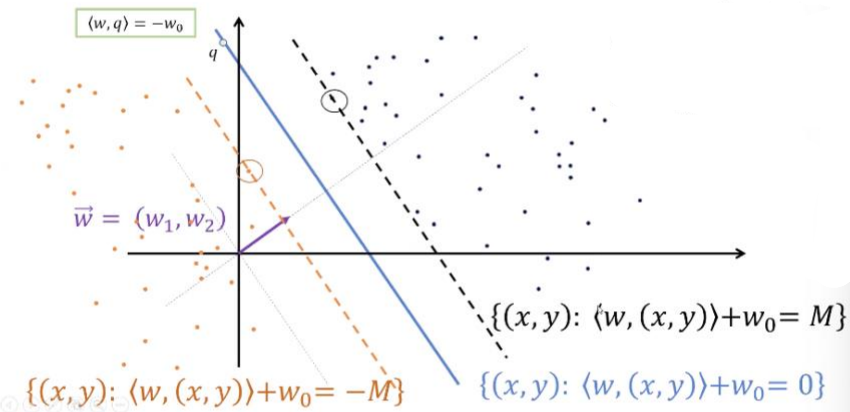
למידה חישובית ממידע: הרצאה 8 – SVM ומבוא לתיאוריה

המשך SVM – Maximal Margins, Optimization and Slack variables

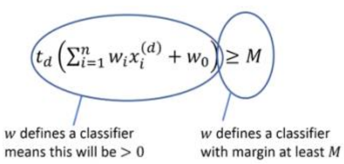
- הגיאומטריה של השוליים
- אופטימיזציה תחת אילוצים
- משתני Slack



פירוש גאומטרי עבור השוליים

בהגדרה הרשמית של שוליים – margin, התייחסנו למרחק מינילי בין הנקודה Xd למישור f(x)=0, כדי למצוא את המרחק הזה אנחנו לוקחים את הנקודה במישור הקרובה ביותר ל-Xd ונחשב עבורה את המרחק (האנך).

נניח כי יש לנו דאטה הניתן להפרדה לינארית ונחפש w שהוא מפריד ליניארי שממזער את השוליים. אם M זה האורך בו השוליים ישיגים (ברי השגה, achievable) אזי קיים וקטור יחידה (באורך 1) כך שמתקיים האיור.



יהי M השוליים הישיגים עבור הדאטה שלנו. אזי קיים וקטור יחידה w כך שלכל נקודות הדאטה מתקיים:

שוליים מקסימליים / Maximum margin

- נסמן את השוליים עבור מישור מועמד בווקטור יחידה w, מהיות M > 0 אנחנו יודעים שעבור כל הדגימות מתקיים האי שוויון לעיל.
- אנחנו מחפשים מסווג בעל שוליים מקסימליים ולכן, אנחנו מחפשים w ו-w0 שיכולים לפתור את בעיית האופטימיזציה הבאה:

$$\max_{M,w,w_0} M \quad \text{subject to} \quad \forall d \quad t_d\left(\frac{\langle w, x^{(d)}\rangle}{\|w\|} + w_0\right) \geq M$$

w מגדיר את הכיוון, w0 מגדיר את המרחק מהכביש, כאשר w הוא וקטור הוא n-ממדי, w0 הוא סקלר, M הוא סקלר. X הוא ווקטור n-ממדי.
הגדרנו את כל אוסף נקודות המגדירים את המגבלות לכל נקודות הדאטה, ונרצה למצוא את M המקסימלי עבורן.

בעיית אופטימיזציה זו שקולה לבעיית האופטימיזציה הבאה:

$$\text{Minimize } \tfrac{1}{2}\|\mathbf{w}\|^2 \text{ subject to :}$$
$$t_d\left(\mathbf{w}\cdot\mathbf{x}^{(d)} + w_0\right) - 1 \geq 0 \quad \forall d$$

(הראינו בהרצאה כיצד הגענו לשקילות)

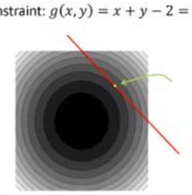
אופטימיזציה תחת אילוצים = כופלי לגרנז'

התבוננות על נקודות הקיצון תחת domain מסוים.

