פרק 5: שיטות ליניאריות לסיווג

מקור: HTF, פרק 3.

מקור: DHS: 5.11, HTF: 4.5.2, 12.2-12.3 .

6.1 פונקציות תיוג ליניאריות

כזכור, בבעיית הסיווג עלינו למצוא מסווג $f : X \to \Omega$ אשר משייך כל קלט $x \in X$ לאחת מ- C המחלקות $\Omega = \{1,\ldots,C\}$. בגישה הפרמטרית, אנו מגדירים משפחה פרמטרית של מסווגים $f(x,\theta)$, ומתמקדים בכוונון (לימוד) וקטור הפרמטרים θ.

מבנה מקובל למסווג עושה שימוש בפונקציות הבחנה פרמטריות. עבור כל מחלקה $j \in \Omega$ נגדיר פונקצית הבחנה (Discrimination Function) $g_j(x,\theta)$ המקבלת ערכים ממשיים. הסיווג יתבצע לפי "החזק מנצח":

$$\hat{j}(x) \equiv f(x,\theta) = \arg\max_{j\in\Omega} g_j(x,\theta)$$

בפרק זה נתמקד בפונקציות הבחנה שהן ליניאריות-בפרמטרים. עבור סט נתון של פונקציות בסיס $\{\phi_m(x)\}_{m=1}^M$ נגדיר

$$g_j(x,\theta) = \sum_{m=1}^{M} \theta_{jm}\phi_m(x), \qquad j \in \Omega$$

מקרה פרטי הינו פונקצית ההבחנה הליניארית (בקלט x): $g_j(x,\theta) = \theta_{j0} + \sum_{i=1}^{d} \theta_{ji} x_i$.

עבור שתי מחלקות, משטח ההפרדה המתקבל במקרה זה הוא על-מישור במרחב x. השימוש בפונקציות בסיס מאפשר לקבל משטחי הפרדה מורכבים בהרבה.

6.2 סיווג באמצעות רגרסיה של פונקצית אינדיקטור

נזכיר כי נתונה סדרת לימוד $D=\{x_k, y_k\}_{k=1}^n$, כאשר $y_k \in \Omega$.

מטרתנו פה תהיה למצוא פונקציות הבחנה אשר מקיימות (בקרוב):

$$g_j(x_k,\theta)=\sum_{m=1}^{M}\theta_{jm}\phi_m(x_k)=\begin{cases}1 : & y_k = j\\ 0 : & y_k \neq j\end{cases}$$

אקויולנטית, נשאף לקיים

$$\sum_{m=1}^{M}\theta_{jm}\phi_m(x_k)\approx \bar{y}_{kj}$$

כאשר

$$\bar{y}_{kj}=\begin{cases}1 : & y_k = j\\ 0 : & y_k \neq j\end{cases}.$$

למציאת הפרמטרים המתאימים, נגדיר עבור כל $j\in\Omega$ את השגיאה הריבועית הבאה

$$E_j(\theta_j)=\sum_{k=1}^{n}\Big(\bar{y}_{kj}-\sum_{m=1}^{M}\theta_{jm}\phi_m(x_k)\Big)^2.$$

אקויולנטית,

$$E_j(\theta_j)=\sum_{k=1}^{n}\Big(\bar{y}_{kj}-\theta_j^T\phi(x_k)\Big)^2$$

כאשר $\phi(x)^T=(\phi_1(x),\ldots,\phi_M(x))$, $\theta_j^T=(\theta_{1j},\ldots,\theta_{j\,1})$. כזכור, וקטור הפרמטרים המביא למינימום שגיאה זו הינו

$$\theta_j^{(opt)}=\left(\sum_{k=1}^{n}\phi(x_k)\phi(x_k)^T\right)^{-1}\sum_{k=1}^{n}\phi(x_k)\bar{y}_{kj}\,,\quad j\in\Omega.$$

- שיטה זו הינה פשוטה יחסית להגדרה וחישוב. עבור סיווג בינארי ($C=2$) היא נותנת תוצאות סבירות (אם כי לא מיטביות), אולם עבור $C>2$ עלולים להתגלות סטיות משמעותיות ממשטחי הפרדה "הגיוניים". הסיבה הבסיסית לכך היא שקריטריון השגיאה הריבועית אינו המדד הטבעי עבור בעיית הסיווג. בפרק הבא (הפרדה ליניארית אופטימאלית) נציע מדד טבעי יותר.

6.3 שיטות לינאריות מבוססות-פילוג

נתאר בקצרה שתי שיטות סיווג לינאריות המבוססות על קירוב פונקציות הפילוג ההסתברותי של הבעייה – דהיינו הפילוג המותנה של הקלט בכל מחלקה $p(x\,|\,\omega_j)$, או פונקציית הפילוג הפוסטריורי $p(\omega_j\,|\,x)$.

