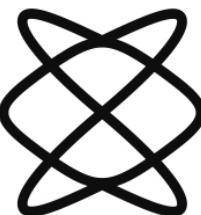


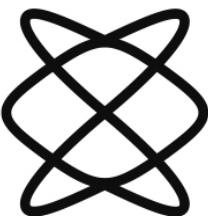
הפקולטה למדעים מדויקים
ע"ש רייןmond וברלי סאקלר
אוניברסיטת תל אביב



החולג למדעי המחשב (0368) מבוא ללמידה חישובית (3235) (גרסה ארוכה)

מרצה: נדב כהן
מתרגלים: אורין לוי, ליעד ארץ
תשפ"ד, סמסטר א' (2024)

מסכם: רועי מעין



The Raymond and
Beverly Sackler Faculty
of Exact Sciences
Tel Aviv University



פרק 1 – למידה מפוקחת (Supervised): תיאוריה

3	מבוא
10	תיאוריה של למידה מפוקחת
18	מחלקות אינסופיות
23	בחירה מודל

פרק 2 – למידה מפוקחת (Supervised): אלגוריתמים

28	אופטימיזציה
36	SVM
49	Decision Trees
54	Regression

פרק 3 – למידה לא מפוקחת (Unsupervised)

58	שיטות נוספות
63	למידה לא מפוקחת



1 – תיאוריה של למידה מפוקחת (Supervised)

מבוא

רקע ללמידה מוכנה

בקע: למידה היא שינוי ההתנהגות לאורך זמן, בצורה שתשתמש מטרה בלהשי (לומדים ללבת, לנוהג, לדבר שפה). בלמידה מוכנה, אנחנו רוצים לבנות מכונות שיכולים ללמידה, להבין את התהילה הלמידה זהה בצורה מתמטית. אין סיבה להגביל את יכולת של מחשבים ללמידה לרמת היכולת של בני אדם, לעתים המכונות "יעקפו" אותו.

דוגמה – פלטוור ספאם: נרצה לסוג מייל, האם הוא ספאם או לא. בגישה האנושית-מקצועית, נרצה לנשח ט של כלים שיעודו לקבוע האם הוא ספאם או לא (אם הכותרת-ב-CAPS זה ספאם, אם זה נשלח מאנשי הקשר שלנו זה לא ספאם...). בגישה זו יש הרבה חסרונות: דרושה הרבה עבודה אנושית, צריך לעדכן ולחזק את החוקים האלה, ולאו דווקא בני אדם הם היכי טובים במציאת האסטרטגייה הנכונה למציאת ספאם (עלינו לעבד במודות גדולות של נתונים ולקבל את החלטות הנכונות).

בגישה ה-ML, נרצה **למד** מוכנה לפטור ספאם, בהתאם על הרובה **דוגמאות**. תוכנת המחשב תלמד מיפוי (פונקציה) שמקבל מיל ומחזר תשובה ספאם כן/לא.

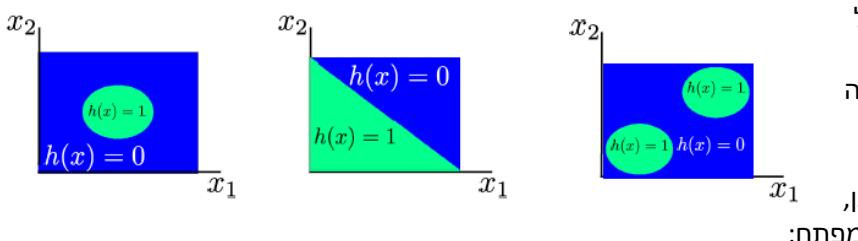


אנחנו ניגש לתחום מהכיוון של **סטטיסטיקה/אופטימיזציה**. יש כיוונים נוספים כמו תורת האינפורמציה/עיבוד אותות, ופילוסופיה/פסיכולוגיה/מדעי המוח. הכלים המתמטיים הבסיסיים בהם נשימוש:

1. **הסתברות** – על מנת למ笪 חוסר וידאות, מה שיש הרבה בלמידה.
2. **אלגבראה לינארית** – כלים כמו מרחבים וקטורים, ביון שהאדואן הוא רב-ממדי.
3. **חו"א ואופטימיזציה** – כלים כמו נגזרות, על מנת למשם אלגוריתמי למידה בצורה יעילה.

למידה מפוקחת:

- **בעיות קלסיפיקציה:** משימות כמו לחתת מייל ולומר אם הוא ספאם, מה מופיע בתמונה (כלב/חתול), איך ספרה מופיעה בתמונה. כל אלו הם מיפויים של **קלט** לקבוצה סופית של **labels**.
- **בעיות רגרסיה:** משימות כמו לחתת סופגנית וЛОוגריטם, ולומר כמה קלוריות יש בה, הסתברות לכך שהלוואה תוחזר, צפיפות על סמן פסיכוןtri ציונים בהתאם. הפלט הוא לרוב **רצייף** (מספר לא סופי של labels), ובנוסח **labels** יש **משמעות נומרית**. אם ניתן פרודיקציה שהיא קרובה לאמת זה יכול להיות מספיק טוב. לעומת זאת קלסיפיקציה בה הטעות היא גדולה יותר.
- **בעיות מורכבות יותר:** ביצוע STT (שמע לטקסט), תרגום משפט מאנגלית לעברית, יצירת תמונה מטקסט. הפלט הוא **רצייף reinforcement learning**, **structured prediction**, **classification**, **structured prediction**.



נייצג מසוג (classifier) בתרו פונקציה (מיפוי) ממוחב של קלטים למרחב של תוצאות אפשריות $\mathcal{Y} \rightarrow \mathcal{X}$: $h: \mathcal{X} \rightarrow \mathcal{Y}$. בקורס רוגרטי ביןארית נחשוב על התוצאה $\mathcal{Y}^d = \mathcal{X}$. נניח לרגע כי הפלט הוא המישור $\mathcal{X} = \mathbb{R}^2$. אז המסוג \mathcal{X} מחלק את המישור לשני חלקים: אחד מתואג עם $label=1$ והשני מתואג עם $label=0$. אם כן, המטרה המרכזית היא **ללמד מסוגים מוגדרים**. שאלות מפתח:

1. **איזה משפחה** של מסוגים לבחור?
2. **בהתנן** משפחה ודואט, איך לבחור את **המסוג** המסוגים?
3. **בהתנן** מסוג, מה **ביצועים** שלו בעולם האמתי? (דוגמאות שלא ראיינו ב-set train-set test).

למידה לא מפוקחת:

לאסוף **labels** או משימה קשה (תמונה שאדם כבר תיאר אותה, ונתן להם **label**). יש הרבה דatasets לא מתווים. נרצה למצוא תיאור מסוים של הדטאא בתרור מבנה, להפיק insights על הדטאא כך שהוא מועיל לשימושים שונים. נושאים שניגע בהם:

- PCA: שיטה להורדת ממד. למשל, נוכל ליצג תמונות 100X100 בMiami הרבה הרבה יותר נמר, לדחוס את המידע.
- Clustering: חלוקת הדטאא לקבוצות.
- Generative models: נרצה ללמידה את התפלגות של הדטאא, מה התהילה ההסתברותי שמייצר אותו. כך נוכל ליצור דאטא חדש בעצמו.



הסתברות והכללה

מרחב הסתברות:

- קבוצת **תוצאות** (outcomes) אפשריות. למשל בנסיון של הטלת מטבע 5 פעמים, יש $32 = 2^5$ תוצאות שונות. כל תוצאה כזו היא רצף של תוצאה אפשרית בכל הטלה (2 תוצאות אפשריות).
- קבוצת **מאורעות** (events), כל אחד כזה הוא תת-קובוצה של התוצאות. למשל, המאורע שההטלה הראשונה היא עז. במקורה הזה המאורע מכיל 16 תוצאות. נחשב על **מרחב המאורעות** zusätzlich של כל קבוצות המאורעות האפשריות (קבוצת חזקה של קבוצת התוצאות).
- **פונקציית הסתברות**: עברו כל מאורע, ניתן מספר בין 0 ל-1 שהוא הסיכוי שהמאורע יקרה. למשל, המאורע שההטלה הראשונה היא עז, ההסתברות של המאורע זהה היא חצי (בהתבהה שהמטבע הוגן, וההטלות הן ב"ת").

משתנה מקרי: משתנה מקרי X היא פונקציה מקבוצת התוצאות למספרים. עברו $\mathbb{R} \in x$ נסמן $b-x = X$ את המאורע שמכיל את התוצאות עליו X מוחזר, כלומר, ככלומר המקור של x .

לדוגמה, נגיד את X להיות מספר הפעמים שיצא עז: עברו רצף מסויים (תוצאה), נוכל לחת ערך ממשי בין 0 ל-5. נסתכל על המקרה שבו $2 = X$, כמה אפשרות באלו יש? לתוצאות שבין יצא 2 פעמים עז – 10 אפשרות כיוון שאנו שוכנים את המיקומים של 2 העז ובשאר המיקומים יש פלי. זה יוצא $10 = \binom{5}{2}$.

אפשר להסתכל על ההסתברות של המאורע שיצא פעמיים עז: $P[X = 2] = \frac{10}{32}$. זאת בהתבססות על הנוסחה $P[A] = \frac{|A|}{|\Omega|}$, ככלומר כמות התוצאות במאורע המדובר, חלקו בכמות התוצאות הכלליות במרחב המודגם.

הקשר ל-ML:

בלמידה מפקחת, אנחנו מקבלים משתנה (למשל תמונה) ואנו שוכנים מתעניינים בערך של משתנה אחר. נסמן את הקלט במשתנה מקרי X (input space), ואת הפלט במשתנה מקרי Y (label space). המטרה: לחוץ את Y בהינתן X . סוג התפלגיות:

- **התפלגות המשותפת** (joint distribution): $[y = X = x]$, ככלומר ההסתברות של המאורע שבו המשתנה המקרי X שווה לערך x , והמשתנה המקרי Y שווה לערך y .
- **התפלגות השולית** (marginal distribution): $\sum_{y \in Y} P[X = x, Y = y]$.
- **התפלגות המותנית** (conditional distribution): $P[Y = y | X = x] = \frac{P[X = x, Y = y]}{P[X = x]}$. אנחנו מגבלים את עצמנו למאורע מסוים $x = X$, ומהוועם החדש הזה נשאל מה ההסתברות שיתקיים $y = Y$.

הנחה זמנית: אנחנו יודעים את ההתפלגות המשותפת $P[X = x, Y = y]$.

אנחנו רוצים לבנות מסווג $\hat{Y} = h(X)$ שמקבל משתנה מקרי חדש, אופטימלי (במובן שבו \hat{Y} קרוב ל- Y). נניח שאנו מקבלים תמונה, וה-label שלנו הוא חתול או כלב. ככלומר, העולם שלנו הוא כמו מבוגת מזל שפolut זוגות (x, y), תמונה ותיוג. נרצה לקחת רק את החלק X (התמונות שמיוצרות בטבע), להפעיל עליה פונקציה דטרמיניסטיבית, ונקבל תיוג \hat{Y} שהוא בדרך כלל כמו התיוג האמיתי Y . נ��וד את המרחק/הקרבה הזה באמצעות *loss function*.

פונקציית הפסד (loss function): פונקציה דטרמיניסטיבית ℓ המקבלת שני תיוגים ומחזירה את המרחק/השוני (discrepancy) ביןיהם, מה המחיר של טעות בתיאור התיוג שקיבלנו. דוגמאות לפונקציות הפסד:

שם	פונקציה	הערות
Zero-one loss	$\ell(y, \hat{y}) = \begin{cases} 1 & y \neq \hat{y} \\ 0 & y = \hat{y} \end{cases}$	הרבה מההתוצאות התאורטיות שנראתה בהתחלה יתיחסו לפונקציה זו. נראה בהמשך שקשה לעבד אותה מבחינה חישובית.
Quadratic loss	$\ell(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$	מתאים לרגרסיה, שם יש לתיוג ערך נומרי. נקרא גם ℓ_2 -square of ℓ_2 .
Domain specific	$\ell(y, \hat{y}) = \begin{cases} ? & y \neq \hat{y} \\ 0 & y = \hat{y} \end{cases}$	כאשר הפדריקציה היא כמו התיוג האמיתי ניתן 0, ובאשר לא, צריך לשקף כמה לא רצוי/חמור לחוזות \hat{Y} כאשר התיוג האמיתי הוא y .

בהינתן פונקציית הפסד ℓ , השאייה שלנו היא למצער את תוחלת הפסד (הפסד הממוצע, הצפוי). אנחנו לוקחים שני משתנים מקרים (X, Y) ומדינים לתוך פונקציית הפסד, שמחזירה גם היא משתנה מקרי. ניקח את התוחלת שלו, ונקבל מספר. התוחלת מבטא את המיצוע של כל-lossים שקיבלנו. אנחנו רוצים **למצוא את המסוג h** שמצער את המספר הזה:

$$\arg \min_h L(h) := \mathbb{E}[\ell(Y, h(X))] = \sum_{x \in X, y \in Y} P[X = x, Y = y] \cdot \ell(y, h(x))$$

בעולם האמיתי, את ההתפלגות המשותפת אנחנו לא יודעים. אנחנו מנסים למצער משהו שאנו לא יכולים לחשב אותו.



מסaggio MAP (Maximum A-Posteriori)

נסתכל לדוגמה על סיווג בינארי (binary classification), באשר $\{0,1\} \in Y$ עם פונקציית loss zero-one. בהינתן $\bar{x} = X$ (תמונה מסוימת), מה הבדיקה הכי טובה (\bar{x})? אין לנו אפשרות על מה שהפונקציה h מוחירה עבור קלטים שונים, וכן אפשר **לפרק את הבעה** לביעות קטנות – נקבע כל פעם את x ובחר את (x) הכי טוב. האלמנטים ב- L שתלויים ב- \bar{x} הם:

$$\begin{aligned} P[X = \bar{x}, Y = 0] &= \ell(0, h(\bar{x})) + P[X = \bar{x}, Y = 1] \cdot \ell(1, h(\bar{x})) \\ P[X = \bar{x}] \left(P[Y = 0|X = \bar{x}] \cdot \ell(0, h(\bar{x})) + P[Y = 1|X = \bar{x}] \cdot \ell(1, h(\bar{x})) \right) &= P[X = \bar{x}] \cdot L \end{aligned}$$

סימנו ב- L את מה שנשאρ בסוגרים, ואנחנו רצאים למצער את L . במקרה יש לבחור האם $1 or 0$

- אם 0 אז $h(\bar{x}) = 0$ כי loss הוא על 1. מתי נרצה לחוץ 0? בשמהיר של 1 נמוך יותר משל 0.
- אם 1 אז $h(\bar{x}) = 1$ כי loss הוא על 0. באן נרצה לחוץ 1 בשמהיר של 0 נמוך יותר משל 1.

אנחנו רצאים ש- L יהיה נמוך אך אנו נבחר $\bar{x} | h(\bar{x}) = \arg \max_{y \in \{0,1\}} P[Y = y | X = \bar{x}]$. אנו ממשמים את מה שלא בחרנו, ונרצה שהמהיר שלו יהיה נמוך יותר. בהתאם נרצה שמה שבחרנו יהיה המחיר הגבוה יותר (נבחר ב-0 כאשר ההסתברות של 0 גבוהה יותר, ונבחר ב-1 כאשר ההסתברות של 1 גבוהה יותר). MAP הוא המסaggio האופטימלי עם פונקציית loss zero-one גם כאשר ב모ת התוצאות (y) גדולה מ-2.

