי ב $f(x)$ התלות בקלט מופיעה רק כמכפלות פנימיות x_i^Tx.

<u>דוגמה: Logistic regression</u>

מודל: ממדלים ישירות את ההתפלגות האפוסטריורית של המחלקה בהנתן הקלט:

$$\mathbb{P}(y\mid x)=\frac{e^{yw^Tx}}{e^{w^Tx}+e^{-w^Tx}}$$

סיווג: מההתפלגות האפוסטריורית (וכמו שראיתם בהרצאה), מקבלים מסווג לינארי:

$$\mathbb{P}(y=1\mid x)>\mathbb{P}(y=-1\mid x)\qquad\Leftrightarrow\qquad w^Tx>0$$

למידה: ראיתם בהרצאה שיש למזער את סכום ה log-loss על סט הלימוד:

$$(*)\qquad w^*=\arg\min_w\left\{\sum_{i=1}^n\log\left(1+e^{-2y_iw^Tx_i}\right)\right\}$$

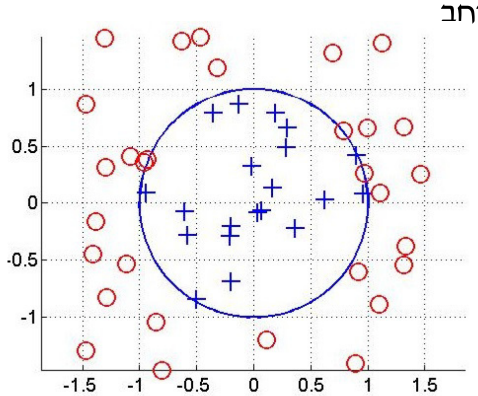
נניח שמצאנו פתרון w^*. נשים לב שאפשר לפרק את w^* לרכיב שמקביל לתת-המרחב שנפרש ע"י סט הלימוד, ולרכיב שניצב לתת-המרחב הנ"ל (זה נכון לכל ווקטור), כלומר $w^*=w^*_{\parallel}+w^*_{\perp}$, כאשר $w^*_{\parallel}\in\text{span}\left(\{x_i\}_{i=1}^n\right)$ ו $w^*_{\perp}\in\text{span}\left(\{x_i\}_{i=1}^n\right)^{\perp}$.

ע״י הצבה ב (*), רואים שאפשר לאפס את הרכיב הניצב $w^*_\perp$ בלי להגדיל את ערכה של פונקציית המטרה, כלומר אפשר תמיד למצוא פתרון שיש לו רק רכיב מקביל $w^*_\parallel$, ולכן הפתרון אכן מקיים את התנאי $w\in \text{span}\left(\{x_i\}_{i=1}^n\right)$.

מסווג לא לינארי

עבור הרבה בעיות מעשיות, משטח החלטה לינארי אינו מפריד בצורה טובה בין המחלקות, ונצפה שמשטח החלטה לא לינארי ישיג ביצועים יותר טובים. דרך אחת להשיג משטח החלטה שכזה היא באמצעות טרנספורמציה לא לינארית ממרחב הקלט למרחב חדש: $x\to \Phi(x)=[\phi_1(x),...,\phi_m(x)]^T$, כאשר $\phi_i(x):\mathbb{R}^d\to\mathbb{R}$. אימון של מסווג לינארי במרחב החדש, שיראה מהצורה $\text{sign}\left(w^T\Phi(x)+b\right)$, שקול למסווג לא-לינארי במרחב המקורי.

דוגמה: נתון קלט $x\in\mathbb{R}^2$. הדוגמאות עם $y=+1$ נמצאות בתוך מעגל עם רדיוס R, והדוגמאות עם $y=-1$ נמצאות מחוץ למעגל. במרחב x אין מפריד לינארי, אך הטרנספורמציה $\Phi(x)=[x_1^2,x_2^2]$ מאפשרת לסווג את הדוגמאות ללא שגיאה עם המסווג הלינארי $\text{sign}\left(-(x_1^2+x_2^2-R^2)\right)$, כלומר $w=\begin{bmatrix}-1\\-1\end{bmatrix},\ b=R^2$.



דוגמה נוספת: טרנספורמציה למרחב הפולינומים ממעלה עד 2, בקואורדינטות של הווקטור x:

$$x\in\mathbb{R}^2$$
$$\Phi(x)=[1,x_1,x_2,\sqrt{2}x_1x_2,x_1^2,x_2^2]^T\in\mathbb{R}^6$$

בעיה: טרנספורמציה למרחב חדש יכולה להיות יקרה חישובית אם המימד של המרחב החדש גבוה מאוד.

פתרון: אם המסווג תלוי רק במכפלות פנימיות בין ווקטורי הקלט, אין צורך לחשב את $\Phi(x)$, אלא רק את המכפלות הפנימיות במרחב החדש, $\Phi(x_i)^T\Phi(x_j)$:

$$\hat{y}=f(\Phi(x))=sign\left(\sum_{i=1}^n\tilde{\alpha}_i\Phi(x_i)^T\Phi(x)+\tilde{b}\right)$$

לצורך זה נציג את פונקציית הגרעין.