<u>א. "אנליזת הבחנה לינארית" LDA</u> (Linear Discriminant Analysis) :

זו למעשה שיטה הסיווג הבייסיאנית-אמפירית, תחת הנחה של פילוג גאוסי בעל קוואריאנס משותף. נתאר פה את השיטה בשילוב וקטור מאפיינים $\phi(x)=(\phi_1(x),\ldots,\phi_M(x))^T$. נניח כי לווקטור המאפיינים פילוג גאוסי : $p(\phi(x)\,|\,\omega_j)\approx N(\mu_j,\Sigma)$, עם מטריצת קוואריאנס משותפת Σ . חוק ההחלטה הבייסיאני האופטימאלי (MAP) הינו :

$$j_{MAP}(x)=\arg\max_{j\in\Omega} g_j(x)$$

כאשר

$$\begin{aligned} g_j(x) &= P(\omega_j)\,p(\phi(x)\,|\,\omega_j) \\ &= P(\omega_j)\frac{1}{\sqrt{(2\pi)^M\,|\Sigma|}}\exp[-\tfrac{1}{2}(\phi(x)-\mu_j)^T\Sigma^{-1}(\phi(x)-\mu_j)] \end{aligned}$$

אקויולנטית, לאחר הוצאת לוגריתם וביטול האיבר הריבועי, מתקבלת פונקציית דיסקרימינציה ליניארית :

$$\tilde{g}_j(x)=\log P(\omega_j)-\tfrac{1}{2}\mu_j{}^T\mu_j+\mu_j{}^T\Sigma^{-1}\phi(x)$$

נותר להעריך את הפרמטרים $P(\omega_j)$, μ_j , Σ מתוך סדרת הלימוד $D=\{x_k,y_k\}_{k=1}^n$. נשתמש במשערכים הסטנדרטיים לפילוג הגאוסי :

$$\hat{P}(\omega_j)=\frac{n_j}{n}\equiv\frac{1}{n}\sum_{k=1}^n I\{y_k=\omega_j\}$$

$$\hat{\mu}_j=\frac{1}{n_j}\sum_{k\,:\,y_k=j}x_k$$

$$\hat{\Sigma}=\frac{1}{n-C}\sum_{j=1}^C\sum_{k\,:\,y_k=j}(x_k-\hat{\mu}_j)(x_k-\hat{\mu}_j)^T$$

ב. <u>רגרסיה לוגיסטית</u> (Logistic Regression)

הרעיון לקרב ישירות את פונקצית הפילוג הפוסטריאורי $P(y=j\,|\,x)$ על ידי אקספוננט ליניארי-בפרמטרים. דהיינו:

$$\hat{P}_\theta(y=j\,|\,x)=\frac{1}{c_\theta(x)}\exp(\sum_{m=1}^{M}\theta_{jm}\phi_m(x))\equiv\frac{1}{c(x)}\exp(\theta_j^T\phi(x))$$

כאשר

$$c_\theta(x)=\sum_{j\in\Omega}\exp(\theta_j^T\phi(x))\,.$$

במקרה זה $\hat{P}(y=j\,|\,x)$ עצמה (עם פרמטרים מתאימים) תשמש כפונקציית ההבחנה, דהיינו $g_j(x)=\hat{P}(y=j\,|\,x)$. אקויולנטית, לאחר לקיחת הלוגריתם וביטול מקדם הנירמול (שאינו תלוי ב- j), נקבל פונקציית הבחנה ליניארית: $\tilde{g}_j(x)=\theta_j^T\phi(x)$.

את וקטור הפרמטרים $\theta=(\theta_{jm})$ ניתן להעריך באמצעות משערך הסבירות המירבית (MLE). פונקציית הסבירות במקרה זה, בהנחה של דוגמאות בלתי תלויות, הינה:

$$L_n(\theta)=\prod_{k=1}^{n}\hat{P}_\theta(y_k\,|\,x_k)$$

הפרמטר האופטימאלי מתקבל כנקודת המכסימום של פונקצית הסבירות. בעיית אופטימיזציה זו אינה ניתנת לפתרון אנליטי, ויש להשתמש באלגוריתמים נומריים לצורך זה.

6.4 מבוא ל- SVM

אנו ממשיכים בפרק זה את הדיון בשיטות ליניאריות לסיווג. כזכור, מדד הביצועים הבסיסי בבעיית הסיווג הינו הקטנת הסתברות השגיאה. בהתאם לכך, קריטריון טבעי בשלב הלימוד הינו (מינימיזציה של) <u>מספר</u> הטעויות בסדרת הלימוד, דהיינו מספר הדוגמאות שאינן מסווגות נכון (בהתאם לתווית שלהם). לקריטריון זה שני חסרונות:

1. הוא אינו מגדיר חד-משמעית את משטחי ההפרדה).
2. בעיית האופטימיזציה המתקבלת קשה לפתרון (פרט למקרה הפרטי של שתי מחלקות הניתנות להפרדה מלאה).