דוגמה – spam filtering: נגידו את התנאים הבאים:

- X – מספר המילים במיל-ב-upper case.
- ידועה ההתפלגות המשותפת $P[X = x, Y = y]$. בנוסף $P[X = x, Y = y] = P[x|Y = y]P[Y = y]$.
- נניח כי $P[X = x|Y = y] \approx 0$ ניתן לקרוב ע"י ההתפלגות גaussian (NORMALITY) עם ממוצע c_y ושונות σ^2 .
- נניח כי $c_1 > c_0$, כלומר ההתפלגות של upper case words שווה יותר ימינה מאשר��ותה.

לפי מסaggio MAP נקבל כי $P[A|B] = \frac{P[B|A]P[A]}{P[B]}$. לפיכך $P[A|B] = \frac{P[Y=1|X=x]P[X=x]}{P[Y=0|X=x]}$. נקבע: $P[Y=1|X=x] = P[Y=1|X=x|h(x) = 1] \Leftrightarrow P[Y=1|X=x] \geq P[Y=0|X=x]$.

$$\begin{aligned} \Leftrightarrow \frac{P[X=x|Y=1] \cdot P[Y=1]}{P[X=x]} &\geq \frac{P[X=x|Y=0] \cdot P[Y=0]}{P[X=x]} \Leftrightarrow P[X=x|Y=1] \geq P[X=x|Y=0] \\ \Leftrightarrow \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-c_1)^2}{2\sigma^2}} &\geq \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-c_0)^2}{2\sigma^2}} \Leftrightarrow (x - c_1)^2 \geq (x - c_0)^2 \Leftrightarrow x^2 - 2xc_1 + c_1^2 \geq x^2 - 2xc_0 + c_0^2 \\ \Leftrightarrow 2x(c_0 - c_1) &\geq c_0^2 - c_1^2 \Leftrightarrow x \geq \frac{c_0 + c_1}{2} \end{aligned}$$

MAP הוא אופטימלי עבור loss zero-one. עבור פונקציות loss מוגדים אופטימליים אחרים. למשל עבור quadratic loss. המסיג האופטימלי עבור סיווג בינארי הוא $x | h(x) = P[Y=1|X=x]$. אנו רצאים למצער את תוחלת הפסד:

$$\mathbb{E}[\ell(Y, \hat{y}) | X = x] = \mathbb{E}[(Y - \hat{y})^2 | X = x] = \mathbb{E}[Y^2 | X = x] - 2\hat{y}\mathbb{E}[Y | X = x] + \hat{y}^2$$

נקבל את המינימום (פרבולת, גוזרים משווים ל-0) כאשר $\hat{y} = \mathbb{E}[Y | X = x] = P[Y=1|X=x]$ וזה יהיה המסיג (x) . פונקציית הפסד משפיעה על המסיג האופטימלי.

שער

בהינתן ההתפלגות המשותפת $P[X = x, Y = y]$ **אפשר תיאורית לקבל את המסיג האופטימלי**. מה קורה אם אנחנו לא יודעים את ההתפלגות המשותפת? זה המקהה המציאות, בו علينا להסיק מה המסיג (explicit or implicit). נניח לדוגמה שאנו רצאים לשערק את ההסתברות של הטלת מטבע – בולומר הפרמטר p של מ"מ ברכנו ל- X :

- מתקיים $p = \mathbb{E}[X = 1] = 1 - \mathbb{E}[X = 0] = P[X = 0].$ בנוסף, $p = \mathbb{E}[X]$.
- אפשר לנקוט m עותקים של המשתנה X_m, X_{m-1}, \dots, X_1 (נניח m הטלות מטבע, ב"ת), וניקח את הממוצע \bar{X}_m (מלינאריות התוחלת): $\mathbb{E}[\bar{X}_m] = \mathbb{E}\left[\frac{1}{m} \sum_{i=1}^m X_i\right] = \frac{1}{m} \sum_{i=1}^m \mathbb{E}[X_i] = p$.
- התוחלת של המשתנה המקרי זהה (מלינאריות התוחלת): $\text{Var}[\bar{X}_m] = \text{Var}\left[\frac{1}{m} \sum_{i=1}^m X_i\right] = \frac{1}{m^2} \sum_{i=1}^m \text{Var}[X_i] = \frac{1}{m^2} \sum_{i=1}^m \text{Var}[X] = \frac{1}{m^2} m \text{Var}[X] = \frac{1}{m} \text{Var}[X]$.
- לגבי השונות: $\text{Var}[\bar{X}_m] = \text{Var}\left[\frac{1}{m} \sum_{i=1}^m X_i\right] = \frac{1}{m^2} \sum_{i=1}^m \text{Var}[X_i] = \frac{1}{m^2} m \text{Var}[X] = \frac{1}{m} \text{Var}[X]$



לכן, אנחנו רואים שהמשערך של \bar{X}_m שהואortalento להיות מ- \bar{X} הוא בעל הממוצע הנכון (biased), והשונות שלו שואפת ל-0 כאשר m שואף לא- ∞ . בכלל שנתיו ויתר פעמים, המשערך יLER ויהיה יותר מדויק.

אנו ממעניין מה קורה לא רק ב- ∞ , בלמידה חישובית אנחנו מסתכלים גם על מספר דוגמאות סופי. נשתמש בחסם הופding (Hoeffding), שנותן חסם על ההסתברות שהמשערך סוטה מ- \bar{X} :

$$P[|\bar{X}_m - p| \geq \varepsilon] \leq 2e^{-2m\varepsilon^2}$$

זה נכון גם כאשר m סופי או אינסופי. הסיבוי שהמשערך רוחק מ- p קטן אקספוננציאלית ככל שמספר הדוגמאות גדול. אנחנו רוצים לבחר m שיבטי שערוך באיכות מסויק טובה. נרצה להבטיח שאנו מ-correct (approximately correct) בפחות הסתברות מסויימת: $\varepsilon \leq |p - \bar{X}_m| \leq \delta$ (with probability w.p. $\geq 1 - \delta$).

אנו יודעים שהסתברות חסומה על ידי $\delta - 1$ ונרצה להבטיח שהוא חסומה על ידי δ . נוכל לדרוש כי:

$$2e^{-2m\varepsilon^2} \leq \delta \Leftrightarrow m \geq \frac{1}{2\varepsilon^2} \log \frac{2}{\delta}$$

ומכך מבטיח כי \bar{X}_m בנראה ($\delta - 1$) נכונה (עד כדי ε). המושג הבסיסי שנראה אותו הרבה הוא **PAC**:

Such m guarantees that \bar{X}_m is **probably** (w.p. $\geq 1 - \delta$) **approximately** (up to ε) **correct**.

תרגול 1 (הסתברות ומסוג אופטימלי)

绰נה על הסתברות:

מושגים בסיסיים:

- **מרחב מדגם:** קבוצה כל התוצאות האפשרות. למשל עבור הטלה מטבע $\{H, T\}^n = \Omega$ ועבור n הטלות $\Omega = \{H, T\}^n$.
- **מאורע:** תת-קבוצה של תוצאות מרחב המדגם $\Omega \subseteq A$.
- **פונקציית הסתברות:** פונקציה $P: 2^\Omega \rightarrow [0, 1]$ המיפה מאורעות להסתברות שלהם (מספר בין 0 ל-1).

צריך להתקיים כי $P[A] = 1$ אם A הסתברות איחוד זו של מאורעות שווה לסכום ההסתברויות. במקרה עבור מאורעות $\Omega \subseteq A, B \subseteq \Omega$ מתקיים $P[A \cup B] = P[A] + P[B]$.

משתנה מקרי (RV, מ"מ): הוא פונקציה $\mathbb{R} \rightarrow \Omega$: המיפה תוצאות מרחב המדגם למספר ממשי. למשל, בניסוי של 3 הטלות מטבע, דוגמה למשתנה מקרי הוא **כמה פעמיים יצא עז**. נפריד בין מ"מ בידיד לרץף.

פונקציית הסתברות מצטברת (CDF): ה-CDF של מ"מ X הוא ההסתברות שיתקיים $x \leq X$ מה הספה עד x , במתה השטח שמצטבר עד x . פורמלית נגיד $F_X(x) = P[X \leq x]$. למשל, אם X הוא מספר הפעמים שיצא עז, בניסוי של 2 הטלות מטבע, נקבל את ה-CDF הבאה.

נאמר כי X הוא מ"מ רציף אם F_X רציפה בכל נקודה. נשים לב כי $P[X = x] = 0$ לכל x , כלומר אין ערכים ספציפיים שהוא מקבל בהסתברות חיובית.

פונקציית צפיפות (PDF): עבור מ"מ רציף אפשר להגיד פונקציית צפיפות. זה ה-f(x) שמתקיים $f_X(x) = dy \int_{-\infty}^x P[X \leq y]$. האינטגרל של ה-PDF נותן את ה-CDF. אפשר לחוש על זה בערך בסיסי לקבל את א: ניקח קטע קטן ליד x , וזה הסיבוי ליפול בקטע זהה חילקו אורוק הקטע (כמו נגזרת).

תוחלת ושונות: התוחלת של מ"מ מוגדרת להיות $E[X] = \sum_{x \in Im(X)} x \cdot P[X = x]$ במקורה המקורי. היא מוגדרת לנו את מיקום מרכז ההתפלגות. למשל אם X הוא מספר העז ב-2 הטלות מטבע: $E[X] = \frac{1}{4} \cdot 0 + \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 = 1$. השונות של מ"מ מוגדרת להיות $Var(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$.

התפלגות משותפת: ברוב הקורס נתעניין בהתפלגות משותפת של שני מ"מ, X והיה האובייקט שהוא רצים לסוג, Y יהיה התויג של אותה תמונה (כלב/חтол). עבור שני מ"מ X, Y אפשר לדבר על ההתפלגות המשותפת שלהם $P[X = x, Y = y]$.

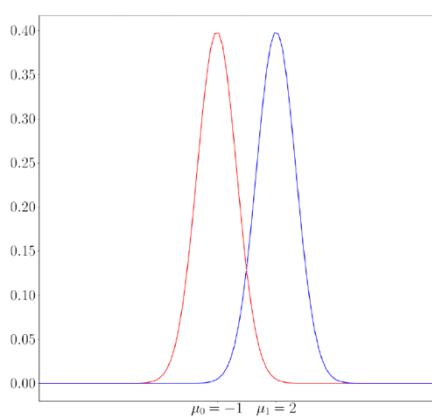


- בהינתן התפלגות הזאת עבר **להתפלגות שולית**, X בshaion לנו מידע על Y : $P[X = x] = \sum_{y \in Y} P[X = x, Y = y]$.
- במקרה הבדיקה, $P[A|B] = \frac{P[A \cap B]}{P[B]}$. שמי כללים חשובים:
 - **בכל השרשרת:** $P[A \cap B] = P[A|B] \cdot P[B]$
 - **בכל ביס:** $P[B|A] = \frac{P[A|B] \cdot P[B]}{P[A]}$.

דוגמה למסוג אופטימלי:

נזכיר תחילת בתפלגות הנורמלית. מ"מ X המתפלג נורמלית הוא בעל פונקציית הצפיפות $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ נסמן $(\mu, \sigma^2) \sim N$ כאשר μ זו התוחלת (קובע את מרכז התפלגות), σ^2 זו השונות.

נניח שאנו יודעים את התפלגות המשותפת D של X, Y . **נרצה למצוא מסוג $Y \rightarrow X: h$ אופטימי.** נרצה למודר את השגיאה של המסוג שמדוברת באופן הבא: $L(h) = E[\Delta(h(X), Y)]$ (Δ היא פונקציית הפסד (loss function)).



- נבעוד עם פונקציית loss 0-1: $\Delta_{0-1}(\hat{y}, y) = \begin{cases} 1 & \hat{y} \neq y \\ 0 & \text{otherwise} \end{cases}$.
- נניח שיש לנו $\mathcal{R} \in \mathbb{R}$ ואנו רוצים לティיג אותו עם $\{0,1\}$ על y .
- בהינתן $0 = Y$ מתקיים $N(\mu_0, \sigma^2)$. ועבור $1 = Y$ מתקיים $N(\mu_1, \sigma^2)$.
- עוד ידוע כי $p_1 = P[Y = 1]$, מה שנקרה הרבה פעמים קודם.

נניח כי התפלגות האדמה היא עבור $0 = Y$ והכולה עבור $1 = Y$. נניח שנ נתונים לנו x שהוא גובה של אדם, ואנו רוצים לסוג אם הוא בן או בת. יש לנו את התפלגות של גובה הבנים באוכלוסייה ושל גובה הבנות. ה-זוווק הוא חצי-חצי, אין לנו עדיפות מראש על מה יותר שכיח באוכלוסייה. לכן, הסוף יהיה באמצעות ממוצע התוחלות, וממנו נקבע 'בן' ומתחייב נקבע 'ב'.

ראינו בשיעור כי המסוג האופטימי במקרה של סיווג ביןארי הוא: $[x = 0|X = x] > P[Y = 1|X = x] \Leftrightarrow P[Y = 1] = 1 \Leftrightarrow h(x) = 1$. אינטואטיבית מסוג 1 אם ההסתברות לקבל 1 יותר גדולה בהינתן ה- x שאנו רואים. בשיעור חישבנו עבור $P[Y = 1] = \frac{1}{2} p_1$.