- השיטה של כופלי לגרנז' (Lagrange multipliers) היא אסטרטגיה למציאת מקסימום/מינימום לוקאליים של פונקציה תחת אילוצי שיוויון. למשל: Maximize $f(x,y)$ subject to $g(x,y) = 0$
- אנחנו מניחים כי גם f וגם g הן בעלות נגזרת ראשונה חלקית רציפה
- השיטה מציגה משתנה חדש (למדא = (λ)) שנקרא כופל לגרנז' ולומדת את פונקציית לגרנז' המוגדרת ע"י: $L(x,y) = f(x,y) - \lambda g(x,y)$
- תנאי הכרחי עבור נקודה p* = (x , y) להיות הפתרון של בעיית האופטימיזציה המקורית הוא שכל הנגזרות החלקיות של פונקציית הלגרנז' הן (הגרדיאנט של L הוא 0): $\nabla L(x,y) = 0$

דוגמה:

- Minimize $f(x,y) = x^2 + y^2$
- Subject to the constraint: $g(x,y) = x + y - 2 = 0$



הקו האדום מייצג את g. העיגולים הם קווי הגובה של הפונקציה f. ככל שהעיגולים שחורים יותר כך הערך של f יותר קטן. נרצה למצוא את הערך המינימלי של f תחת המגבלה של הקו האדום. לכן נלך לאורך הקו האדום נגד הגרדיאנט עד שנתכנס למינימום.

דוגמה נוספת:

נתונים אליפסה f ומעגל ועלינו למצוא את המינימום של האליפסה f תחת המעגל.

נשתמש במשפט לגרנז' והתקבלו 4 נקודות מועמדות לקיצון, נציב ב- f במטרה למצוא את הקיצון:

Find the min and max values of $f(x,y) = x^2 + 2y^2 - 4y$ subject to $x^2 + y^2 = 9$.

<u>Solution</u>

Set three equations as follows

$\nabla f = \lambda \nabla g \quad \Rightarrow \quad 2x = \lambda 2x \quad , \quad 4y - 4 = \lambda 2y$

and the constraint implies $x^2 + y^2 = 9$.

$x = 0$ $y = \pm 3$	$\lambda = 1$ $y = 2$ $x = \pm\sqrt{5}$

Plugging these 4 points into the function we get:

$$f(0,3) = 6$$
$$f(0,-3) = 30$$
$$f(\sqrt{5},2) = f(-\sqrt{5},2) = 5$$

בחזרה ל-SVM

נזכיר את המטרה שלנו למזער את $\frac{1}{2}\|\mathbf{w}\|^2$ תחת האילוצים הבאים $t_d\left(\mathbf{w}\cdot\mathbf{x}^{(d)} + w_0\right) - 1 \geq 0 \quad \forall d$. נשים לב כי לגרנז' לא אומר לנו שום דבר לגבי אי-שוויון. בנוסף, אין התייחסות לקיום התנאי עבור כל האילוצים לכל הנקודות. אז קיימת הרחבה אשר אומרת שמשפט לגרנז' מתקיים גם עבור אי שוויונים וגם עבור אילוצים על כל נקודות הדאטה d.