כדי להתגבר על חסרונות אלה נתאר בפרק זה מדד של הפרדה אופטימאלית במובן של "שוליים מירביים", ונתעכב בקצרה על תכונותיו. מדד זה עומד בבסיסה של שיטת הסיווג הידועה בשם Support Vector Machine (SVM), שהיא בין שיטות הסיווג המתקדמות ביותר מבחינת ביצועיה. מסווגים מסוג זה נקראים גם Maximal Margin Classifiers .

הרעיון: היינו רוצים שיתקיים $\mathrm{sign}\{w'x+b\}=y$ עבור רוב הזוגות (x,y) כאשר $x\in R^d$ וכן $y\in\{-1,+1\}$ כאשר רוב מוגדר תחת פילוג קבוע שאינו ידוע על הזוגות $(x,y)\sim D$. זו דרישה קשיחה שכן איננו לוקחים בחשבון את הערך הממשי של $w'x+b$ שיתכן שקרוב לערכו של y ויתכן שרחוק. תנאי יותר "רך" הוא הדרישה כי $w'x+b$ יהיה שונה משמעותית מ $-y$ דהיינו בעל סימן מתאים וכן בעל ערך מוחלט גבוה ככל האפשר.

6.5 דוגמאות ניתנות להפרדה

בפרק זה נתמקד בבעיית הסיווג הבינארית, דהיינו סיווג לשתי מחלקות. נציין את שתי מחלקות אלה בערכים המספריים $+1,-1$. כרגיל נתונה סדרת הלימוד דוגמאות $\{x_k,y_k\}_{k=1}^n$, כאשר הפעם $y_k\in\{-1,+1\}$.

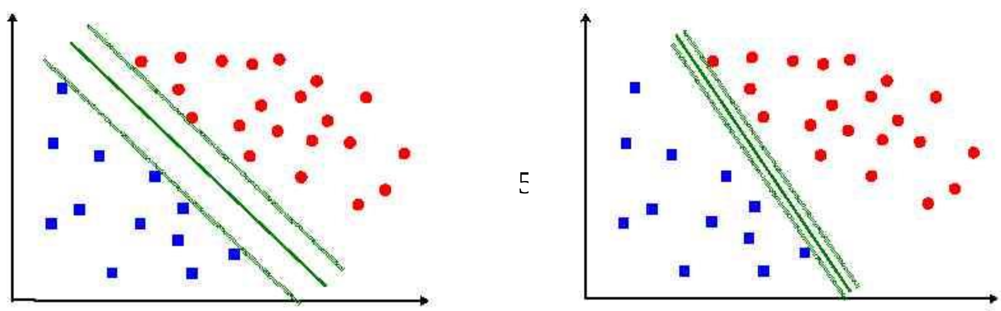
הגדרה : סדרת הדוגמאות $\{x_k,y_k\}_{k=1}^n$ ניתנת להפרדה ליניארית אם קיים על-מישור במרחב x אשר מפריד באופן מלא בין הדוגמאות בהתאם לסיווגן.

נזכיר כי על-מישור מוגדר על ידי השוויון $\sum_{i=1}^d w_i x^{(i)}+b=0$, או בקיצור $w'x+b=0$ (כאשר $w'=w^T$) . דרישת ההפרדה תתקיים באם :

$$\mathrm{sign}\{w'x_k+b\}=y_k\ , \quad k=1,\ldots,n$$

נניח לעת עתה כי הדוגמאות ניתנות להפרדה ליניארית באמצעות משטח הפרדה ליניארי (על-מישור) מתאים. במקרה זה, יהיו באופן כללי אינסוף משטחי הפרדה ליניאריים אשר מקיימים זאת (ראה ציור). במי מהם נבחר?

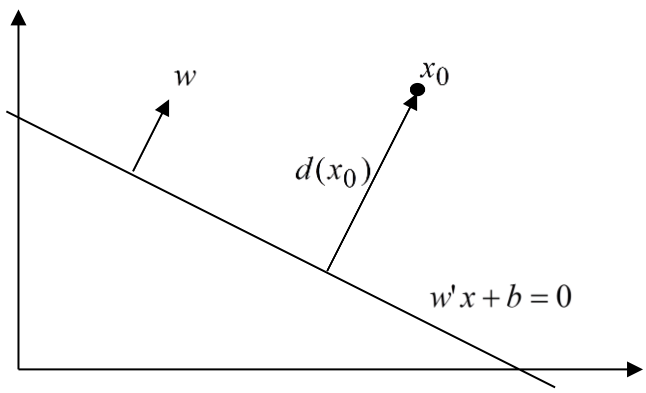
<u>הצעה</u> : נבחר במשטח ההפרדה אשר נותן את "מרווח הביטחון" הגדול ביותר :



<u>מעט גיאומטריה</u>: פונקציית המרחק בין נקודה $x_0 \in \Re^n$ למשטח $w'x+b=0$ הינה

$$d(x_0)=\frac{w'x_0+b}{\|w\|}\ .$$

הערך המוחלט של גודל זה הינו המרחק האאוקלידי בין הנקודה למשטח. בנוסף, ל- $d(x_0)$ סימן חיובי אם x_0 בכיוון הוקטור w (יחסית למשטח), וסימן שלילי אחרת.