לשם זה נשתמש בכל ביס $P[Y = y|X = x] = \frac{P[Y = y|X = x] \cdot P[X = x]}{P[Y = y]}$, נכתוב $P[X = x|Y = y] = \frac{P[Y = y|X = x] \cdot P[X = x]}{P[Y = y]}$ כיוון ש- X רציף ו- Y בדיד. **לכל אורך הקורס, אם יש התפלגות רציפה נשים f במקומ P ונבעוד עם זה כך.** נניח בה"כ $\mu_0 > \mu_1$ ונקבל:

$$\begin{aligned} P[Y = 1|X = x] > P[Y = 0|X = x] &\Leftrightarrow \frac{f_X(x|Y = 1) \cdot P[Y = 1]}{f_X(x)} > \frac{f_X(x|Y = 0) \cdot P[Y = 0]}{f_X(x)} \\ &\Leftrightarrow \frac{f_X(x|Y = 1)}{f_X(x|Y = 0)} > \frac{P[Y = 0]}{P[Y = 1]} \Leftrightarrow \frac{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_0)^2}{2\sigma^2}}} > \frac{1-p_1}{p_1} \Leftrightarrow e^{-\frac{(x-\mu_1)^2}{2\sigma^2} + \frac{(x-\mu_0)^2}{2\sigma^2}} > \frac{1-p_1}{p_1} \\ &\Leftrightarrow \frac{(x-\mu_0)^2 - (x-\mu_1)^2}{2\sigma^2} > \log\left(\frac{1-p_1}{p_1}\right) \Leftrightarrow -2x\mu_0 + \mu_0^2 + 2x\mu_1 - \mu_1^2 > 2\sigma^2 \log\left(\frac{1-p_1}{p_1}\right) \\ &\Leftrightarrow 2x(\mu_1 - \mu_0) > 2\sigma^2 \log\left(\frac{1-p_1}{p_1}\right) + \mu_1^2 - \mu_0^2 \Leftrightarrow x > \frac{\sigma^2 \log\left(\frac{1-p_1}{p_1}\right) + \mu_1^2 - \mu_0^2}{(\mu_1 - \mu_0)} = \frac{\sigma^2 \log\left(\frac{1-p_1}{p_1}\right)}{(\mu_1 - \mu_0)} + \frac{\mu_1 + \mu_0}{2} \end{aligned}$$

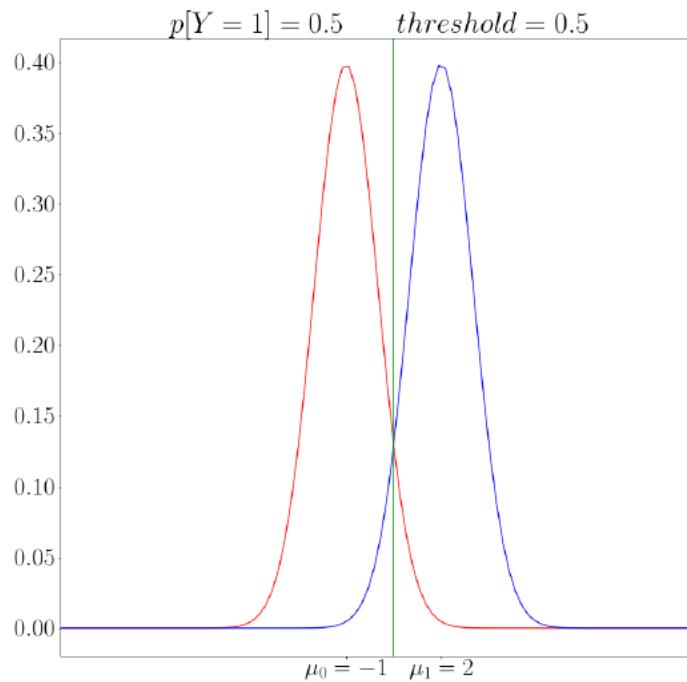
לסיכום, הנחנו שלכל תיוג שלנו $i = Y$ מתקיים $N(\mu_i, \sigma^2)$. הבנו שהתנאי שלנו למסוג האופטימי הוא $[x = x] > P[Y = 0|X = x]$:

$$x > \frac{\sigma^2 \log\left(\frac{1-p_1}{p_1}\right)}{(\mu_1 - \mu_0)} + \frac{\mu_1 + \mu_0}{2}$$

אחרת $y = 0$.

נסתכל על כמה מקרים.

1. עברו $\frac{1}{2} = p_1$ נקבל $x > \frac{\sigma^2 \log(\frac{1-0.5}{0.5})}{(\mu_1 - \mu_0)} + \frac{\mu_1 + \mu_0}{2} = \frac{\mu_1 + \mu_0}{2}$ כלומר הסוף הוא בדיק האם המוצע בין התוחלות.

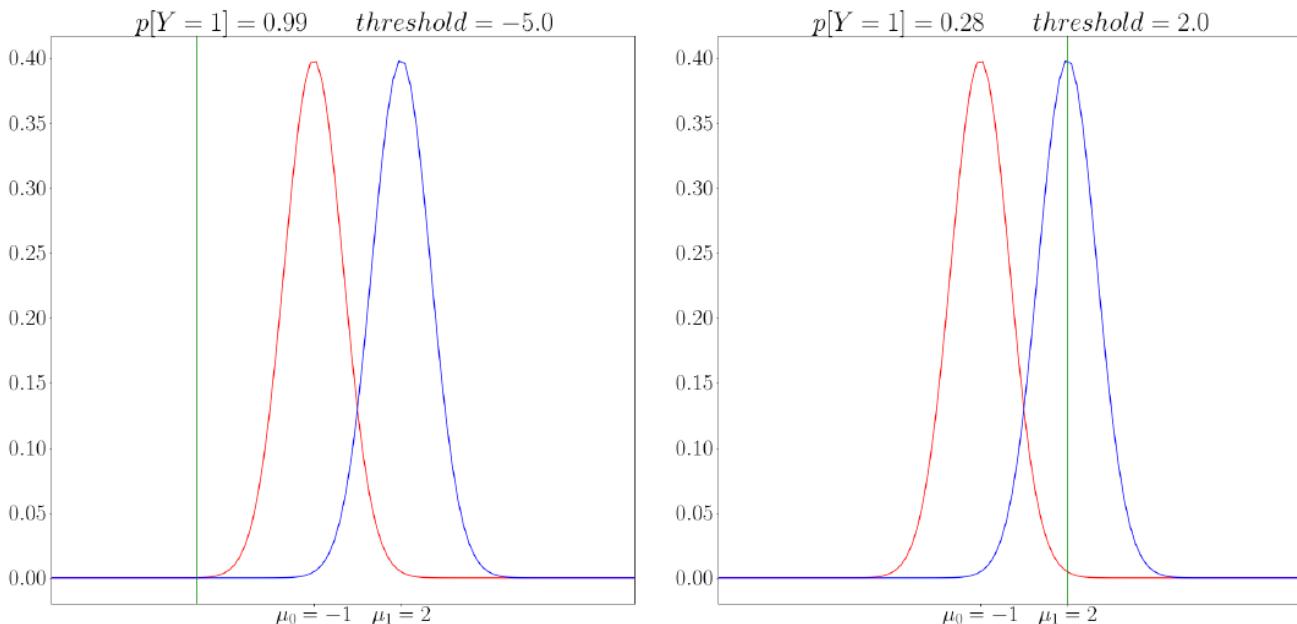


2. עברו $\mu \approx \mu_0$ (במיעט אותה הטעות, ונניח X – מספר האחים שלך, Y – לסוג אם אתה בן או בת, אין קורלציה). המבנה הוא מספר חיובי מאוד קטן, אז הגורם הראשון יהיה הרובה יותר דומיננטי בין המוחוברים, והוא נקבע לפי סימן ה-log.

a. אם $\frac{1}{2} > p_1$ אז $\lim_{\mu_0 \rightarrow \mu_1} \frac{\sigma^2 \log(\frac{1-p_1}{p_1})}{(\mu_1 - \mu_0)} < 0$ ו- ∞ $\log\left(\frac{1-p_1}{p_1}\right)$ אז התנאי לסוג 1 יהיה $-x > a$. זה הגיוני אם הפסיכיר שלנו לא נתונים עוד מידע על התיאוג, מותר לנו להסתכל רק על ה-prior, ולא על x . זה מקרה מבודן שבו x לא נתונים לנו מידע עליו.

b. אם $\frac{1}{2} < p_1$ אז $\lim_{\mu_0 \rightarrow \mu_1} \frac{\sigma^2 \log(\frac{1-p_1}{p_1})}{(\mu_1 - \mu_0)} > 0$ ו- ∞ $\log\left(\frac{1-p_1}{p_1}\right) > 0$ ותמיד נסוויג 0.

במקרה שבו ה-prior נושא יותר לתיאוג Y , הסוף זו ימינה. בכיוון השני של נתיחה לתיאוג Y הסוף זו ממש שמאליה.





חסמי ריבוב:

במציאות אנחנו לא יודעים את ההתפלגות המשותפת, והדבר הבci טוב שאנחנו יכולים הוא לדוגם הרבה דגימות ב"ת של משתנה שמעוניינו אותנו, ולקוטה שהמבחן הזה מייצג את המציאות בצורה כזו שהיא טוב על המבחן, יהי טוב על המציאות. נרצה ראשית להבין כמה מ"מ יכול להיות וחוק מהתוחלת שלו? כמה דגימות צריך לדוגם עם מ"מ כדי שניהה בטוחים שהממוצע של הדגימות האלו קרוב לתוחלת שלו?

לרוב יהיו לנו X_i משתנים שהם IID שייהו את הדגימות שלנו, ונתנו ע"י ממוצעם $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. הממוצע הוא גם משתנה מקרי, אין סיבה לצפות שהוא יצא בדיק התוחלת, והוא נסמן $\mathbb{E}[\bar{X}] = \mu$. נחשב על ניסוי של ח הטלות מטבע, באשר הסיכוי לקבל "עז" הוא הפרמטר k המוגדר עבור מ"מ (μ) $P[\bar{X} \geq p + \varepsilon] \leq \frac{\mathbb{E}[\bar{X}]}{p + \varepsilon}$. נstable על הממוצע של המשתנים $X_i \sim B(p)$ ועל התוחלת שלו $\mu = \mathbb{E}[\bar{X}]$. נרצה לחשב את הסיכוי **שהממוצע סופה מהתוחלת ביותר מ-** μ , כלומר $P[\bar{X} \geq p + \varepsilon]$. לפי חוק המספרים הגדולים, ככל שמספר הדגימות שלנו גדול, כך הממוצע קרוב לתוחלת. מעניין אותנו באיזה קצב הוא מתקרב.

חסם	נוסחה	בסיס הטלת מטבעות
מרקוב (Markov)	$P[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$ עבור מ"מ $0 \leq X$ ובכל $a > 0$: $\mathbb{E}[\bar{X}] = \mu$ ונקבל:	$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ $P[\bar{X} \geq p + \varepsilon] \leq \frac{\mathbb{E}[\bar{X}]}{p + \varepsilon} = \frac{p}{p + \varepsilon}$ חומר לא תלוי בכמות הדגימות n , אז אין לו משמעות מעניינת עבורנו (ככל שנגדל את n).
צ'בישב (Chebyshev)	$P[X - \mathbb{E}[X] \geq b] \leq \frac{\sigma^2}{b^2}$ נכיח שהשונות סופית: $\infty < \sigma^2$.	$.Var(\bar{X}) = Var\left(\frac{1}{n} \sum_{i=1}^n \bar{X}_i\right) = \frac{1}{n^2} \sum_{i=1}^n Var(\bar{X}_i) = \frac{p(1-p)}{n}$ לכן נקבל: $P[\bar{X} - p \geq \varepsilon] \leq \frac{p(1-p)}{n\varepsilon^2} = O\left(\frac{1}{n}\right)$ הסתברות קטנה רק לנארית ב- ε .
הופдинג (Hoeffding)	נכיח שיש לנו א"מ X_i ב"ת כך שמתקיים $P[a \leq X_i \leq b] = 1$ $P[\bar{X} - \mu > \varepsilon] \leq 2e^{-\frac{2\varepsilon^2}{(b-a)^2 n}} = e^{-\Omega(n)}$	באו הסתברות קינה אקספוננציאלית ב- ε .

חסם על השגיאה: נניח שיש לנו א"מ IID כך שערכיהם בין 0 ל-1. נגדיר את הממוצע $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ ואת התוחלת $\mu = \mathbb{E}[\bar{X}]$. נקבע δ כלשהו קטן מאוד (רמת הביטחון, חסם על הסיכוי לשגיאה), מה **נוכל לומר על השגיאה** $|\mu - \bar{X}| < \delta$ (ה- ε , כמה אנחנו רוחקים מהתוחלת) **בפונקציה של n** בסיסי גבוהה של $\delta - 1$? לפי הופдинג נרצה לדרוש שהסתטיה תהיה קטנה מ- δ (הסתברות של המאrove הרע תהיה לכל היוטר δ):

$$P[|\bar{X} - \mu| \geq \varepsilon] \leq 2e^{-\frac{2\varepsilon^2}{(1-0)^2 n}} = 2e^{-2n\varepsilon^2} \leq \delta$$

$$\Leftrightarrow e^{-2n\varepsilon^2} \leq \frac{\delta}{2} \Leftrightarrow -2n\varepsilon^2 \leq \log\left(\frac{\delta}{2}\right) \Leftrightarrow \varepsilon^2 \geq \frac{\log\left(\frac{2}{\delta}\right)}{2n} \Leftrightarrow \varepsilon \geq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}$$

נקבל כי בהסתברות של $\delta - 1$ (הסתברות גבוהה):

$$|\bar{X} - \mu| \leq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}} = O\left(\frac{1}{\sqrt{n}}\right)$$

דוגמה (על ההסתברות): נתיל 100 = n מטבעות באשר $p = 0.5$, מה ההסתברות של לפחות $\frac{3}{4}$ מהם יהיו "עז"? אינטואיטיבית זו הסתברות אפסית. נשים לב שכאנו $\varepsilon = 0.25$ (ε הtoutלת היא 0.5 ואנחנו רוצים לסטות ב-0.25).

- **מרקוב:** $P[\bar{X} \geq p + \varepsilon] \leq \frac{p}{p + \varepsilon} = \frac{0.5}{0.75} = \frac{2}{3}$
- **צ'בישב:** $P[\bar{X} \geq p + \varepsilon] \leq P[\bar{X} \geq p + \varepsilon] + P[\bar{X} \leq p - \varepsilon] = P[|\bar{X} - p| \geq \varepsilon] \leq \frac{p(1-p)}{n\varepsilon^2} = \frac{0.5^2}{100 \cdot 0.25^2} = 0.04$
- **הופдинג:** $P[\bar{X} \geq p + \varepsilon] \leq 2e^{-2n\varepsilon^2} = 2e^{-2 \cdot 100 \cdot 0.25^2} \approx 7.4 \cdot 10^{-6}$

דוגמה (על השגיאה): בעת נדרש מטבע יש לבצע? כאשר $|\bar{X} - p| \geq 0.01$ ≤ 0.1 , והסיכוי הוא 0.9 (כלומר $\delta = 0.1$). לפי הופдинג: $P[|\bar{X} - p| \geq 0.01] \leq 2e^{-2 \cdot 0.01^2 n} \leq 0.1 \Leftrightarrow n \geq \frac{\ln\left(\frac{100}{5}\right)}{2 \cdot 0.01^2} = 14979$

תיאוריה של למידה מפוקחת

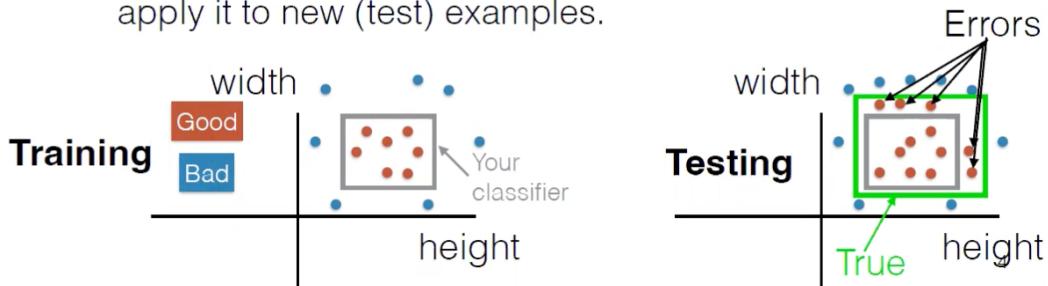
מושגים בסיסיים

מבוא:

אנחנו רוצים ללמידה פונקציה שmaps בין X ל- Y , דרך דוגמאות. נתמקד במקרה שבו Y הוא בינארי (0 או 1, לא ספאם/ספאם). הפונקציה היא כל סיווג שננסמן באותו מצבו הבא: $\{0, 1\} \rightarrow \mathcal{Y}$. נרצה לבצע תהליכי שבו אנו **לומדים את המסוג** זהה. המסוג הזה צריך להיות טוב – על דברים **שלא ראיינו** במהלך האימון.