<u>משתני סלאק Slack =</u>

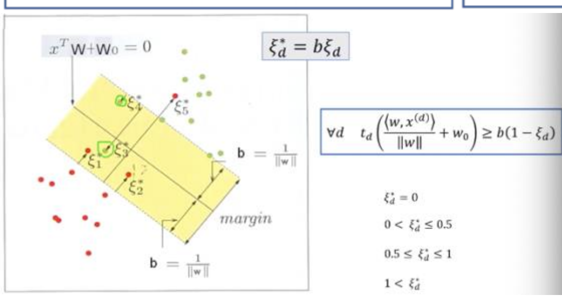
נעניק לכל נקודת דאטה d, משתנה סלאק שמסומן באות קסי היוונית (נראית כמו נחש), שהוא אי שלילי. ונוצרת לנו בעיית אופטימיזציה דומה, מממד שונה – ממד אחד יותר, יש לנו קסי לכל דגימה.

$$\min_{w,w_0} \|w\|^2$$
subject to
$$\forall d \quad t_d\left(\langle w, x^{(d)}\rangle + w_0\right) \geq (1 - \xi_d)$$
$$\xi_d \geq 0, \qquad \sum_{d\in D} \xi_d \leq C$$

$$\max_{M,w,w_0} M$$
subject to
$$\forall d \quad t_d\left(\frac{\langle w, x^{(d)}\rangle}{\|w\|} + w_0\right) \geq M(1 - \xi_d)$$
$$\xi_d \geq 0, \qquad \sum_{d\in D} \xi_d \leq C$$

<u>המשמעות של משתני סלאק:</u>

ככל שהתיקון קטן, הקסי קטן וככל שהתיקון גדול, הקסי גדול.

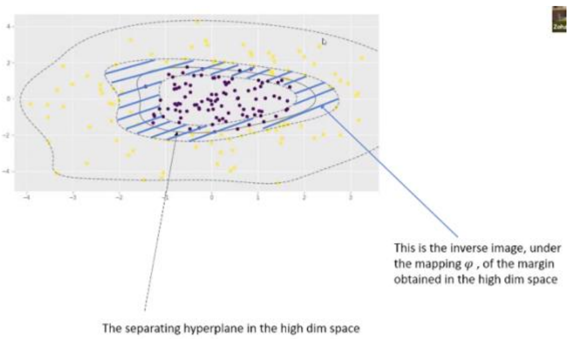


סיכום SVMs

- האלגוריתם המבצע שלנו דורש support vectors, מקדמים וקרנל:

$$class(\vec{x}) = \text{sgn}\left(\sum_{d\in SV} a_d\, t_d K(\vec{x}_d, \vec{x})\right)$$

- הלמידה דומה לקרנל פרספטרון אבל משתמשת ב- convex optimization ובגישת לגרנז' / KKT (ההרחבה של לגרנז').
- בלמידה אנחנו מוצאים מפריד ליניארי בעל שוליים רחבים במרחב מממד גבוה.
- ה-Kernel מייצג את המכפלה הפנימית בממד הגבוה.
- SVMs מאפשרים מיס-קלסיפיקציות במידה מסוימת שנשלטת על ידי היפר-פרמטר.



Introduction to Learning Theory / מבוא לתיאוריית הלמידה

<u>בלמידה חישובית ממידע –</u>

אנחנו ממדלים היפותזה h ממרחב ההיפותזות H למידע שאנחנו רואים (training set). לכן, זוהי בעיית שיערוך – אנחנו נרצה שהטעות של ההיפותזה שלנו תהיה הכי קטנה שאפשר על ה-training set. לרוב זה נקרא "in-sample-error". אבל, בלמידה אנחנו לא באמת מתעניינים ב-"in--sample-error" אלא בטעות שאנחנו לא רואים! זה נקרא לרוב "out-of-sample-error" או טעות ההכללה (generalization error).

<u>שיערוך מול הכללה / Approximation vs. Generalization :</u>

- שיערוך / Approximation : מודד כמה טוב ההיפותזה ממדלת את ה-training data.
- הכללה / Generalization : מודדת כמה טוב ההיפותזה צפויה למדל דאטה חדש.

בלמידה אנחנו מעוניינים בהכלה ולכן תהליך הלמידה הוא קשה. אנחנו נרצה דרך להעריך את הביצועים של ההכללה מתוך ה-sample data.