<u>המרחק לדוגמאות</u>: ברצוננו להגדיל את המרחק בין המישור המפריד לדוגמא x_k, וזאת כמובן בצד הנכון של המישור. מרחק זה יהיה לכן $y_k d(x_k)$, כאשר y_k דואג לסימן המתאים (חיובי כאשר הדוגמא בצד הרצוי, ושלילי אחרת). לפיכך, "מרווח הביטחון" נתון על ידי

$$\frac{\min_{1\le k\le n}\{y_k(w'x_k+b)\}}{\|w\|}$$

ומטרתנו להביא מרווח זה למכסימום:

$$\max_{w,b}\left\{\frac{\min_{1\le k\le n}\{y_k(w'x_k+b)\}}{\|w\|}\right\}$$

<u>נירמול</u>: כל וקטור פרמטרים (w,b) ניתן לנירמול בקבוע חיובי ללא שינוי המישור המפריד. נבחר לנרמל את הפרמטרים כך שיתקיים $\min_{1\le k\le n}\{y_k(w'x_k+b)\}=1$. נרמול זה מוביל לבעיית האופטימיזציה הבאה:

$$\max_{w,b}\left\{\frac{1}{\|w\|}\right\},\quad \text{subject to}\quad \min_{1\le k\le n}\{y_k(w'x_k+b\}=1$$

לבסוף, בעיה זו ניתנת לכתיבה כך:

$$\boxed{\begin{array}{l}\min_{w,b}\tfrac{1}{2}\|w\|^2 \\ \text{s.t. : } \ y_k(w'x_k+b)\ge 1,\quad k=1,2,\ldots,n\end{array}}$$

הגענו לבעיה של מינימיזציית מחיר ריבועי, כפוף לאילוצי אי-שוויון ליניאריים. זוהי בעיית תכנות ריבועי (קונווקסי), שעבורה קיימים אלגוריתמים נומריים יעילים למציאת המינימום (הגלובלי). בעיה זו נקראת הבעייה הראשונית (פרימאלית).

<u>תכונות הפתרון</u>:

משפט: הוקטור w האופטימאלי ניתן לביטוי באופן הבא:

$$w=\sum_{k=1}^{n}\alpha_k y_k x_k$$

כאשר $\alpha_k\ge 0$, ו- $\alpha_k\ne 0$ רק אם $y_k(w'x_k+b)=1$. כמו כן, $\sum_{k=1}^{n}\alpha_k y_k=0$.

הוכחה (*): נעזר התוצאה מתורת האופטימיזציה לגבי בעיה עם אילוצי אי-שוויון. התנאים ההכרחיים מסדר ראשון לקיום מינימום מקומי לבעיה עם אילוצי אי-שוויון הם תנאי Kuhn-Tucker (KT), שהם הכללה של "כופלי לגרנז'" המוכרים מבעיות עם אילוצי שוויון. בבעיה שלנו, תנאים אלה גם מספיקים עקב הקונווקסיות. נגדיר ראשית את פונקצית הלגרנז'יאן:

$$L(w,b,\alpha)=\tfrac{1}{2}\|w\|^2-\sum_{k=1}^{n}\alpha_k(y_k(w'x_k+b)-1)$$

תנאי KT לקיום אופטימום בנקודה (w,b) הינם קיום קבועים <u>אי שליליים</u>, $\alpha_k\ge 0$, כך ש:

א. $\{\alpha_k\}$ מביאים את $L(w,b,\alpha)$ למכסימום.

ב. (w,b) מביאים את $L(w,b,\alpha)$ למינימום.

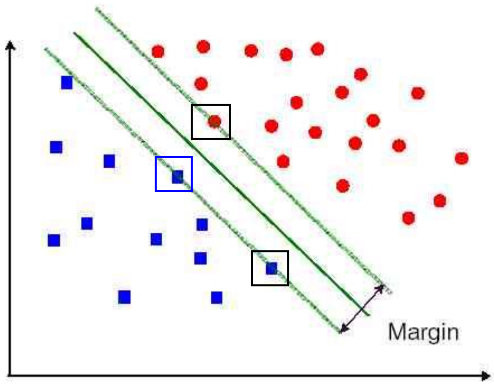
ניתן לראות כי תנאי (א) שקול ל: $\alpha_k \neq 0$ רק אם $y_k(w'x_k+b)-1=0$.