דוגמה: יש לנו אוסף של פרטיות בעיר, חלק רוחב וגובה ורחב, ככלומר $\mathbb{R}^2 = \mathcal{X}$. אנחנו בוחרים מסוג כלשהו שmbased על training data, test data. ואז מימושים אותו על דוגמאות חדשות. אנחנו רוצים לדעת האם הפרטיה רעליה או לא: $\{1, 0\} = \mathcal{Y}$. בתמונה ניתן לראות שב-test גילינו שלא סייגנו בצורה נכונה כמה פרטיות שהן Good, לא אכלנו מהן.

- You pick a classifier based on training data and then apply it to new (test) examples.



האלגוריתם הבci נאיבי כדי לסואג נקודה הוא להחזיר את ה-label של השכן הקרוב ביותר. יש וריאציה של k-nearest neighbors: מודול פשטוט, עובד טוב במקרים רבים. חסרונות: יש להציג מטריקה למרחק, מסתבר במדדים גובהים, יקר חישובית.

MORECOMPLEX: ללמידה מסוג זו אינה בעיה פשוטה. במקרה של $\mathcal{X} = \mathbb{R}^2$ ניתן לבנות מסוג שהוא קו ישר (linear) או קו עקום (curved). בבעיות ורגסיה (regression) הפלט לא חייב להיות בינארי, אפשר למצוא מספר כללים מתאימים כדי לקבל את התשומות הנכונות ב- \mathcal{X} , מספר מסוגים. בקיצור, מדובר בעסק מורכב.

שאלה המפתח: מה אנחנו יכולים להגיד לגבי הטעות ב-test? זה תלוי ב:

- משפחת המסוגים שבה אנו משתמשים.
- بما data יש לנו. לעיתים שיפור בכמות ה-data יכולה להשפיע מאוד על הביצועים הסופיים.
- מה המסוג האמיתי.

מושגים:

בצד נבנה מודל למידה? אלו הם אובי הרכיבים:

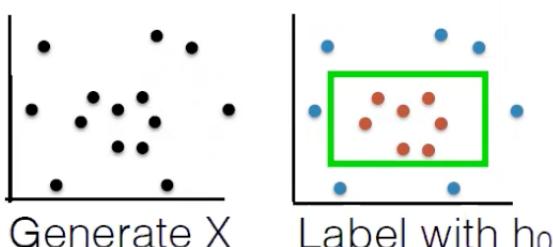
חישוב שההתקפות שמייצרת את ה-data, train-data, תהיה זהה לו שמייצרת גם את ה-test-data. ללא ההנחה הזו (בלומר יש mismatch בין train-test), לא נלמד בצורה נכונה/רלוונטיות. אפשר לחשב על P אליו הוא ראיית מיצרת את X , ואז את Y בהינתן X (לעתים התהילה זהה יכול להיות טובסטי, לא דטרמיניסטי. יכול להיות שהוא 2 פרטיות עם אותו גובה ורוחב כך שאחת רעליה והשנייה לא).	ההתפקידות "האמתית" שמייצרת צמדים (Y, X)
הפונקציות/המסוגים שهما מיתן לבחור עד כמה ביצענו טוב בזמן האימון, השגיאה האמפירית.	חלוקת ההיפותזה
עד כמה ביצענו טוב בזמן האימון, השגיאה האמפירית.	שגיאת האימון (training error)
התוצאה בזמן המבחן	שגיאת האמת (true error)

נתון לנו **training set** של n צמדים של קלט-תיאוג: $(X_1, Y_1), \dots, (X_n, Y_n) = S$, לעתים מסומן S . זהו אוסף של $2n$ מקרים.סביר להניח שהצמדים האלה הם ב"ת, והוגלו באופן IID (Independent, identically distributed), נסמן את מה שמייצר את S (הדוגמא) בטור P_n . חיבורים להניח משווה על הדוגמה, אחרת כל הדוגמאות יכולות להיות עותקים של אותו הדבר.

חלוקת ההיפותזה: הפונקציות שאנחנו רוצים לשקל ולבחור מתוכן את המסוג שלנו. אינטואיטיבית, לא נרצה שהפונקציה תהיה מושכנת ומצווצמת מדי ("עשירה" מספיק כדי להכיל h "מדוייק"), ולא שרירותית ורחבה מדי (מרחיב היפותחות גדול מדי יכול להיות רע עבור הלמידה וליצור overfit). קר שזה לא מכיל דוגמאות שלא ראיינו. אפשר לסייע את המידע בכל דרך אפשרי, ولكن אפשר לסוג אותו רנדומלית). דבר זה מכונה "[bias-variance tradeoff](#)". נסמן אתחלוקת ההיפותזה האפשריות ב- \mathcal{H} , והיפותזה מסוימת תהיה $H \in \mathcal{H}$. נקבע את הקבוצה זו לפני שאנחנו רואים את הדטא.

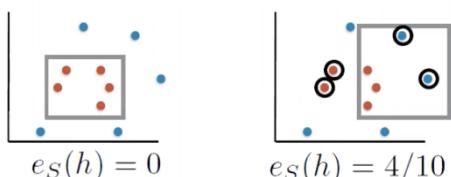
דוגמאות לסוגים של היפותזות:

- מלבדים מלבנים מלבניים לציריהם: $\mathcal{X} = \mathbb{R}^d, \mathcal{H} = \{h(x, R) | R - rectangle in the plane\}$
 - חיצית המרחב לשניים על ידי יישר: $\mathcal{X} = \mathbb{R}^d, \mathcal{H} = \{h(x, w, b) = sign(\langle x, w \rangle - b) | w \in \mathbb{R}^d, b \in \mathbb{R}\}$. מלבנים כל וקטור x במשקל w ומחסירים b (bias).
 - קבוצת המיפויים שמקבלים מסידור שונה של משקولات עבור ארביטקטורה של רשת נירונית.
 - נניח שיש לנו מידע בינארי: x_d, \dots, x_1 . ניתן לבנות פונקציות OR: $x_5 = (x_1 \vee x_2 \vee \dots \vee x_d)$, פונקציות בצורת CNF וכו'.
- $h_0 \in \mathcal{H}$** – התפלגות האמת P מיצרת את \mathcal{H} באמצעות **היפותזה שייכת למחלקה שמנא לומדים: \mathcal{H}** .
- המקרה הזה נכון לאנליה, והוא יותר סביר במקרה \mathcal{H} יותר גדולה. נסמן את כל ההתפלגותים שהן ב- \mathcal{H} .



- An example of a P that is realizable for axis aligned rectangles.

21



(Empirical) Training Error: בהינתן h , אפשר למדוד כמה הצלחנו על S (ה-training data) בזמן האימון. נסמן את השגיאה האמפירית על המדגם S של ההיפותזה h ב- $e_s(h)$. בדוגמה מימין ניתן לראות כי יש לנו 2 שגיאות בתוך הריבוע (נקודות שסוווגנו לא נכון), ו-2 שגיאות מחוץ לריבוע (2 נקודות שפספסנו), לכן קיבל 4 שגיאות מתוך 10.

- באופן יותר פורמלי: $(Y_i, X_i) \in S$ כאשר $e_s(h) = \frac{1}{n} \sum_{i=1}^n \Delta_{zo}(h(X_i), Y_i) = \begin{cases} 0 & y = \hat{y} \\ 1 & y \neq \hat{y} \end{cases}$ היא פונקציית ה-zero-one loss שלנו, אם h -labels זהים היא 0, ואם הם שונים מקבלים קנס של 1. היא סופרת האם יש טעות או לא עבור המקרה הנוכחי משווים. נסכום את כל השגיאות ונחלק ב- n .
- במקרה ה- P , יש תמיד $h_0 \in \mathcal{H}$ שעובד ייש שגיאה 0 (זו שתיגה את P), בפועל יש הרבה כאלה.

(Test/Expected/Population) True Error: בהינתן h , אפשר לבדוק אותה מול דוגמאות מ- P , בלוגר אנחנו בודקים את התוחלת של ה-loss עבור דוגמאות מ- P . בזמן אימון, דגמנו קצת דוגמאות מ- P , ובזמן test מעבינים אותן מה היא התוחלת האמיתית של השגיאה. **זה באמת הדבר שמעניין אותנו**, ונרצה להציג למצב זהה יהיה נמור בכל האפשר:

$$e_P(h) = \mathbb{E}_P[\Delta_{zo}(h(X), Y)] = \sum_{(X,Y) \in \mathcal{X} \times \mathcal{Y}} P[X, Y] \cdot \Delta_{zo}(h(X), Y)$$

פרשניות:

- זהה תוחלת השגיאה כאשר מסוגים בעזרת h מידע שמאגי מ- P .
- עבור משתנה מקרי ביןארי Z , $\mathbb{E}[Z] = \sum_{z \in \{0,1\}} z \cdot P[Z = z]$, لكن נגדיר משתנה ביןארי חדש שאומר האם h טעה על הצמד (X, Y) , יש לו 2 ערכים, 0 או 1: $Z = [h(X) \neq Y]$. כלומר:

$$e_P(h) = \mathbb{E}_{Z(h)=\Delta_{zo}}[Z(h)] = \text{binary var. } P[Z(h) = 1] = P[h(X) \neq Y] \text{ by def.}$$

כלומר, שגיאת האמת של h היא ההסתברות שנגזריל צמד (X, Y) ש- h תטעה עליו. אם $0 = e_P(h)$ זה אומר שאין אף צמד שנחקרו טועים עליו.

- גאומטרית: נניח שההתפלגות האמת P מסווגת על פי המלבן הירוק, וה- h היא המלבן הכלול. מה שבאופןם הם האזוריים שבינם h טועה. האזורי הלבן הוא האזורי שבו h צודקת. $e_P(h)$ היא ההסתברות להגריל נקודה מתחומי האזורי האפור.



24



אלגוריתם ERM

שגיאות אמפיריות (train) ושגיאות האמת (true)

מה הקשר בין השגיאה האמפירית שלנו (עד כמה ההיפותזה עובדת טוב על מוגם האימון) לבין שגיאת האמת (עד כמה ההיפותזה עובדת טוב באמצעותות האמיתית)?

- הגדכנו את השגיאה האמפירית (training error). אפשר לחשב עליה בעל סכום של משתני ברגולרי. הערכים האפשריים של Δ_{zo} הם 0 או 1, והוא מושתנה מקרי כי הוא פונקציה של שני משתנים מקרים X, Y . כמובן, המשתנים $\{0,1\}$ הם דגימות IID של $Z(h) = \Delta_{zo}(h(X), Y)$.
- בollowmore, השגיאה האמפירית עבור מוגם בגודל n היא ממוצע של n משתני ברגולרי שווים התפלגות:

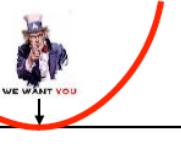
$$e_S(h) = \frac{1}{n} \sum_{i=1}^n \Delta_{zo}(h(X_i), Y_i) = \frac{1}{n} \sum_i Z_i(h)$$

אם יודעים בברכיו (h) $\mathbb{E}[Z(h)] = P[h(X) \neq Y] = e_P(h)$, ולבן:

$$\mathbb{E}[e_S(h)] = \mathbb{E}\left[\frac{1}{n} \sum_i Z_i(h)\right] = \frac{1}{n} \sum_i \mathbb{E}[Z_i(h)] = \frac{1}{n} \sum_i \mathbb{E}[Z(h)] = \frac{1}{n} n \cdot \mathbb{E}[Z(h)] = e_P(h)$$

מטרת הלמידה היא למצוא h שambil את $e_P(h)$. אבל אנחנו לא יודעים את P !^{*} לכן הדבר הכי טוב שאנו יכולים לעשותו הוא לקבל עבור ה- $e_P(h)$ true error שהוא $e_P^* = \min_{h \in \mathcal{H}} e_P(h)$. זה גם מכונה approximation error, כיון שהוא מייצג את היכולת של מחלוקת ההיפותזה לשער את הסתברות P . המטרה שלנו בהמשך תהיה לבנות אלגוריתם למידה, לצד הבוחות תאוריות כדי שנוכל לשער כמה אנחנו קרובים ל-^{*}"תוצאה הכי טובה שיבולנו לקבל אילו ידענו את P ". בollowmore ל-^{*} e_P .

מבחן קונספטואלי אנחנו מוצאים את התהיליך הבא באימון ולמידה:



True Error(h) = $e_P(h)$

- מקבלים training set עם מידע מותג S שנದם IID מתוך P .
- משתמשים באלגוריתם A בליוויו מקבל את S ומחדיר ההיפותזה $\mathcal{A} \in \mathcal{H}$.
- שגיאת האמת של ההיפותזה שלמדו היא $e_P(\mathcal{A}(S))$.

חצים שהתוכאה שקיבלנו ($\mathcal{A}(S)$) תהיה קרובה ל-^{*} e_P .

אם $e_S(h)$: האלגוריתם מקבל את המוגם S , והוא יתן לנו את ההיפותזה שמשמעותה את (h) (Empirical Risk Minimization) ERM. נרצה שייהו לנו אמירות על הביצועים בעולם האמיתי ($e_P(ERM(S))$), הוא מוגדר באופן הבא:

הצדקה לא פורמלית ל-ERM: אלגוריתם ERM מבוסס על האמונה שהשגיאה האמפירית ($e_S(h)$) "קרובה" לשגיאת האמיתית ($e_P(h)$). אינטואיטיבית, נראה סביר שככל שיש יותר דוגמאות הם יותר קרובים. בollowmore, נראה ששה קורה כאשר $\infty \rightarrow n$. נשים לב כי h קבוע, השגיאה ($e_S(h)$) תליה ב-set-training. מתקיים:

$$\mathbb{E}[e_S(h)] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \Delta_{zo}(h(X_i), Y_i)\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_P[\Delta_{zo}(h(X_i), Y_i)] = \frac{1}{n} \sum_{i=1}^n e_P(h) = e_P(h)$$

מתקיים לגבי השונות של:

$$Var(e_S(h)) = Var\left(\frac{1}{n} \sum_{i=1}^n \Delta_{zo}(h(X_i), Y_i)\right) = \frac{1}{n^2} \sum_{i=1}^n Var(\Delta_{zo}(h(X_i), Y_i)) = \frac{n}{n^2} \cdot Var(\Delta_{zo}(h(X), Y)) = \frac{con.}{n} \xrightarrow{n \rightarrow \infty} 0$$

מכך אנחנו מסיקים שהמ"מ ($e_S(h)$) הוא בעל תוחלת ($e_P(h)$) ושונות שווהpta ל-0 במספר הדוגמאות n שואף לאינסוף. המשער שלים הולך ונhea יותר מדויק. אמנם, מה שעשינו פה לא כל כך מייצג את המציאות.