סיבוכיות הדגימה – כמה training data נחוצה עבור רמה מסויימת של ביצועים.

דוגמה: למידת פונקציה בוליאנית

נרצה ללמוד פונקציה בוליאנית (קלסיפיקציה בינארית) מעל 4 משתני אינפוט : $f(x_1, x_2, x_3, x_4) = t \in \{0,1\}$

ונתנו לנו 7 דגימות – training examples, עליהן אנחנו יודעים את ה-label.

במרחב שלנו יש 16 נקודות, ומכיוון שנתנו לנו 7, יש לנו 9 נקודות שאנחנו לא יודעים עליהן כלום – יש לנו 2^9 דיכוטומיות (פונקציות בוליאניות) אפשריות שהן קונסיסטנטיות עם ה-training data.

Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

Note:
$X = \{0,1\}^4$ and therefore $|X| = 16$.
If all Boolean functions are in our hypotheses space H then $H = 2^{16}$

אין ארוחות חינם / "No Free Lunch"

- טעות ההכללה (שמסומנת $Err_{GEN}(h)$) של היפותזה היא מדד הטעות של h עבור כל ה-non-training examples.
- יהי F מרחב כל הקונספטים האפשריים עבור y = f(x) שהן קונסיסטנטיות עם training dataset מסוים.
- **משפט: עבור כל היפותזה h, טעות ההכללה הממוצעת(*) מעל כל הקונספטים ב-F היא 0.5:** $\frac{1}{|\mathbb{F}|}\sum_{f\in\mathbb{F}} Err_{GEN}(h) = 1/2$
- כאשר ממוצעת = בהנחה שכל ההיפותזות הקונסיסטנטיות h הן סבירות במידה שווה.

<u>**הסבר עבור המשפט אין ארוחות חינם:**</u> יש 512 **הרחבות אפשריות** ל-training data. נתבונן בהרחבה מספר 3 שמסומנת ב-f3 ונגדיר את '(f3) להיות – הנגטיב של f3. כמה טעויות f3 יכולה לעשות? במקסימום 9 ובמינימום 10. למעשה הטעויות של '(f3) משלימות את הטעויות של f3, כך שאם f3 עשתה 4 טעויות, '(f3) עשתה 5 טעויות. לכן מספר הטעויות הממוצע של f3 ו-'(f3) הוא בדיוק 4.5! מכיוון שלכל היפותזה יש את "ההיפותזה-תגי" שלה, ההנחה של אין ארוחות חינם היא להצמיד לכל היפותזה את ההיפותזה ההפוכה לה, ולכן **הסתברות השגיאה על כל ההרחבות היא חצי.**

- משפט ה-NFL מניח כי כל הקונספטים שהם קונסיסטנטים עם ה-training הם בעלי סבירות שווה בהינתן ה-training.
- במציאות לא כל הקונספטים סבירים באותה המידה.
- תופעות ריאליסטיות אינן דגימות שמתפלגות יוניפורמית בכל ההרחבות האפשריות של הדאטה. קונספטים אמתיים (טבעיים, פרי-אדם, סוציולוגיים) הם בעלי רגולריות, חוקים, מבנה.
- דגימת training עם ערכי-אטריביוט נתונים נותנת אינדיקציה לגבי הקלאס האמיתי של דגימת ה-non-training שהיא בעלת ערכי-אטריביוט דומים. כפי שראינו – אנחנו מנצלים תכונות אלה.

<u>ה-Set-up הכללי שלנו</u> : ננסה ללמוד קונספט c (פונקציה, קלסיפיקציה – דיכוטומיה : כאשר c מוכל במרחב הדגימות). יש לנו training data : דגימות ממרחב האינסטנסים X. נרצה להציע היפותזה (אלגוריתם מבצע) בעל צורה מאפיינת בהתאם למשימה. עבור מסווגים אנחנו משתמשים במרחב ההיפותזות H – אשר משרה תת-קבוצה של קבוצת החזקה של X.