לגבי תנאי (ב), מינימיזציה של הפונקציה הקונווקסית $L(w,b,\alpha)$ שקולה ל:

$$\frac{d}{dw}L(w,b,\alpha)=0, \quad \frac{d}{db}L(w,b,\alpha)=0$$

הטענה נובעת מיידית על ידי חישוב הנגזרות. □

וקטורי הקלט x_k שעבורם מתקיים $y_k(w'x_k+b)=1$ נקראים "וקטורי התמיכה" (Support Vectors). מהטענה לעיל נובעת התכונה החשובה כי הפתרון האופטימאלי ל-w הוא צרוף ליניארי של וקטורי תמיכה בלבד (שמספרם קטן יחסית באופן טיפוסי).



6.6 הבעייה הדואלית

בסעיף זה נרשום נתאר בעיית אופטימיזציה שהיא הדואלית לבעיה הפרימאלית בסעיף הקודם. בתורת האופטימיזציה הקונבקסית, הבעיה הדואלית הוא בעיה שמתוך פתרונה ניתן לגזור גם את פתרון הבעייה הדואלית. במקרה שלנו, הבעייה הדואלית תתן באופן ישיר את המקדמים $\{\alpha_k\}$. על יתרונות הבעיה הדואלית נעמוד בהמשך.

מטרתנו לחשב את הקבועים $\{\alpha_k\}$ מתוך בעיית האופטימיזציה שהופיעה בהוכחת הטענה הקודמת:

$$\min_{\alpha} L(w,b,\alpha)$$

כמו כן נציב $w=\sum_{k=1}^n \alpha_k y_k x_k$, ונוסיף את האילוצים $\alpha_i \geq 0$, $\sum_{k=1}^n \alpha_k y_k = 0$. נקבל את בעיית האופטימיזציה הבאה - הבעיה הדואלית:

$$\boxed{\begin{array}{l}\max\limits_{\alpha} \ \sum\limits_{k=1}^{n}\alpha_k - \dfrac{1}{2}\sum\limits_{k,l=1}^{n}\alpha_k\alpha_k y_k y_k\langle x_k, x_l\rangle \\ \text{s.t. :} \quad \alpha_k \ge 0, \quad k=1,2,\dots,n \\ \qquad\qquad \sum_{k=1}^{n}\alpha_k y_k = 0\end{array}}$$

בבעיה זו :

א. הוקטורים $\{x_k\}$ מופיעים רק באמצעות המכפלות הפנימיות $\langle x_k, x_l\rangle$ – אבחנה שתהיה רבת ערך בהמשך.

ב. מימד הבעיה (מספר המשתנים) הוא כמספר הדוגמאות.

לאחר חישוב הווקטור α, ניתן לחשב את $w=\sum_{k=1}^{n}\alpha_k y_k x_k$, ואת b מתוך השוויון :

$$\alpha_k \ne 0 \quad \Rightarrow \quad y_k(w'x_k+b)=1 .$$

נשים לב, בנוסף, כי $w'x=\sum_{k=1}^{n}\alpha_k y_k\langle x_k, x\rangle$, כאשר הסכום הוא אפקטיבית על וקטורי תמיכה בלבד.

נשווה בין שני הניסוחים : הניסוח הראשוני, פרימלי, כולל d משתנים ו- n אילוצים לינארים מורכבים, הקלט מיוצג ע"י מטריצה בגודל $n\times d$. הניסוח הדואלי כולל n משתנים ו- n אילוצים לינארים פשוטים , הקלט מיוצג ע"י מטריצה בגודל $n\times n$. מה עדיף?

6.7 המקרה הכללי – דוגמאות שאינן ניתנות להפרדה

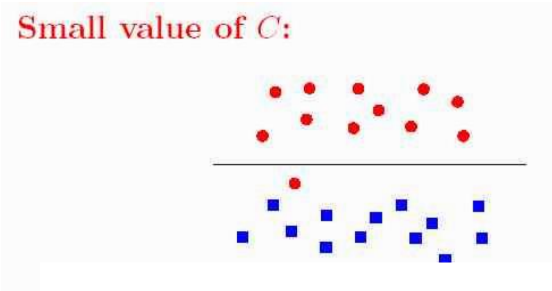
במקרה הכללי איננו יכולים לצפות כי הדוגמאות תהינה תמיד ניתנות להפרדה ליניארית. במקרים אלה לבעיה שתיארנו קודם לא יהיה פיתרון. על מנת לקבל בעיה ברת משמעות, נחליש את אילוצי ההפרדה הקשיחים $y_k(w'x_k+b)\ge 1$, על ידי החלפתם בדרישה :