בעיות:

$$\begin{aligned} \text{לפנינו}: & f_n(x) \rightarrow f(x) \text{ כאשר } x \in \mathcal{X} \text{ ו } n \in \mathbb{N} \text{ נסוברים.} \\ \text{לפנינו}: & f_n(x) \rightarrow f(x) \text{ כאשר } x \in \mathcal{X} \text{ ו } n \in \mathbb{N} \text{ נסוברים.} \\ \text{לפנינו}: & |f_n(x) - f(x)| < \epsilon \Leftrightarrow N > n \text{ ו } \forall x \in \mathcal{X} \text{ נסוברים.} \\ \text{לפנינו}: & f_n(x) \rightarrow f(x) \text{ כאשר } x \in \mathcal{X} \text{ ו } n \in \mathbb{N} \text{ נסוברים.} \\ \text{לפנינו}: & |f_n(x) - f(x)| < \epsilon \Leftrightarrow N > n \text{ ו } \forall x \in \mathcal{X} \text{ נסוברים.} \end{aligned}$$

1. מטפל בהיפותזה ספיציפית – הטיעון שננתנו מטפל ב- \mathcal{H} נקודתי, בעוד שאנחנו צריכים שהשגיאה האמפירית והאמיתית יהיו קרובות עבור כל היפותזה במחלקה (יכול להיות שגם היפותזה הש- \mathcal{H} שלהם יהיה רחוק מאוד מ-0, לא משנה כמה דוגמאות יש לנו).
 2. במקרה של **training data סופי** – במקרה ש- \mathcal{H} סופי, ברור שהשגיאה האמפירית יכולה להשתנות מדגימה לדגימה, והוא לרוב שונת מהשגיאה האמיתית.
- ERM במקרה ש- \mathcal{H} סופי:** יש מספר שאלות שאנו מתחכמים בהן:
- כמה דוגמאות מאנחנו צריכים כדי להבטיח שההיפותזה שנלמדת בעלת error low true?
 - متى מובטח ש- $\text{ERM}(\mathcal{H})$ יעבוד?
 - איך בוחרים את המחלקה \mathcal{H} ואיזה tradeoffs נדרש לבצע?

למידה במקרה PAC עם Realizable

Realizable Setting:案 אナンכו נניכ בקיימת היפותזה h^* שעבורה השגיאה האמיתית היא אפס: $0 = \mathbf{e}_P(h^*) \in \mathcal{H}$. מוכנת המודל של הטבע, מיצרת דוגמאות ומסתמשת בהיפותזה זו כדי ליצר label. לעומת P מייצרת דוגמאות בצורה הבאה:

1. עבור התפלגות P_X מעל \mathcal{X} .
2. נקבע $Y \sim h^*(X)$ עבור $\mathcal{H} \in h^*$ שמתיגת את X.

זה נקרא ה-realizable setting. נשים לב שמתקיים:

- עבור קלט X תמיד תתקבל התוצאה $(X)^* = h^* = Y$, כלומר **P במובן זהה דטרמיניסטי (Y בהינתן X)**.
 - לכל דוגימה $P \stackrel{\text{IID}}{\sim} S$ מתקיים $\mathbf{0} = \mathbf{e}_S(h^*)$, כי לפי הגדרה:
- $$e_S(h^*) = \frac{1}{n} \sum_{i=1}^n \Delta_{zo}(h^*(X_i), Y_i) = \frac{1}{n} \sum_{i=1}^n \Delta_{zo}(h^*(X_i), h^*(X_i)) = 0$$
- ביוון ש- ERM ממציא את השגיאה האמפירית נסיק:
- $$0 \leq e_S(\text{ERM}(S)) \leq e_S(h^*) = 0 \Rightarrow e_S(\text{ERM}(S)) = 0$$

למידות PAC: המטרה שלנו היא להבטיח שהשגיאה האמיתית (S) היא נמוכה עבור מספר דוגמאות גדול מספיק. ההבטחה תהיה בקירות (approximate): במקרה $\text{P}(\text{ERM}(S)) \leq \epsilon$ (ולא ניתן לצפות יותר טוב מזה), $w.h.p$:probably (w.h.p) δ – חדש הסתברות δ – 1 עבור δ קטן). פורמלית, נדרש $N \in \mathbb{N}(\epsilon, \delta)$ -realizable, שאם P היא P אז יתקיים:

$$P[e_P(\text{ERM}(S)) \leq \epsilon] \geq 1 - \delta \Leftrightarrow P[e_P(\text{ERM}(S)) > \epsilon] \leq \delta$$

אם קיים זהה (δ, ϵ) נאמר ש- \mathcal{H} היא PAC (probably-approximately-correct) learnable via ERM. התשובה לשאלת הוא תלוי במחלקה \mathcal{H} .

- במקרים פשוטים, המשמעות היא שניתן להבטיח שהשגיאה נמוכה כרצוננו, בהסתברות גבוהה כרצוננו, אם נהייה מוכנים לשלים בכך שישו מספיק דוגמאות.
- באופן כללי ניתן להרחיב לאלגוריתם כלשהו (לאו דווקא ERM), ולומר שהמחלקה היא **PAC learnable**.
- $N(\epsilon, \delta)$ נקרא **sample complexity**.

משפט: תהי \mathcal{H} מחלקה סופית בעלת $|\mathcal{H}|$ היפותזות שונות. לכל \mathcal{P} , \mathcal{H} סופית, P הוא PAC learnable אם \mathcal{H} realizable, $\epsilon, \delta \in (0, 1)$, $n \geq \frac{1}{\epsilon} \cdot \ln\left(\frac{|\mathcal{H}|}{\delta}\right)$.

היפותזה קונסיסטנטית: היפותזה $\mathcal{H} \in \mathcal{H}$ היא **קונסיסטנטית (consistent)** עם S , אם $\mathbf{0} = \mathbf{e}_S(h)$. נשים לב ש- ERM תמיד מחייב היפותזה קונסיסטנטית (קיימת כזו שמתאימה בצורה מושלמת, ERM מחייב את השגיאה האמפירית המינימלית ולכן יחזיר אותה).

היפותזה רעה: היפותזה $\mathcal{H} \in \mathcal{H}$ שעבורה $\epsilon > e_P(h)$. לפי הגדרה: $[e_P(\text{ERM}(S)) \in \mathcal{H}_{bad}^\epsilon(P)]$ כאשר P כאשר נניח ש- $\mathcal{H}_{bad}^\epsilon(P)$ היא הקבוצה כל היפותזות הרעות, זה אובייקט דמיוני מבחיננו.



נשים לב כי הדריך היחידה שבה ERM יבחר היפותזה רעה, היא אם קיימת אחת כזו שהיא קונסיסטנטית עם ה-set train S:

$$P[ERM(S) \in \mathcal{H}_{bad}^\varepsilon(P)] \leq P[\exists h \in \mathcal{H}_{bad}^\varepsilon(P). e_s(h) = 0]$$

נשים לב שזה אי-שוויון ולא שווין, כי יש מצב שבו קיימת היפותזה רעה שكونסיסטנטית עם S, אבל ERM מעדיף היפותזה טובה. בפרט, ניעזר ב-**union bound**: ההסתברות של איחוד מאורעות (קיימת היפותזה), קטן או שווה לסכום ההסתברויות של המאורעות.

$$P[\exists h \in \mathcal{H}_{bad}^\varepsilon(P). e_s(h) = 0] = P\left[\bigcup_{h \in \mathcal{H}_{bad}^\varepsilon(P)} \{e_s(h) = 0\}\right] \underset{\text{union bound}}{\leq} \sum_{h \in \mathcal{H}_{bad}^\varepsilon(P)} P[e_s(h) = 0]$$

עבור היפותזה רעה כלשהו $\varepsilon > h$ מתקיים:

$$\begin{aligned} P[e_s(h) = 0] &= P[h(X_1) = Y_1, \dots, h(X_n) = Y_n] \underset{\{(X_i, Y_i)\} \text{ indep.}}{=} \prod_{i=1}^n P[h(X_i) = Y_i] = \prod_{i=1}^n P[h(X) = Y] \\ &= \prod_{i=1}^n (1 - P[h(X) \neq Y]) \underset{\text{by def.}}{=} \prod_{i=1}^n (1 - e_P(h)) = (1 - e_P(h))^n \underset{e_P(h) > \varepsilon}{\leq} (1 - \varepsilon)^n \underset{\forall \alpha > 0, 1 - \alpha \leq e^{-\alpha}}{\leq} e^{-n\varepsilon} \end{aligned}$$

נחזיר לביטוי הראשון ונקבל:

$$P[\exists h \in \mathcal{H}_{bad}^\varepsilon(P). e_s(h) = 0] \leq \sum_{h \in \mathcal{H}_{bad}^\varepsilon(P)} e^{-n\varepsilon} = |\mathcal{H}_{bad}^\varepsilon(P)| \cdot e^{-n\varepsilon} \leq |\mathcal{H}| \cdot e^{-n\varepsilon}$$

זה"ב קיבלנו שכל שאנחנו מגדים את זה, הסיכוי לבישלון הולך ויריד בקצב אקספוננציאלי. נדרש קטן מ-δ ונקבל:

$$P[e_P(ERM(S)) > \varepsilon] \leq |\mathcal{H}| \cdot e^{-n\varepsilon} \leq \delta \Leftrightarrow n \geq \frac{1}{\varepsilon} \ln \frac{|\mathcal{H}|}{\delta}$$

הערות:

1. ניתן להכפיל את המשפט לכל loss \mathcal{L} ולבב $\Delta(y, y) = 0 \forall y \in [0,1]$ label space
 2. אם מתייחסים ל- \mathcal{H} קבוע, ושאלים איזה שגיאה אפשר להבטיח δ – $n \geq \frac{1}{\varepsilon^2} \ln \frac{|\mathcal{H}|}{\delta}$
 3. קיבלנו ממש לקחת את $\varepsilon = \frac{1}{n} \ln \frac{|\mathcal{H}|}{\delta}$: sample complexity
- בפרט אפשר ממש לקחת את $\varepsilon = \frac{1}{n} \ln \frac{|\mathcal{H}|}{\delta}$.
- a. אם ε יורד אז $(\varepsilon, \delta) N$ גדול – צריך יותר דוגמאות בשבייל sha-he-*confidence* עליה.
 - b. אם δ יורד אז $(\varepsilon, \delta) N$ גדול – צריך יותר דוגמאות כדי שה-*confidence* עליה.
 - c. אם $|\mathcal{H}|$ יורד, אז $(\varepsilon, \delta) N$ יורד – מחלוקת קטנות נזנחות הבטחה טוביה יותר.

למידה במקורה עם Unrealizable

אם לא נניח realizability, יכול להיות שיתקיים: $e_p^*(h) < \min_{h \in \mathcal{H}} e_p(h)$. הכו טוב שאנחנו יכולים לקוות זה ללמידה $\mathcal{H} \in h$ שעבורו השגיאה האמיתית (h) לא גדולה בהרבה מהשגיאה האופטימלית e_p^* . נטפל במקורה הזה באמצעות התכנסות במידה שווה (uniform convergence).

למה: $|e_p(h) - e_p^*(h)| \leq 2 \sup_{h \in \mathcal{H}} |e_s(h) - e_p^*(h)|$. כאמור, השגיאה האמיתית של ERM פחותה השגיאה האמיתית האופטימלית, הוא לכל היותר פעמיים המרחק הגדול ביותר בין השגיאה האמפירית לשגיאה האמיתית.

הוכחה: נסמן ב- h^* היפותזה אופטימלית, כלומר $h^* \in \arg \min_{h \in \mathcal{H}} e_p(h)$ (אנחנו לא במקורה ה-*realizable*)

$$e_p(ERM(S)) - e_p^* = e_p(ERM(S)) - e_p(h^*) = e_p(ERM(S)) - e_s(ERM(S)) + e_s(ERM(S)) - e_p(h^*)$$

$$\begin{aligned} &: (e_s(ERM(S)) \leq e_s(h^*)) \\ &\leq e_p(ERM(S)) - e_s(ERM(S)) + e_s(h^*) - e_p(h^*) \leq |e_p(ERM(S)) - e_s(ERM(S))| + |e_s(h^*) - e_p(h^*)| \\ &\leq 2 \sup_{h \in \mathcal{H}} |e_s(h) - e_p(h)| \end{aligned}$$

מציאת חסם על סיבוכיות הדגימה:

עכשו נטפל בביטוי של הסופריםום, בדומה למה שכבר עשינו – זה הסיכוי שקיימת היפותזה כלשהי שמקיימת את התנאי, ונכתבו את זה מידע כמו קודם קודם בטור אויחוד מאורעות:

$$P \left[\sup_{h \in \mathcal{H}} |e_S(h) - e_P(h)| > \varepsilon \right] = P \left[\bigcup_{h \in \mathcal{H}} \{|e_S(h) - e_P(h)| > \varepsilon\} \right]$$

נקבע היפותזה $\mathcal{H} \in h$. מתקיים כי $e_S(h)$ הוא הממוצע של n עותקים IID של המ"מ $(Y, Z) := \Delta(h(X))$, שמקבל ערכים ב-[0,1]. עכשו נשתרמש **בחסם הופדינג**:

$$P[|e_S(h) - e_P(h)| > \varepsilon] = P \left[\left| \frac{1}{n} \sum_{i=1}^n Z_i - \mathbb{E}[Z] \right| > \varepsilon \right] \leq 2e^{-2n\varepsilon^2}$$

ניקח שוב **union bound** על \mathcal{H}

$$P \left[\bigcup_{h \in \mathcal{H}} \{|e_S(h) - e_P(h)| > \varepsilon\} \right] \underset{\text{union bound}}{\leq} \sum_{h \in \mathcal{H}} P[|e_S(h) - e_P(h)| > \varepsilon] \leq \sum_{h \in \mathcal{H}} 2e^{-2n\varepsilon^2} = 2|\mathcal{H}|e^{-2n\varepsilon^2}$$

משפט: עבור \mathcal{H} סופית, לכל $\delta \in (0,1)$ מתקיים $\delta \geq \frac{2|\mathcal{H}|}{\varepsilon^2} \ln \left(\frac{2|\mathcal{H}|}{\delta} \right)$

הוכחה: בעזרת הלמה והחסם שמצאנו נקבל:

$$P[e_P(\text{ERM}(S)) - e_P^* > \varepsilon] \underset{\text{lemma}}{\leq} P \left[2 \sup_{h \in \mathcal{H}} |e_S(h) - e_P(h)| \geq \varepsilon \right] = P \left[\sup_{h \in \mathcal{H}} |e_S(h) - e_P(h)| \geq \frac{\varepsilon}{2} \right] \leq 2|\mathcal{H}|e^{-\frac{n\varepsilon^2}{2}}$$

נחשב על ידי δ ונקבל:

$$2|\mathcal{H}|e^{-\frac{n\varepsilon^2}{2}} \leq \delta \Leftrightarrow n \geq \frac{2}{\varepsilon^2} \ln \left(\frac{2|\mathcal{H}|}{\delta} \right)$$

הערות:

1. אפשר להראות שההתוצאה תופסת לפחות loss ε .
2. כאן $O(\varepsilon N) = O\left(\frac{1}{\varepsilon^2}\right)$ במקרה ה-*realizable*. זה המחיר שאנו משלמים על אגנוסטיות.
3. הערת 3 מקודם תופסת באותה הצורה.

תרגול 2 (למידות PAC)**למידות PAC:**

בلامידה מפוקחת המטריה שלנו היא ללמידה מסווג טוב $\mathcal{Y} \rightarrow \mathcal{X}$: h . אנחנו מנהנים את האלגוריתמים המתאימים באמצעות PAC. P היא התפלגות מעל $\mathcal{Y} \times \mathcal{X}$ (זה העולם שלנו). נזכיר בהגדרות:

- סט אימון: tuples של דוגמאות מתוצאות $S = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$ מההתפלגות P . נתעסק בסיווג בינהרי.
- נתיחס לתיאוג 1 בחיבור, ותיאוג 0 כשליל.
- מחלוקת היפותזה: אנחנו מוגבלים למחלוקת נתונה $\mathcal{Y} \rightarrow \mathcal{X} \subseteq \mathcal{H}$. ברגע נגדיר $\{0,1\} = \mathcal{Y}$.
- התפלגות P היא *realizable* אם קיים מסווג $h_0 \in \mathcal{H}$ שמתאים לתיאוג הנכון: $(X, Y) = h_0(X)$.
- ביחס לכל היפותזה אנחנו מודדים שתי שגיאות:
 - שגיאת אמת (בטסט): $e_P(h) = \mathbb{E}_P[\Delta_{zo}(h(X), Y)]$.
 - שגיאה אמפירית (באימון): $e_S(h) = \frac{1}{n} \sum_{i=1}^n \Delta_{zo}(h(X_i), Y_i)$.