פונקציית גרעין (Kernel)

נגדיר פונקציית גרעין על קבוצה X (תת-קבוצה של $\mathbb{R}^d$) כפונקציה $K: X \times X \to R$ שהינה:

א. סימטרית: $K(x,z) = K(z,x)$

ב. לכל קבוצה סופית של נקודות $\{x_1, x_2, ..., x_n\}$, המטריצה $\left\{K_{kl}\right\} = K\left(x_k, x_l\right)$ היא מטריצה אי שלילית מוגדרת (PSD).

אזי, תחת תנאים טכניים סבירים, קיים מרחב $\Phi(x)$ כך שפונקציית הגרעין הינה מכפלה פנימית מהצורה $K(x_k, x_l) = \Phi^T(x_k)\Phi(x_l)$.

תכונה זו מאפשרת לנו לחשב ישירות את $K(x_i, x_j)$, במקום את $\Phi(x_i)^T\Phi(x_j)$ שיכול להיות יקר לחישוב.

דוגמאות לפונקציות גרעין:

א. גרעין גאוסי: $K(x,z) = \exp(-\left\|x-z\right\|^2 / c)$.

הפונקציות $K(x,x_l)$: גאוסיאנים, כאשר c פרמטר שיש לקבוע ידנית.

ב. גרעין פולינומיאלי: $K(x,z) = (1 + x^T z)^p$, כאשר $p \geq 1$ פרמטר שיש לקבוע ידנית. הפונקציות $K(x,x_l)$: פולינומים רבי-משתנים מסדר עד p באיברי הווקטורים x, x_l.

<u>שאלה 1</u>

א. עבור $x \in \mathbb{R}^2$, $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ נגדיר את וקטור המאפיינים הבא:

$\phi(x) = \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix} \in \mathbb{R}^3$. הוכיחו כי עבור בחירת הגרעין $k(y,z) = (y^T z)^2$,

מתקיים: $k(y,z) = \left\langle \phi(y), \phi(z) \right\rangle$ כאשר $\left\langle u, v \right\rangle = u^T v$.

(כלומר הוכיחו כי הגרעין הנ״ל מתאים למכפלה הפנימית של וקטור המאפיינים).

ב. במקרה הכללי יש לנו קלט במימד d, כלומר $x\in\mathbb{R}^d$, ורכיבי הווקטור $\Phi(x)$ במרחב החדש הם $\phi_m(x)=c_m x_1^{\alpha_1}x_2^{\alpha_2}\cdot\ldots\cdot x_d^{\alpha_d}$, כך ש α_i מספר שלם אי-שלילי ו- $\sum_{i=1}^{d}\alpha_i = p$, כלומר כל רכיב הוא מכפלה של p מרכיבי x, כאשר יתכנו חזרות. לדוגמא, בסעיף א' $d=2, p=2$.

ידוע כי עבור בחירת הגרעין $k(y,z)=\left\langle y,z\right\rangle^p$, מתקיים $k(y,z)=\left\langle \phi(y),\phi(z)\right\rangle$. מהו המימד של $\phi(x)$ עבור d,p כלליים? לאור זאת, מהי החשיבות של שימוש בפונקציית גרעין?

<u>פתרון</u>

א. נחשב את המכפלה הפנימית:

$$\Phi(y)^T\Phi(z)=[y_1^2,\sqrt{2}y_1y_2,y_2^2]\cdot\begin{bmatrix} z_1^2\\ \sqrt{2}z_1z_2\\ z_2^2\end{bmatrix}=y_1^2z_1^2+2y_1y_2z_1z_2+y_2^2z_2^2=$$

$$\left(y_1z_1+y_2z_2\right)^2=(y^Tz)^2=K(y,z)$$

ב. מכיוון שסכום החזקות בכל קואורדינטה הוא p , $\sum_{i=1}^{d}\alpha_i = p$, כל קואורדינטה תיקבע ע"י חלוקה של p חזקות בין d רכיבים. זה שקול לחלוקה של p כדורים בין d תאים – מספר הכדורים בתא ה-i הוא α_i, עבור $i=1,...,d$, תחת האילוץ $0\le\alpha_i\le p$. מצב זה מודגם בציור הבא:

$$|\,\alpha_1\,|\,\alpha_2\,|\ldots|\,\alpha_i\,|\ldots|\,\alpha_d\,|$$

את מספר האפשרויות לחלק p כדורים בין d תאים אפשר לחשב באופן הבא: d תאים מוגדרים ע"י $d+1$ מחיצות, אבל שתי המחיצות הקיצוניות לא משפיעות על החלוקה, לכן אפשר להתייחס ל $d-1$ מחיצות ועוד p כדורים. יש $(d+p-1)!$ אפשרויות לסדר את התאים והכדורים בשורה, אבל צריך לחלק במספר האפשרויות של הסידור הפנימי בין המחיצות לבין עצמן, ובין הכדורים לבין עצמם. בסה"כ נקבל שמספר האפשרויות הוא $\binom{d+p-1}{d-1}=\frac{(d+p-1)!}{p!(d-1)!}$, וזהו המימד של $\Phi(x)$. מכיוון שמספר זה יכול להיות מאוד גדול, השימוש בפונקציית גרעין יכול להוריד את החישוביות באופן משמעותי.