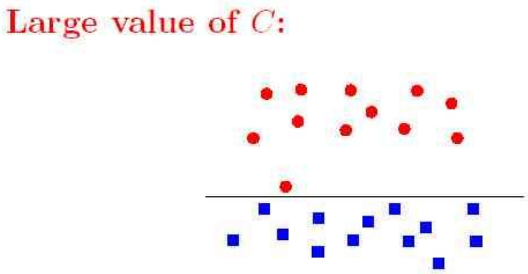
$$y_k(w'x_k+b)\ge 1-\xi_k, \quad \xi_k \ge 0, \qquad k=1,2,\dots,n$$

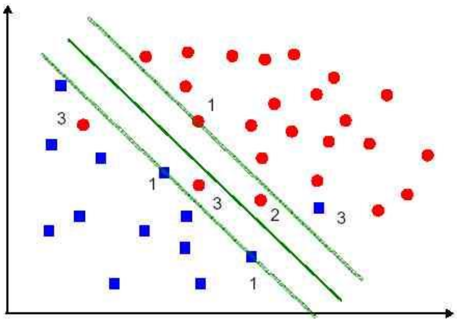
המשתנים החדשים נקראים "משתני מרווח" (slack variables), ומטרתנו כמובן שיהיו קטנים ככל האפשר. לשם כך נוסיף אותם לבעיית האופטימיזציה הפרימאלית, שתהפוך להיות :

$$\min_{w,b,\xi} \tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i$$

הקבוע C קובע את החשיבות היחסית של גודל השוליים לעומת גודל החריגה המותרת.







וקטורי התמיכה יכללו עתה גם את הנקודות $\{x_i\}$ אשר אינן מסווגות נכון על ידי המשטח.

הבעיה הדואלית נשארת ללא שינוי, פרט להחלפת האילוץ $\alpha_k \geq 0$ באילוץ $0 \leq \alpha_k \leq C$.

את b ניתן לחשב מתוך: $y_k(w'x_k + b) = 1 \quad \Rightarrow \quad 0 < \alpha_k < C$.

הערה: הקריטריון הנ"ל מתחשב בדוגמאות שאינו מסווגות נכונה על ידי מדידת מרחקן ממשטח ההפרדה. כתחליף לכך, ניתן היה לחשוב על קריטריון אשר פשוט סופר את מספר הדוגמאות האלה. קריטריון כזה ייתקל בקשיים חישוביים עקב המשפט הבא (שלא נוכיח): עבור מדגם שאינו פריד ליניארית לא ניתן למצוא אלגוריתם יעיל (פולינומי במימד הקלט d) למציאת מסווג בעל מספר שגיאות מזערי.

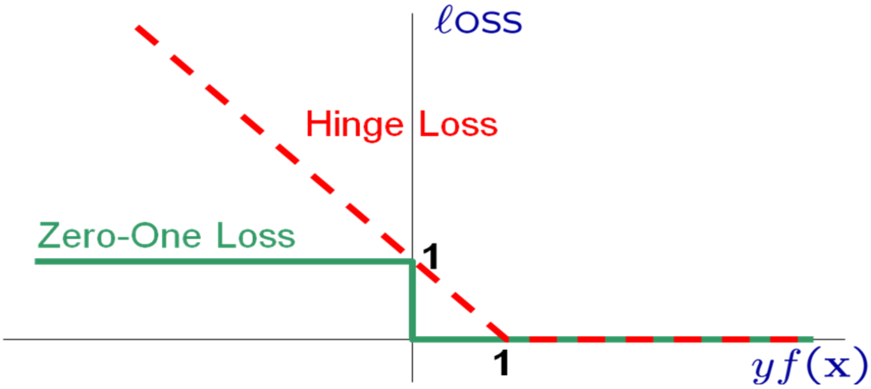
הערה: ניתן לרשום את בעיית הלמידה באופן שקול גם כ-

$$\min_{w,b,\xi} \tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \max\left\{0, 1 - y_i\left(w'x_i\right)\right\}$$

האיבר הימני הוא סכום של חסמים על שגיאה התיוג, חסמים אילו נקראים Hinge loss.

האיבר השמאלי נקרא רגולריזציה. מטרתו לאפשר לאלגוריתם הלמידה להכליל גם כאשר מס הפרמטרים הוא רב מאוד.

ראו תמונה



6.8 שילוב פונקציות בסיס

שיפור משמעותי של יכולת ההפרדה בעזרת משטחים ליניאריים יתקבל על ידי שילוב פונקציות בסיס, או מאפיינים, במימד גבוה ממימד הקלט x. דהיינו, נחליף את $x=(x_1,\dots,x_d)^T$ בווקטור המאפיינים $\phi(x)=(\phi_1(x),\dots,\phi_M(x))^T$, כאשר $M>>d$. את ההפרדה הליניארית נבצע במרחב המאפיינים, ולא במרחב הקלט.