אלגוריתם למידה A מקבל בקלט סט אימון S וgiveשה למחלוקת היפותזה \mathcal{H} , הוא מחזיר היפותזה מהמחלקה $\mathcal{H} \in A(S)$. המטריה שלנו, שהוא יחזיר לנו היפותזה שעושה מינימיזציה לשגיאת האמת e_P . אבל אנחנו לא יודעים אותה כי אין לנו את P . לכן אנחנו נSEARCH**להתחרות בשגיאה האופטימלית** (ששgiaת האמת המינימלית שהיפותזה כלשהי מושגתה לנו): $\min_{h \in \mathcal{H}} e_P(h) = e_P^*$. במקרה ה- realizable , הערך הזה הוא 0, כי המשוו שעליו שגיאת האמת היא 0 נמצא בחלוקת.



לכן, אנחנו מזמנים כי אלגוריתם שיעשה **מינימיזציה לשגיאה האמפירית (על המדגם - סט האימון)**, יחויר לנו מסוג שהוא בעל שגיאת אמת שתיהה קרובה לשגיאה האופטימלית. לגישה זו קראנו ERM שעשוה זאת: $h \in \arg \min_{h \in \mathcal{H}} e_S(h)$.

למיצות PAC (Realizable): יש לנו מחלוקת \mathcal{H} . נאמר **שהיא למידה PAC** ע"י אלגוריתם A שלומד אותה, אם קיימת פונקציית דגימה שמקבלת $(0,1) \in \delta, \epsilon$ ומחייבת לנו מספר מסוים של דוגימות, כך שכל התפלגות realizable בלשאלה P, אם נריץ את A על סט אימון S שמכיל לפחות $\delta, \epsilon \geq N$ דוגימות מ- \mathcal{P} :

- מתקיים: $\delta \leq \epsilon > (e_P(A(S)) - e_P^*)$.
- הוכחנו בשיעור שבعزורת ERM זה מתקיים עבור: $n \geq \frac{1}{\epsilon} \ln \left(\frac{|\mathcal{H}|}{\delta} \right)$

لمיצות PAC (Agnostic): במקרה זה לא נכון ש- \mathcal{P} realizable, לכן יכול להתקיים $0 > e_P^*$, וכך נקבל כי:

- מתקיים: $\delta \leq \epsilon > (e_P(A(S)) - e_P^*)$.
- הוכחנו בשיעור שבعزורת ERM זה מתקיים עבור: $n \geq \frac{2}{\epsilon^2} \ln \left(\frac{2|\mathcal{H}|}{\delta} \right)$

מחלקה היפותזות סופית – פונקציות OR:

נקבע $\mathcal{H}_{OR} = \{h: \mathcal{X} \rightarrow \mathcal{Y} | h = \beta_1 \vee \dots \vee \beta_m\} = \{0,1\}^m$. נתבונן במחלקה היפותזות הבאה $\{h = \beta_1 \vee \dots \vee \beta_m | \beta_i \in \{0,1\}\}$. נראתה **הוכחה שהיא למידה PAC** בנסיבות realizable. נראה **אלגוריתם**, ופונקציה $N(\epsilon, \delta)$.

נשים לב כי הקלט שלנו הוא וקטורים בעלי m קואורדינטות x_1, \dots, x_m . הפלט שלנו הוא OR על פני m משתנים כאשר כל משתנה x_j וריאציה על ערך של קואורדינטה מסוימת. במשתנה j יכול לקבל אחד מהVALUES: $\{0, 1, x_j - 1, x_j\}$. בולמר, או את הבנייה x_j , או את השילילה של x_j , או $x_j - 1$, או להתעלם ממנו (0), או לחתוך אותו תמיד (1). למשל בהינתן הוקטור $X = \{x_1, x_2, x_3, x_4\}$ הפוקנצייה $\beta_1 = x_1, \beta_2 = 1 - x_2, \beta_3 = 0, \beta_4 = x_4 \vee (1 - x_2)$ מתחילה למשתנים הבאים: $\beta_1 = x_1, \beta_2 = 1 - x_2, \beta_3 = 0, \beta_4 = x_4$. נראה אלגוריתם ERM על המחלוקת \mathcal{H}_{OR} הוא שמבצע אלימינציה ונקרא **ELIMINATE**.

אלגוריתם ELIMINATE:

1. נתחיל עם קבוצת כל המשתנים ושליהם $Z = \{x_1, \bar{x}_1, x_m, \bar{x}_m, \dots, x_i, \bar{x}_i\}$.
2. נעביר דוגמה אחר דוגמה במדגם S שלנו, בולמר עבור $i = 1, \dots, m$, בהינתן דוגמה $S \in \{(x^i, y^i) | i = 1, \dots, m\}$.
 - a. אם היא בעלת תיוג חיובי, $y^i = 1$, נציג עליה.
 - b. אם היא בעלת תיוג שלילי, $y^i = 0$. נעביר על הקבוצה Z, וכל **משתנה x_j או $x_j - 1$** ($j \neq i$ בלשאלה) שగורם לשערוך של x^i להיות 1 אנחנו נוריד. אם הדוגמה שלילית, והופיע לנו בקבוצה משתנה שגורם לשערוך של x^i להיות 1, בשיםוקח OR הוא ייתן לנו את הערך 1 (בי מודובר-ב-OR). אבל הדוגמה שלילית ולכן לא נרצה לקבל 1.
3. נזכיר $Z = \bigcup_{z \in Z} \text{ELIM}(S)$, ביצוע OR על פני כל המשתנים שנשארו בקבוצה.

טענה: תחת המקרה realizable, ההיפותזה קונסיסטנטית על המדגם, מושגה שגיאה 0 עליו: $e(S) = y^i$. כלומר, הטענה נכונה, מכיון ש- \mathcal{P} realizable.

דוגמה: נניח $m = 3$. המדגם שלנו הוא $S = \{(011,0), (000,1), (001,0)\}$.

1. ההיפותזה ההתחלתית שלנו היא $h(x) = (x_1) \vee (1 - x_1) \vee (x_2) \vee (1 - x_2) \vee (x_3) \vee (1 - x_3)$. מתחילה עם כל המשתנים והשליליה שלהם.
2. נעביר אחר הדוגמאות במדגם ונפסול בהתאם:
 - a. $(011,0)$ – תיוג שלילי. לכן נפסול את x_1, x_2, x_3 .
 - b. $(000,1)$ – תיוג חיובי. נמשיך.
 - c. $(001,0)$ – תיוג שלילי. לכן נפסול את x_2 .

הוכחה: יהי מדגם S , ונסמן $\mathcal{H}_{OR} = \{h: \mathcal{X} \rightarrow \mathcal{Y} | h = \beta_1 \vee \dots \vee \beta_m\}$. אנחנו יודעים שבכבר הינו קיימת ב- \mathcal{H} , כי אנחנו במקרה ה- \mathcal{P} realizable. נסמן ב- h את הפלט של האלגוריתם. נרצה להוכיח ש- h קונסיסטנטית על S , זה אומר שהוא ERM.

- עבור דוגמה $S \in \{(x^i, y^i) | i = 1, \dots, m\}$ שהיא שלילית – אז לפי האלגוריתם כל המשתנים שימושיים ל-1 לא יהיו ב- h . כל מי שstrand משערכו 0 על הדוגמה הזאת: $y^i = 0 = h(x^i)$. הפונקציה שלנו היא OR על פני דברים שימושיים 0, ולכן גם 0.
- עבור דוגמה $S \in \{(x^i, y^i) | i = 1, \dots, m\}$ שהיא חיובית.
 - נניח בשליליה שהפונקציה שלנו נתנה עליה תיוג שלילי: $0 = h(x^i)$.
 - אנחנו יודעים כי $1 = h(x^i)$ כי היא מתאיגת לנו. לכן, קיימים לפחות משתנה j אחד שמשערכו אותו ל-1.
 - המשטנה הזה נפסל בשלב בלשאלה באלגוריתם שלנו. בולמר קיימת דוגמה שלילית $S \in \{(x^k, y^k) | k = 1, \dots, m\}$, שהעיפה אותו כי הוא שיערך אותה ל-1.
 - בולמר $1 = h(x^k)$, בסתיו לקונסיסטנטיות של h .

פונקציה (ε, δ) :

בכמחלוקת היא OR על פwi ומשתנים, לכל אחד 3 אופציות (0, $x_j - 1$, x_j) בתוספת 1 על הפונקציה הקבועה שמתיגת הכל-ב-1. לפि המשפט שראינו בשיעור, כל מחלוקת סופית למידה PAC ע"י אלגוריתם ERM כאמור:

$$N(\varepsilon, \delta) = \frac{1}{\varepsilon} \ln \left(\frac{|\mathcal{H}|}{\delta} \right) = \frac{1}{\varepsilon} \ln \left(\frac{3^m + 1}{\delta} \right) \leq \frac{1}{\varepsilon} \ln \left(\frac{3^{m+1}}{\delta} \right) = \frac{1}{\varepsilon} \ln \left(\frac{1}{\delta} \right) + \frac{1}{\varepsilon} \ln(3^{m+1}) = \frac{1}{\varepsilon} \ln \left(\frac{1}{\delta} \right) + \frac{1}{\varepsilon} (m+1) \ln 3$$

מחלקת היפותזות אינסופית – מלבים במרחב:

קבע $\{0,1\}^2 = \mathcal{X}$. נתבון במחלקת ההיפותזות הבאה $\{y \in \mathbb{R}^2 \mid a_1 \leq b_1, a_2 \leq b_2\} = \mathcal{H}$. באשר:

$$h_{(a_1, b_1, a_2, b_2)}((x_1, x_2)) = \begin{cases} 1 & a_1 \leq x_1 \wedge a_2 \leq x_2 \leq b_2 \\ 0 & \text{otherwise} \end{cases}$$

בכמחלוקת אינסופית, כיוון 4 נקודות במרחב שייצרו מלבן. כל מי שבתו על המלון מקבל תיוג "+", כל מי שבוחז מקבל "-".

נסמן ב-R את המלון המוגדר על ידי (a_1, b_1, a_2, b_2) , ואת ההיפותזה המתאימה לו ב- h_R .

נרצה להוכיח שהיא למידה PAC בקרה-realizable, הפונקציה (ε, δ) .

נראה אלגוריתם A_{rec} שלומד את \mathcal{H}_{rec}^2 . האלגוריתם מקבל מבחן של n נקודות $\{x^i, y^i\} \in \mathbb{R}^2$ שהגלו IID מ-P שהוא $h_R \in \mathcal{H}_{rec}^2(X)$, כאשר R הוא מלון האמת שלנו. האלגוריתם שלמו יתנו $e_P(h_R) \leq \delta$.

נראה שקיים $N(\varepsilon, \delta) \geq N$ כך שכאשר נירץ את A_{rec} על $n \geq N(\varepsilon, \delta)$ דוגמאות IID מ-P נקבל היפותזה $\mathcal{H}_{rec}^2 \in h_R$ כך שבסיכוןו לפחות $1 - \varepsilon$ מתקיים $e_P(h_R) \leq \delta$.

אלגוריתם A_{rec} : בהינתן מבחן בגודל n , הוא מוצאת המלון הקטן ביותר R שהוא קונסיסטנטי עם המבחן. כפיו על המלון לכבוד את כל הנקודות החשובות בתוכו, ואת כל השיליות מחוץ לו. הוא יכול לטעת רק על ידי סיווג נקודות חיויבות בשליליות (זה קורה כאשר המלון R קטן מדי – R_0). שגיאת האלגוריתם היא $e_P(h_R) = P[R_0 \setminus R]$.

הוכחה: נניח לשם פשוטות שההתפלגות על X היא רציפה, ומתקיים $\varepsilon > P[R_0] > 1 - \varepsilon$ (אחרת $e_P(h_R) = P[R_0 \setminus R] \leq P[R_0] \leq 1 - \varepsilon$ שגיאת האמת). בעת נכוחה את סיבוכיות הדגימה:

התפלגותינו שלנו רציפה, לכן ככל שמתוקדים עליה צוברים מסה. לכן יש לנו בתוך המלון מסה של לפחות ε (הסיכוי שנקודה טיפול במלון). נגדיר 4 פסים על פנוי R_0 . נתחילה בקצתה של המלון, עד שצברנו מסה של $\frac{\varepsilon}{4}$. בתוך הפסים יש רק נקודות חיויבות.

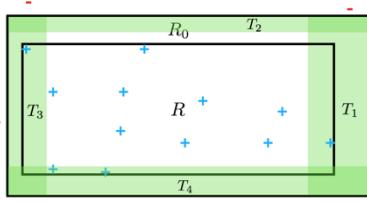
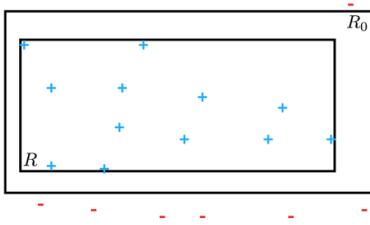
אבחןה: אם בכל אחד מהפסים נפלה לנו נקודה חיובית, אז השטח של מסגרת השגיאות שלו, חייב להיות מוכל באיחוד של הפסים. כלומר לכל $1 \leq j \leq 4$ קיים $i \leq n$ כך שיש נקודה חיובית $x^i \in T_j \subseteq \bigcup_{i=1}^n T_i \setminus R_0$. לכן:

$$e_P(h_R) = P[R_0 \setminus R] \leq P \left[\bigcup_{i=1}^n T_i \right] \leq \sum_{i=1}^n P[T_i] \leq \varepsilon$$

באופן שקול, אם $\varepsilon > e_P(h_R)$ גורר שקיים $1 \leq j \leq 4$ כך שלבכל $i \leq n$ מתקיים $x^i \notin T_j$ (אף נקודה לא נפלה בו). יש לנו תנקודות שנדגמו IID, לכל נקודה הסיכוי שהיא לא בפס הוא $1 - \frac{\varepsilon}{4}$.

$$P[e_P(h_R) > \varepsilon] \leq P[\exists j. \forall i. x^i \notin T_j] \leq 4 \left(1 - \frac{\varepsilon}{4}\right)^n \leq 4e^{-\frac{\varepsilon n}{4}} \leq \delta$$

$$N(\varepsilon, \delta) = \frac{4}{\varepsilon} \ln \frac{4}{\delta} \cdot 4e^{-\frac{\varepsilon n}{4}} \leq \delta \Leftrightarrow n \geq \frac{4}{\varepsilon} \ln \frac{4}{\delta}$$



מחלקות היפותזה אינסופיות

VC Dimension

מה קורה כאשר $= |\mathcal{H}|$ או מוד גודל? (אם מייצגים בעדרת מספרים ממשיים, יש אינסוף היפותחות אפשריות). במחלקות סופיות, לא ללחנו בחשבון בכלל את מבנה הבעה. בחלק זה $= \mathcal{H}$. בturn נגידר מושג **מדד של גודל מחלוקת היפותזה**, שלוקח בחשבון את המבנה של \mathcal{H} ויש מקרים שבהם הוא נותן חסם דגימות יותר טוב. במה הגדרות:

- **דיבוטומיות (Dichotomies)**: נגידר את \mathcal{H}_C להיות הצטצום של \mathcal{H} לקבוצה $\mathcal{X} \subseteq \{x_1, \dots, x_n\}$ כ $\mathcal{H}_C = \{h(x_1), h(x_2), \dots, h(x_n)\}$. \mathcal{H}_C היא קבוצת הדיבוטומיות של- \mathcal{H} משרה על C : $h \in \mathcal{H}_C : [h(x_1), h(x_2), \dots, h(x_n)]$. זה **אוסף התוצאות האפשרים**. ניקח כל היפותזה, נפעיל אותה על כל הנקודות ונקבל וקטור בינארי. היפותזה נוספת נוספה. אלה כל הדיבוטומיות שהמחלקה משרה. נשים לב כי $|\mathcal{H}_C| \leq 2^{|C|}$.
- **קבוצה מנוטצת (Shattered set)**: אם קבוצה סופית $\mathcal{X} \subseteq C$ מקיימת $2^{|C|} = |\mathcal{H}_C|$ בולמר אפשר להשרות על C את כל התוצאות הבינאריים, אז נאמר ש- C -מנוטצת על ידי המחלוקת \mathcal{H} .
- **VC Dimension**: ה- n -ה מקסימלי כך שקיימת קבוצה C בגודל n שמנוטצת על ידי \mathcal{H} . נסמן $VCdim(\mathcal{H})$ (בולם לא קיימת קבוצה מנוטצת גדולה יותר). אם אין כזו אז מתקיים $\infty = VCdim(\mathcal{H})$.