בחזרה לדוגמה שלנו של הפונקציה הבוליאנית:

נגביל את מרחב ההיפותזות שלנו, כך שלא יהיה לנו יותר 512 הרחבות אפשריות של ה-training data. יש רק היפותזות מסוג מסוים הן כשרות, נניח כי גימום conjuctions (משפט וגם) מגדיר את הקונספט שלנו ונציג את כל משפטי ה"וגם" האפשריים עם 4 משתנים.

משפט "וגם" הריק לא מייצג את הדאטה שלי מפני שיש עבורו דוגמות נגדיות.

נמשיך ונשלול את כל משפטי "הוגם" האפשריים הנ"ל, מפני שעבור כולם יש לנו ב-training דוגמות נגדיות, ונגיע לכך **שאין מודל מתוך מרחב ההיפותזות שגדרנו שהוא קונסיסטנטי (מסכים) עם הדאטה (ה-training).**

Rule	Counterexample
$\Rightarrow y$	1
$x_1 \Rightarrow y$	3
$x_2 \Rightarrow y$	2
$x_3 \Rightarrow y$	1
$x_4 \Rightarrow y$	7
$x_1 \wedge x_2 \Rightarrow y$	3
$x_1 \wedge x_3 \Rightarrow y$	3
$x_1 \wedge x_4 \Rightarrow y$	3
$x_2 \wedge x_3 \Rightarrow y$	3
$x_2 \wedge x_4 \Rightarrow y$	3
$x_3 \wedge x_4 \Rightarrow y$	4
$x_1 \wedge x_2 \wedge x_3 \Rightarrow y$	3
$x_1 \wedge x_2 \wedge x_4 \Rightarrow y$	3
$x_1 \wedge x_3 \wedge x_4 \Rightarrow y$	3
$x_2 \wedge x_3 \wedge x_4 \Rightarrow y$	3
$x_1 \wedge x_2 \wedge x_3 \wedge x_4 \Rightarrow y$	3

Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

לכן, ננסה להגדיר מרחב היפותזות אחר, m-out-of-n, ונבדוק עבורו. להלן הדוגמות הנגדיות עבור מרחב ההיפותזות הנ"ל – ונראה כי לכל אחת מהאפשרויות נקבל דוגמות נגדיות.

עבור ה-* אנחנו כן מסכימים עם הדאטה!**

רק אחד מהם הוא קונסיסטנטי עם הדאטה – למדנו את הקונספט האמיתי בהנחה שאין טעויות ובהינתן מרחב היפותזות מוגבל!

אבל הגענו לקונספט האמיתי בכך שהגבלנו את מרחב ההיפותזות שלנו. ישנם עוד מודלים שהם קונסיסטנטים עבור הדאטה שלנו.

variables	Counterexample 1-of	2-of	3-of	4-of
$\{x_1\}$	3	–	–	–
$\{x_2\}$	2	–	–	–
$\{x_3\}$	1	–	–	–
$\{x_4\}$	7	–	–	–
$\{x_1, x_2\}$	3	3	–	–
$\{x_1, x_3\}$	4	3	–	–
$\{x_1, x_4\}$	6	3	–	–
$\{x_2, x_3\}$	2	3	–	–
$\{x_2, x_4\}$	2	3	–	–
$\{x_3, x_4\}$	4	4	–	–
$\{x_1, x_2, x_3\}$	1	3	3	–
$\{x_1, x_2, x_4\}$	2	3	3	–
$\{x_1, x_3, x_4\}$	1	***	3	–
$\{x_2, x_3, x_4\}$	1	5	3	–
$\{x_1, x_2, x_3, x_4\}$	1	5	3	3

Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

Example	x_1	x_2	x_3	x_4	y
1	0		1	0	0
2	0		0	0	0
3	0		1	1	1
4	1		0	1	1
5	0		1	0	0
6	1		0	0	0
7	0		0	1	0

הגבלת מרחב ההיפותזות:

ע"י בחירת סוג הפונקציה אנחנו מגבילים את מרחב ההיפותזות. להלן חסרונות ויתרונות:

- חסרון: הפונקציה/קונספט האמתיים עשויים לא להיות שייכים בכלל למרחב ההיפותזות שבחרנו, וכן היפותזות מורכבות יותר אינן תמיד טובות יותר.
- יתרון: היפותזות פשוטות הן טובות יותר עבור הכללה. בכך אנו נמנעים מ-overfitting, למרות שהדאטה שלנו יכול לכלול טעויות. הן יכולות "לתפוס" מבנה חבוי ועל כמו מאפשרות למידה, והן יותר קלות ללמידה (מבחינה חישובית).

מרחבי היפותזות באלגוריתמי למידה שלמדנו:

Algorithm/Problem	Hypothesis Space
Regression	Linear functions, polynomial functions etc
Decision Trees	Trees/axis aligned boundaries
Perceptron	Linear decision boundary (+ we can map to higher dimensions)
KNN	Complex separating boundaries – depends on K
SVM	Separating hyperplane in φ–space (depends on Kernel choice)
Logistic Regression	A linear decision boundary (+ we can map to higher dimensions)

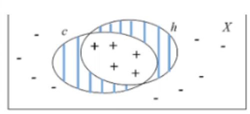
<u>הטעות האמתית (במסווגים)</u>

<u>היא ההסתברות למיס-קלסיפיקציה:</u>

הטעות האמתית של היפותזה h ביחס ל-target concept, c, היא ההסתברות ש-h יסווג באופן שגוי דגימה שהוגרלה רנדומלית מתוך התפלגות הדאטה F:

$$error_F(h) = Pr_{x \sim F}[c(x) \neq h(x)]$$

הטעות תלויה ביותר בהתפלגות הדאטה F!



<u>שיערוך טעויות</u>

- שאלת המפתח שעולה כאן היא: האם טעות ה-in-sample יכולה להגיד לנו משהו אודות טעות ה-out-of-sample? ובאופן יותר כללי: האם נוכל להגדיר ולומר משהו על הטעות האמיתית או על הטעות המצופה / expected error של ההיפותזה שלנו?
- כאשר נפתח את תיאוריית סיבוכיות הדגימה נבדיל בין שני המקרים הבאים:
  1. הקונספט האמתי נמצא במרחב ההיפותזות.
  2. הקונספט האמתי אינו במרחב ההיפותזות.
- נשתמש ב-test set כדי לשערך את הטעות האמתית של היפותזה מועמדת. אם ה-test set הוא יתרת X (מרחב הדגימות) אזי נדע את הטעות האמתית! כמובן שזהו מקרה לא ריאלי – אנחנו חייבים להסתמך על הדגימות ונגדיר את טעות הדגימה, עבור קבוצת דגימות באופן הבא: $error_S(h)$ = the ratio of misclassified samples in S
- ככל ש-S גדולה יותר כך ההערכה טובה יותר.

תהליך שיערוך סטטיסטי / Statistical Estimation Procedure –

נשתמש ב-test set בגודל |S| ונניח כי ראינו k שגיאות. נוכל להראות כי משערך עבור טעות ההכללה יהיה k / |S|. כאשר אנו תלויים בגודל של ה- test set שלנו, נוכל לקבל הבטחה סטטיסטית כמו : בוודאות של 95% אנו יודעים שהשגיאה האמיתית היא קטנה מ- eps + |S| / k. וזה נובע מחישוב אינטרוול הוודאות.

לסיכום ההרצאה:

- NFL Thm
- Hypothesis spaces and how they allow for learning
- “out-of-sample” error and “in-sample” error
- Statistically assessing errors