מטרתנו לבצע סיווג בעזרת הפונקציה

$$\hat{f}(x)=sign(w'\phi(x))$$

עם מקדמים מתאימים w. "מרווח הביטחון" אותו נביא למכסימום יימדד עתה במרחב המאפיינים:

$$d_{w,b}(x_0)=\frac{w'\phi(x_0)}{\|w\|}$$

ניתן עתה לחזור על כל הפיתוחים שלעיל, עם ההחלפות הבאות:

א. x מוחלף בוקטור $\phi(x)$

ב. $\langle x,z\rangle$ מוחלף ב- $K(x,z)\equiv\langle\phi(x),\phi(z)\rangle$.

<u>הבעיה הדואלית</u> המתקבלת:

$$\max_{\alpha} \ \sum_{k=1}^{n} \alpha_k - \frac{1}{2} \sum_{k,l=1}^{n} \alpha_k \alpha_l y_k y_l K(x_k, x_l)$$

$$\text{s.t.}: \ \ 0 \le \alpha_k \le C, \quad k = 1, 2, \ldots, n$$

$$\sum\nolimits_{k=1}^{n} \alpha_k y_k = 0$$

לאחר מציאת המקדמים (α_k) ניתן לחשב את $w = \sum\nolimits_{k=1}^{n} \alpha_k y_k \phi(x_k)$, ומכאן את פונקצית המסווג:

$$y = sign(w' \underline{\phi}(x)) = \underline{sign\left(\sum\nolimits_{k=1}^{n} \alpha_k y_k K(x_k, x)\right)}$$

נציין כי:

- מימד הבעיה הדואלית לא השתנה, למרות הגדלת מרחב המאפיינים.
- וקטור המקדמים (α_k) צפוי להיות דליל: $\alpha_k \neq 0$ רק אם $\phi(x_k)$ הוא וקטור תמיכה, או אם x_k אינו מסווג נכון.

6.9 שילוב פונקציות גרעין (The Kernel Trick)

נוסיף עתה גורם נוסף (ואחרון) אשר מאפשר ליישם את ההפרדה האופטימאלית הנ"ל במרחב מאפיינים במימד גדול מאוד, ואף אינסופי.

נשים לב כי בבעיית האופטימיזציה הדואלית של הסעיף האחרון, לא מופיעות פונקציות הבסיס באופן ישיר כי אם באמצעות המכפלות הפנימיות $K(x,z) \square \langle \phi(x), \phi(z) \rangle$. הדבר נכון גם לגבי המסווג האופטימאלי המתקבל. לפיכך, באם נוכל לחשב את המכפלה הפנימית באופן יעיל, מימד $\phi(x)$ לא ישפיע על סיבוכיות החישוב.

הרעיון הוא כי עבור אוספים מסוימים של פונקציות בסיס $\{\phi_m(x)\}$, למכפלה הפנימית $\langle \phi(x), \phi(z) \rangle$ יש צורה אנליטית סגורה, כך שפונקציה $K(x,z)$ ניתנת לחישוב ישיר.

<u>הבסיס התיאורטי</u>:

פונקציית גרעין $K(x,z)$ על המרחב $X = \mathbb{R}^d$ היא פונקציה רציפה $K : X \times X \to \mathbb{R}$ שהינה:

א. סימטרית: $K(x,z)=K(z,x)$

ב. חיובית מוגדרת: המטריצה $\bar{K}=\{K(x_k,x_l)\}_{k,l=1}^{n}$ הינה אי-שלילית מוגדרת ($\bar{K}\geq 0$ לכל אוסף סופי של נקודות $(x_1,\cdots x_n)$).

משפט מרסר (Mercer 1909): כפוף לתנאים טכניים מסויימים, פונקציית גרעין ניתנת לביטוי באמצעות הסכום הבא:

$$K(x,z)=\sum_{m=1}^{\eta}\phi_m(x)\phi_m(z)$$

כאשר η עשוי להיות אינסופי, ו- $\{\phi_m(x)\}$ פונקציות בסיס מתאימות.

מכאן כי כל פונקציית גרעין $K(x,z)$ מגדירה מכפלה פנימית בין פונקציות בסיס.

מהן פונקציות הבסיס המתאימות לפונקצית גרעין נתונה? למעשה השאלה החשובה היא: מהו המרחב הנפרש על ידי פונקציות בסיס אלה. ניתן לוודא כי זהו המרחב הנפרש על ידי אוסף הפונקציות $\{K(x,z_0)\ :\ z_0\in X\}$ (מדוע?).

פונקציות הגרעין הנפוצות כוללות את הבאות:

א. גרעין גאוסי: $K_\lambda(x,z)=\exp(-\|x-z\|/\lambda)$.