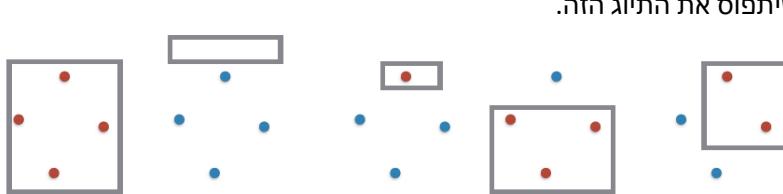
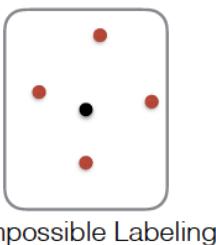
חישוב חסם מחלוקת \mathcal{H} : בהינתן מחלוקת \mathcal{H} נוכחים כי $k = VCdim(\mathcal{H})$ באופן הבא:

1. נמצא קבוצה $\mathcal{X} \subseteq \{x_1, \dots, x_k\}$ שמנוטצת על ידי \mathcal{H} .
2. נראה שאין תת-קבוצה של \mathcal{X} בגודל $k > n$ שמנוטצת על ידי \mathcal{H} . מספיק להראות שאין קבוצה כזו בגודל $1 + k$, ביוון שנייה כל קבוצה גדולה יותר, בהכרח יכול ניתוץ תת-קבוצה שלה מגודל $1 + k$.

דוגמאות:

1. **סוגים לינאריים מעלה הממשיים**: $\mathcal{H} = \{h(x) = sign(wx - b) : w, b \in \mathbb{R}\}$, כאשר $\{h(x) = sign(wx - b) : w, b \in \mathbb{R}\} = \mathcal{X}$. איך נראית היפותזה בחלוקת הזה? יש לנו קו ישר $b - wx = 0$ אחריו שמשמים על זה חזון נתון 1 בשני-עչובי ו-0 כשהוא שלילי. אפשר לבחור נקודה על הישר, משמאלו אליה הכל חיובי ומימינו הכל שלילי, או להיפך.
2. **קיימת קבוצה בגודל 2 שנוכבל לנתק:** למשל $C = \{x_1 = 3, x_2 = 4\}$. נוצר step function שהירידה היא בדיק בין הנקודות האלה, $+/-$, $+/-$, $+/-$. לכן $2 \geq VCdim(\mathcal{H})$.
3. **אין קבוצה בגודל 3:** למשל 3 נקודות בלבד, לא ניתן למשם את התיאוג הבא: $+/-, +/-, +/- = h(x)$.

2. **מלבנים מושרים לצירם מעלה המישור:** $\mathcal{H} = \{h_R(x) = \begin{cases} 1 & x \in R \\ 0 & x \notin R \end{cases} : R \subseteq \mathbb{R}^2\}$, כאשר $R \subseteq \mathbb{R}^2$. איך נראית היפותזה זו? יש לנו מלבן R במישור. היפותזה נותנת + לכל נקודה בתוך המלבן, - לכל נקודה מחוץ למלבן.
3. **קיימת קבוצה בגודל 4 שנוכבל לנתק:** לא משנה איזה תיוג נרצה, אפשר לבנות מלבן שיתפס את התיאוג הזה.



4. **אין קבוצה בגודל 5:** נקרא לנו "נקודה טובה" אם היא ה- i ה- i -קיצונית בכוון בלבדו (ימינה, מעלה). מתוך 5 יכולות להיות מוקסימים 4 נקודות טובות. אי אפשר לישם זהה מלבן, כי אם הוא יוכל את כל הנקודות הטובות, הוא יוכל גם את הנקודה הלא טובה.
5. **סוגים לינאריים מעלה ממד בלבדו (הכללה):** $\mathcal{H} = \{h(x) = sign(\langle w, x \rangle - b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$, כאשר $\{h(x) = sign(\langle w, x \rangle - b) : w \in \mathbb{R}^d, b \in \mathbb{R}\} = \mathcal{X}$. באנ מתקיים $VCdim(\mathcal{H}) = d + 1$ (d הוכחה בתרגול).

הערות:

- לא תמיד יש קשר יפה בין כמות הפרמטרים של המחלוקת $VCdim$:
- $VCdim(\mathcal{H}) = \{h_\alpha(x) = sign(\sin(\alpha x)) : \alpha > 0\} = \{0,1\}$. באנ מתקיים $\infty = VCdim(\mathcal{H})$ (הוכחה בתרגול).
- אם \mathcal{H} סופי אז $VCdim(\mathcal{H}) \leq \log_2 |\mathcal{H}|$.
- $VCdim(\mathcal{H}) = \infty$ אם $\mathcal{H} = \mathcal{X}^\mathcal{Y}$.



פונקציית גידול: נסמן על ידי (n) את המספר הגדל ביותר של דיבוטומיות ש- \mathcal{H} משרה על קבוצה בגודל n :

$$\Pi_{\mathcal{H}}(n) = \max_{C \subseteq \mathcal{X}, |C|=n} |\mathcal{H}_C|$$

נסתכל על זה בתור פונקציית הגדיל (growth function) של \mathcal{H} : $\mathbb{N} \rightarrow \mathbb{N}$. נשים לב כי:

- לכל $n \in \mathbb{N}$ מתקיים $2^n \leq \Pi_{\mathcal{H}}(n)$ – יש לפחות 2^n תוצאות אפשריות על קבוצה בגודל n .
- אם $\Pi_{\mathcal{H}}(n) = 2^n$ אז $VCdim(\mathcal{H}) = d$ לכל $d \leq n$.

למה: אם $d = \sum_{i=0}^d \binom{n}{i} > d$ ו- $VCdim(\mathcal{H}) = d$

מסקנה: לכל $d > n$ מתקיים: $\Pi_{\mathcal{H}}(n) \leq \left(\frac{en}{d}\right)^d$. יש פה התנהגות פולינומיאלית. ה- d -growth function בהתחלה גדול אקספוננציאלי, אחריו שמאט ופתח את המשווה. נשים לב כי מתקיים $1 < \left(\frac{d}{n}\right)^d$ ולכן נגדיל את הביטוי:

$$\left(\frac{d}{n}\right)^d \sum_{i=0}^d \binom{n}{i} = \sum_{i=0}^d \left(\frac{d}{n}\right)^d \binom{n}{i} \stackrel{i \leq d}{\leq} \sum_{i=0}^d \left(\frac{d}{n}\right)^i \binom{n}{i} \stackrel{\text{binom}}{=} \left(1 + \frac{d}{n}\right)^n \leq \left(1 + \frac{d}{n}\right)^{\frac{n}{d} \cdot d} = \left(1 + \frac{d}{n}\right)^{\frac{n}{d} \cdot d} \leq e^d$$

מכאן נקבל: $\sum_{i=0}^d \binom{n}{i} \leq \left(\frac{n}{d}\right)^d e^d = \left(\frac{en}{d}\right)^d$

Finite VC Dimension

משפט (Vapnik-Chervonenkis): נניח שאנו מבצעים סיווג ביןארי עם פונקציית loss zero-one, ו- \mathcal{H} מחלקת היפותחות. אז, לכל התפלגות P , קבוע שגיאה $\epsilon \in \mathbb{R}$ וב모ת דוגמאות $\mathcal{N} \in \mathbb{N}$ מתקיים:

$$P_S \left[\sup_{h \in \mathcal{H}} |e_S(h) - e_P(h)| \geq \epsilon \right] \leq 4\Pi_{\mathcal{H}}(2n) e^{-\frac{n\epsilon^2}{8}}$$

מדובר בתוצאה של התכנסות במ"ש (uniform convergence): לכל היפותחות ביחס. ההסתברות שקיימת היפותזה כלשהי שבעזרת המרחק בין השגיאה האמפירית שלה (train) והא יותר מ- ϵ – חסומה על ידי חסם מסוים. נסמן אם ϵ ימי היה קטן: יש ביטוי שדוען אקספוננציאלית $e^{-\frac{n\epsilon^2}{8}}$, ואו פונקציית הגידול $\Pi_{\mathcal{H}}(2n)$.

אם פונקציית הגידול לא גדרה אקספוננציאלית, כל הביטוי ידוע מהר, והדרך להבילה היא קקרה. עקב בכך נוכן להסיק, כי אם נזער את השגיאה האמפירית, זו דרך טובה להבטיח שהשגיאה האמיתית תהיה נמוכה.

הערה: עבור \mathcal{H} סופית טענו כי מתקיים: $P_S \left[\sup_{h \in \mathcal{H}} |e_S(h) - e_P(h)| \geq \epsilon \right] \leq 2|\mathcal{H}|e^{-2n\epsilon^2}$. בעצם, במקום $|\mathcal{H}|$ יש לנו את פונקציית הגידול שלה ($\Pi_{\mathcal{H}}$).

מסקנה: קיימים קבוע C כך שמתקיים הדבר הבא: לכל \mathcal{H} , התפלגות P , קבוע שגיאה $\epsilon \in (0,1)$, וקבוע ביחסו $(0,1)$, אם מספר דוגמאות האימון מקיימים $\frac{c}{\epsilon^2} \geq n$ כאשר $d = VCdim(\mathcal{H})$ אז, מובטח לנו כי:

$$P_S \left[e_P(ERM(S)) - \inf_{h \in \mathcal{H}} e_P(h) > \epsilon \right] \leq \delta$$

בפרט, \mathcal{H} למידה PAC עם סיבוכיות דוגמה $N(\epsilon, \delta)$

הערות:

1. \mathcal{H} סופית: $\ln \frac{1}{\delta} \leq \ln \left(\frac{2|\mathcal{H}|}{\epsilon} \right) = \frac{2}{\epsilon^2} \ln \left(\frac{2|\mathcal{H}|}{\delta} \right) = \frac{2}{\epsilon^2} \left(\ln 2|\mathcal{H}| + \ln \frac{1}{\delta} \right)$. עד כדי קבועים לוגריתמים, d החליף את $(|\mathcal{H}|) \ln$.
2. אפשר לקבל תוצאה במקרה ה- \mathcal{H} -realizable, ושם $N(\epsilon, \delta) = \frac{c}{\epsilon} \left(d \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta} \right)$.

הוכחה: ראיינו בעבר כי $|e_P(ERM(S)) - e_P^*(S)| \leq 2 \sup_{h \in \mathcal{H}} |e_S(h) - e_P(h)| \cdot \epsilon$. לכן:

$$\begin{aligned} P \left[e_P(ERM(S)) - \inf_{h \in \mathcal{H}} e_P(h) > \epsilon \right] &\leq P \left[2 \sup_{h \in \mathcal{H}} |e_S(h) - e_P(h)| > \epsilon \right] = P \left[\sup_{h \in \mathcal{H}} |e_S(h) - e_P(h)| > \frac{\epsilon}{2} \right] \\ &\leq 4\Pi_{\mathcal{H}}(2n) e^{-\frac{n\epsilon^2}{32}} \underset{\text{assume } d < 2n}{\leq} 4 \cdot \left(\frac{2ne}{d} \right)^d \cdot e^{-\frac{n\epsilon^2}{32}} < \delta \end{aligned}$$



- כדי לסייע את ההוכחה צריך למצאו $0 > C \ln \frac{d}{\varepsilon} + \ln \frac{1}{\delta}$ והוא זה יתקיים.
- נרצה להיפטר מההנחה שעשינו כי $2n < d$. אם C שקיבלנו הוא מספר מספיק גדול $\frac{1}{2} > C$ למשל – זה בבר קורה אוטומטית: קיום של התנאי $\ln \frac{d}{\varepsilon} + \ln \frac{1}{\delta} > n$ מבטיח ש- $\frac{d}{2} > n$. אחרת נחליף את C בקבוע גדול יותר.

Infinite VC Dimension

ראינו כי **Finite VC Dimension מבטיח למידות!**

טענה (No Free Lunch): נניח כי $\infty = \text{VCdim}(\mathcal{H})$ אז לכל אלגוריתם למידה \mathcal{A} , ומספר דוגמאות אימון m , קיימת התפוגות P כך שהאלגוריתם לא ייחזר היפותזה טובעה: ההסתברות שההפרשות בין השגיאה האמפירית של תוצאה A לבין שגיאת האמת המינימלית גדולה מ- $\frac{1}{8}$, היא לפחות $\frac{1}{16}$ (בך עבור $\frac{1}{16} = \delta_0 = \frac{1}{8}$), לא יוכל לדודת מתחה להה.

$$P_S \left[e_P(\mathcal{A}(S)) - \min_{h \in \mathcal{H}} e_P(h) > \varepsilon_0 \right] \geq \delta_0$$

בנוסף, התפוגות P יכולה להיות אפילו realizable (נחמדה), זה עדין יתקיים. כל אלגוריתם יצליח להביא תוצאה טובה יותר.

המשפט היסודי של למידות PAC: נניח שאנו מבצעים סיווג ביןארי עם פונקציית loss zero-one, \mathcal{H} מחלוקת היפותחות. אז:

$$\mathcal{H} \text{ is PAC learnable by ERM} \Leftrightarrow \mathcal{H} \text{ is PAC learnable} \Leftrightarrow \text{VCdim}(\mathcal{H}) < \infty$$

הערה: במקרה של multiclass (סיווג של יותר מ-2 תיוגים) הסיטואציה עדינה יותר, ואין אנלוגיה מדוייקת.

Bias-Variance Tradeoff: אנחנו רוצים ש- \mathcal{H} יהיה "עשיר" כדי שהשגיאה האופטימלית (H) תהיה קטנה. מצד שני, אנחנו רוצים שהוא "פשוט" (עם VCdim קטן) כדי שסבירויות הדוגימה תהיה קטנה (לא צריך הרבה דוגמאות כדי להבטיח שגיאת הכללה מסוימת).