הפונקציות $K(x,z_0)$ הן גאוסיאנים רדיאליים בעלי רוחב נתון. מרחב פונקציות הבסיס הוא המרחב הנפרש על ידי כל גאוסיאנים אלה (זהו מרחב אינסוף-מימדי).

המסווג המתקבל במקרה זה יהיה מהצורה:

$$y=\text{sign}\left(\sum\nolimits_{k=1}^{n}\alpha_k y_k\exp(-\|x-x_i\|^2/\lambda)\right)$$

ב. גרעין פולינומיאלי: $K(x,z)=(1+x^Tz)^L$ (כאשר $L\geq 1$).

הפונקציות $K(x,z_0)$: פולינומים רבי-משתנים (multinomials) מסדר L ברכיבי הווקטור x.

לדוגמא, עבור $x\in\mathbb{R}$, $L=2$ נקבל:

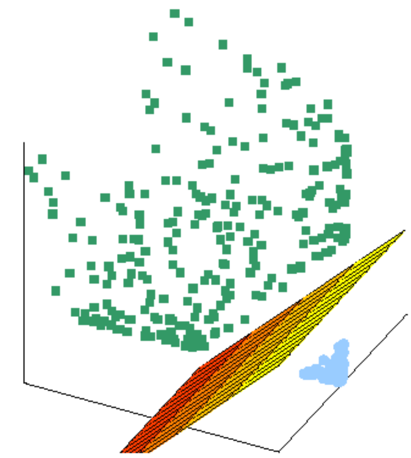
$$K(x,z)=(1+xz)^2=1+2xz+x^2z^2=\left\langle\begin{pmatrix}1\\ \sqrt{2}x\\ x^2\end{pmatrix},\begin{pmatrix}1\\ \sqrt{2}z\\ z^2\end{pmatrix}\right\rangle$$

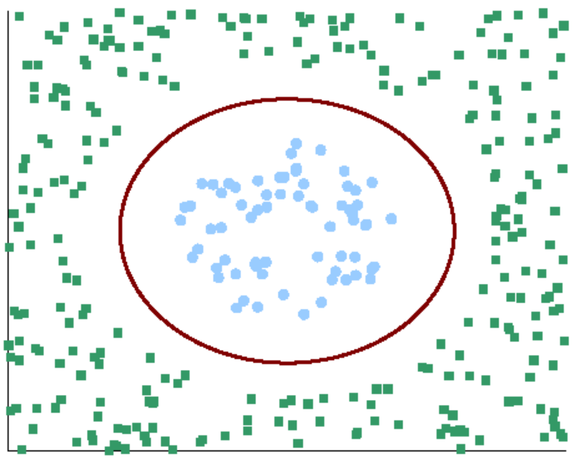
צורת המסווג המתקבל תהיה:

$$y=\operatorname{sign}\left(\sum_{k=1}^{n}\alpha_k y_k (1+x_k^{'}x)^L\right)$$

והאיבר בסוגריים הוא פולינום מסדר L.

דוגמא:





שימוש זה בפונקצית גרעין לחישוב מכפלות פנימיות במימד גבוה מכונה ה- Kernel Trick. הוא שימושי גם לתחומים נוספים בלמידה ממוחשבת כגון PCA ועוד.

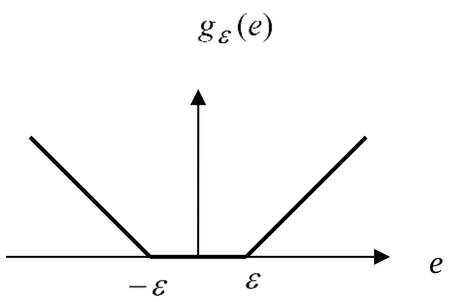
<u>הערות נוספות:</u>

1. SVMs (עם או בלי שימוש ב-Kernel Trick) הם כיום שיטה מובילה לסיווג.
2. קיימות הרחבות למרבית תחומי הלמידה הממוחשבת – בעיות תיוג רב-מחלקתיות, קרוב פונקציונלי (רגרסיה), PCA, אישכול, וכו'.
3. היישום לבעיית הקרוב הפונקציונלי, מתבסס על החלפת המחיר הריבועי

$$E=(y_i-\hat{f}(x_i))^2$$

בפונקציית מחיר ליניארי עם "מרווח אי-רגישות":

$$E=g_\varepsilon(y_i-\hat{f}(x_i)).$$



4. העמקה בנושאי פונקציות גרעין ו-SVM ניתן למצוא בספרי הלימוד, וכן בספרים:

N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, 2000

Smola and B. Schölkopf, *Learning with Kernels*, 2002.

מאמרי סקירה וחומר נוסף ניתן למצוא באתר : http://www.kernel-machines.org .