תרגול 3 (VC Dimension

:VC dimension

ראינו שככל מחלוקת היא למידה PAC. ראיינו גם דוגמה למחלוקת אינסופית שהיא למידה PAC. אמנם, ברור **שלא כל מחלוקת אינסופית היא למידה PAC** (חלוקת כל הסבירויות בעולם). נרצה לקבוע קритריון שיקבע מתי מחלוקת היא למידה PAC, והוא נובע מה- VCdim . נתרגם את הקритריון הזה לשיבוכיות דוגימה. אנחנו עדין **מתמקדים בסיווג ביןארי, עם loss zero-one**.

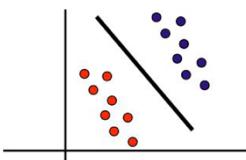
חזקה על מושגים:

- דיבוטומיות:** בהינתן קבוצה C מוגדל מ- \mathcal{H} מחלוקת \mathcal{H} , נסתכל על הצמצום של \mathcal{H} לקובוצה C : עולם בו קבוצה C מהווה את הקלטים שלנו. נסתכל על וקטורי התווג שהמחלקה משורה על העולם זה – קבוצת התווגים האפשריים של המודם C . נקבע קבוצה של וקטורים באורך m , באשר הכוונה בכל וקטור היא הפעלת $(x_1, \dots, x_n) \mapsto h(x_1), \dots, h(x_n)$: $\mathcal{H}_C = P\{[h(x_1), \dots, h(x_n)] : h \in \mathcal{H}\}$.
- מתקיים $|C| \leq |\mathcal{H}_C|$, אנחנו חסומים על ידי כל הוקטורים הבינאריים באורך m .
- נאמר ש- C -**מנוטצת על ידי** \mathcal{H} אם $|C| = |\mathcal{H}_C|$. המחלוקת ממנוטצת את כל וקטורי התווגים האפשריים על C .
- VC dimension:** הגודל המקסימלי של קבוצה C , שמנוטצת על ידי \mathcal{H} . זה מגדיר סיבוכיות של מחלוקת, בכל שהוא יותר גבוה, יש מוגם יותר גדול שמנוטצח על ידי \mathcal{H} , הוא ממנוטצת את כל התווגים על מוגם יותר גדול.
- הוכחת $k = \text{VCdim}(\mathcal{H})$:**
 - נראה קבוצה אחת בגודל k שמנוטצת ע"י \mathcal{H} .
 - נראה שלא קיימת קבוצה מוגדל $k+1$ שמנוטצת ע"י \mathcal{H} .
- המשפט היסודי של למידות PAC:** מחלוקת היא PAC $\Leftrightarrow \text{VCdim}(\mathcal{H})$ שלו הוא סופי.
- אם $\text{VCdim}(\mathcal{H}) = d$ סופי אז ניתן ללמידה את \mathcal{H} עם סיבוכיות דוגימה $N(\varepsilon, \delta)$.
- במקרה האגנוטטי מתקיים $N(\varepsilon, \delta) = \frac{C}{\varepsilon^2} \left(d \ln \frac{1}{\varepsilon} + \ln \frac{1}{\delta} \right)$.



מסוגים לינאריים מעל \mathbb{R}^d (תשפ"ג – לא הומוגנים, הפיע במליה 2, בתשפ"ד – הומוגנים):

נסתכל על מחלוקת המסוגים הלינאריים: $\mathcal{H}_d = \left\{ h(x) = sign(\langle w, x \rangle - b) = \begin{cases} 1 & \langle w, x \rangle - b \geq 0 \\ 0 & otherwise \end{cases} : w, x \in \mathbb{R}^d, b \in \mathbb{R} \right\}$



- במקרה של $b = 0$, כל המחלוקת היא מסוגים לינאריים שעוברים דרך ראשית הצירים. נהוג לבנות אותה "סוגים לינאריים הומוגניים".
- נסמן: $\langle w, x \rangle - b = [w, b]^T [x, 1]$.
- . $VCdim(\mathcal{H}_d) = d + 1$

קבוצה מגודל 1 + d שמנוטצת ע"י \mathcal{H}_d :

נבחר $e_i = x$ (הבסיס הסטנדרטי של \mathbb{R}^d) לכל $d \leq i \leq 1$ (d הוקטורים הראשונים יבחרו כך שהכונסה ה-0 שליהם היא 1 כאשר השאר 0), ובנוסף $e_{d+1} = x$. נראה שלכל תיוג מתאימה היפותזה שמנוטצת אותו. נקבע תיוג כללי $s_{d+1} \in \{0, 1\}, s_1, \dots, s_d$, כל תיוג הוא 0/1. נתבונן בתיוג של הוקטור האחרון ונפצל למקרים:

- אם $s_{d+1} = 0$, נגדיר היפותזה h עם המשקלות $w = [2s_1, \dots, 2s_d, b = 1]$. נראה שהיא מושגיה לנו את התיוג:
 - $h(x_i) = sign(\langle w, e_i \rangle - b) = sign(w_i - 1) = sign(2s_i - 1) = s_i$ ○
 - $.sign(2s_i - 1) = sign(-1) = 0 = s_i$ אם $s_i = 0$ ▀
 - $.sign(2s_i - 1) = sign(1) = 1 = s_i$ אם $s_i = 1$ ▀
 - $.h(x_{d+1}) = sign(\langle w, 0 \rangle - b) = sign(-1) = 0 = s_{d+1}$ ○
- אם $s_{d+1} = 1$, נגדיר היפותזה h עם המשקלות $w = [2s_1 - 2, \dots, 2s_d - 2, b = -1]$. באופן דומה:
 - $h(x_i) = sign(\langle w, e_i \rangle - b) = sign(w_i - (-1)) = sign(2s_i - 2 + 1) = sign(2s_i - 1) = s_i$ ○
 - $.h(x_{d+1}) = sign(\langle w, 0 \rangle - b) = sign(1) = 1 = s_{d+1}$ ○

קבוצה מגודל 2 + d לא מנוטצת ע"י \mathcal{H}_d :

נקבע קבוצה של $d + 2$ וקטורים: $v_i := [x_i, -1] := [(x_i)_1, \dots, (x_i)_d, -1] \in \mathbb{R}^{d+1}$. בולם, אנחנו משרשים 1 – ומוסיפים אותו בקואורדינטה האחורונה של הוקטור. $\{v_1, \dots, v_{d+2}\}$ היא קבוצה של $d + 2$ וקטורים ב- \mathbb{R}^{d+1} . גם, הם בהכרח ת"ל: קיימים מקדים $\alpha_{d+2}, \alpha_{d+1}, \dots, \alpha_1$ (לא כולם אפס) כך שמתקיים:
 $\sum_{i=1}^{d+1} \alpha_i v_i = \alpha_{d+2} v_{d+2}$.

בנוסף, נניח בה"ב כי $\alpha_{d+2} \neq 0$: לא כל הוקטורים הם 0, אז נבחר את v_{d+2} להיות זה שהמקדם שלו לא 0, ואז נחלק את כל המקדים האחרים במקדם α_{d+2} α_{d+2} ונקבל שהמקדם של v_{d+2} הוא 1.

נניח בשילhouette שהקבוצה $\{v_1, \dots, v_{d+2}\}$ כן מנוטצת ע"י המחלוקת.

כל תיוג אפשרי עליו מושג על ידי היפותזה כלשהי במחלוקת. נסתכל על התיוג: $1 = s_i = 0$ אם $\alpha_i \geq 0$ אחרת 0 . $s_i = 0$ מההנחה בשילhouette קיימת היפותזה h שמשגגה את $[s_1, s_{d+1}, \dots, s_{d+2}]$. נראה שההיפותזה תתייג את 1 , s_{d+2} , וזה טעה:

$$[w, b]^T v_{d+2} = \sum_{i=1}^{d+1} \alpha_i [w, b]^T v_i = \sum_{i=1}^{d+1} \alpha_i [w, b]^T [x_i, -1] = \sum_{i=1}^{d+1} \alpha_i (\langle w, x_i \rangle - b)$$

נראה כי הביטוי הנ"ל תמיד או-שלילי:

- נניח $0 \geq \alpha_i$: אז $s_i = 1$ כלומר $1 = sign(\langle w, x_i \rangle - b) \geq 0$ מ ממשת את התיוג. לכן, $0 = \langle w, x_i \rangle - b$ כולם בביטוי בצד ימין או-שלילי.
- נניח $0 < \alpha_i$: אז $s_i = 0$ כלומר $0 = sign(\langle w, x_i \rangle - b) < 0$ בביטוי בצד ימין או-שלילי. (מכפלה של גורם שלילי בגורם שלילי).

לכן: $0 = \langle w, x_{d+2} \rangle - b \geq 0$ ולכן $\langle w, x_{d+2} \rangle = b$ בסתירה.



שאלה מבחון (תשפ"א – מועד א):

שאלה 1

תהא \mathcal{H} מחלקה היפותזות של חיתוך k מפרידים לינאריים לא הומוגניים מעל $\mathcal{X} = \mathbb{R}^d$:
 $(d \geq 2)$ $\mathcal{H} = \left\{ x \mapsto \prod_{i=1}^k I[\langle w_i, x \rangle - b_i] : w_1, \dots, w_k \in \mathbb{R}^d, b_1, \dots, b_k \in \mathbb{R} \right\}$
 כאשר $I[z]$ שווה לאחד אם $z \geq 0$ וללאו אחרת.

א. (13 נק') הוכיחו כי ($VCdim(\mathcal{H}) \leq O(dk \log(dk))$)

במילים:

- לפי הلمה של סאורה, מחלוקת היפותזות שמייד ה-VC שלה הוא $3 \geq c$ משרה לכל היוטר m^c תוצאות על הקבוצה בגודל m כאשר $c > m$.
- חסמו מלמעלה את מספר התוצאות ש- \mathcal{H} משרה על כל שחי כתלות במספר התוצאות שמשרה מפריד לינארי לא הומוגני על אותה קבוצה.
- ניתן להשתמש בעובדה שאם $m < 2a \log(a)$ אז בהכרח $a \log(m) < a$.

נסמן ב- \mathcal{H}_d את מחלוקת המסוגים הלינאריים שראינו קודם. נקבע מדגם C בגודל m : $C = \{x_1, \dots, x_m\}$. עבור היפותזה h , נסמן את וקטור התוצאות שהיא משרה על הקבוצה C : $\bar{h}_C = (h(x_1), \dots, h(x_m))$. אם $\bar{h}_C = (h(x_1), \dots, h(x_m))$.

ניחסם את מספר התוצאות ש- \mathcal{H} משרה על הקבוצה C :

$$|\mathcal{H}_C| = |\{(h(x_1), \dots, h(x_m)) : h \in \mathcal{H}\}| =$$

כל $h \in \mathcal{H}$ היא מכפלה של k מסוגים לינאריים. כי h מובעת על ידי מכפלה של k היפותזות $h' \in \mathcal{H}_d$. לכן:

$$\begin{aligned} &= \left| \left\{ \left(\prod_{i=1}^k h'_i(x_1), \dots, \prod_{i=1}^k h'_i(x_m) \right) : h'_1, \dots, h'_k \in \mathcal{H}_d \right\} \right| = \left| \left\{ \begin{pmatrix} h'_1(x_1) & \dots & h'_k(x_1) \\ \vdots & \ddots & \vdots \\ h'_1(x_m) & \dots & h'_k(x_m) \end{pmatrix}^T : h'_1, \dots, h'_k \in \mathcal{H}_d \right\} \right| \\ &= \left| \left\{ \begin{pmatrix} h'_1(x_1) \\ \vdots \\ h'_k(x_m) \end{pmatrix} \circ \dots \circ \begin{pmatrix} h'_k(x_1) \\ \vdots \\ h'_k(x_m) \end{pmatrix} : h'_1, \dots, h'_k \in \mathcal{H}_d \right\} \right| = \left| \left\{ (\bar{h}'_1)_C \circ \dots \circ (\bar{h}'_k)_C : h'_1, \dots, h'_k \in \mathcal{H}_d \right\} \right| \\ &= \left| \left\{ (\bar{h}'_1)_C \circ \dots \circ (\bar{h}'_k)_C : (\bar{h}'_1)_C, \dots, (\bar{h}'_k)_C \in (\mathcal{H}_d)_C \right\} \right| \stackrel{\text{פונקציה על }}{\leq} |\mathcal{H}_d|^k \stackrel{\text{פונקציה על }}{\leq} VCdim(\mathcal{H}_d)^k = m^{(d+1)k} \end{aligned}$$

נזכיר כי $VCdim$ מוגדר להיות גודל הקבוצה הגדולה ביותר שמנוטצת על ידי \mathcal{H} . אם C מנוטצת על ידי \mathcal{H} אז:

$$\begin{aligned} |\mathcal{H}_C| &= 2^{|C|} = 2^m \leq m^{(d+1)k} \Leftrightarrow m \stackrel{\log(ab)=bloga}{\leq} (d+1)k \log m < 2(d+1)k \log m \\ &\Leftrightarrow m < 4(d+1)k \log(2(d+1)k) = O(dk \log(dk)) \end{aligned}$$



בחירה מודל

Bias-Variance Tradeoff

עד כה, עברו הPUR בין השגיאה האמיתית של מה ש-ERM מוחזר, לבין השגיאה האמיתית האופטימלית, פיתחנו חסמים עליונים שתומפסים בהסתברות גבוהה על פני מודם האימון (תחת הנחה שהdagimot הן IID מ-P). בחסמים אלו, הופיע "ഗודל" שמכמת את המורכבות של המחלקה \mathcal{H} . במקרה שהוא סופית – הבמות היתה מנות היפותזות. במקרה של המחלקות האינסופיות – ראיינו את ה-VCdim שהוא ממד מתאים. נכתבו את השגיאה האמיתית של ההיפותזה שלמדו בקורס קצת אחרת:

$$e_P(ERM(S)) = \left[\min_{h \in \mathcal{H}} e_P(h) \right]_{approximation\ error, bias} + \left[e_P(ERM(S)) - \min_{h \in \mathcal{H}} e_P(h) \right]_{estimation\ error, variance}$$

יש לנו כאן שני מקורות שגיאה, שונים מאוד אחד מהשני:

- ה-bias מבטא את יכולת המחלקה להשיג שגיאה נמוכה (להתאים להתרפות של העולם) והוא לא תלוי בשום צורה ב-set training ולא בקורס שאנו משתמשים בו.
- ה-variance זה הגורם שהתעסקנו בו, והוא נובע מהלמידה. אם יש למידה מושלמת הביטוי כולל יצא 0. הגורם הזה נובע מכך שאנו שוכן לא ידעים מה מודם אימון, מקבלים מודם קיטו, ואנו עושים מינימיזציה של השגיאה עליו.

באופן כללי – יש לנו הבטחה על ה-variance, אנחנו ידעים לחסום אותו מלמעלה אבל לא מלמטה. יכול להיות שבמקרה מה ש-ERM מוחזר הוא עם שגיאה אמיתית נמוכה. אנחנו לא ידעים כי יש לנו רק חסם עליון. לכן אם \mathcal{H} מחלקה קטנה, ה-bias עליה וה-variance יכול להיות רק יותר קטן (יותר טוב). לעומת זאת, אם נגדיל את \mathcal{H} ה-bias עולה (או נשאר זהה), וה-variance יכול רק לעלות. כך אנחנו קוראים ה-**bias-variance tradeoff**:

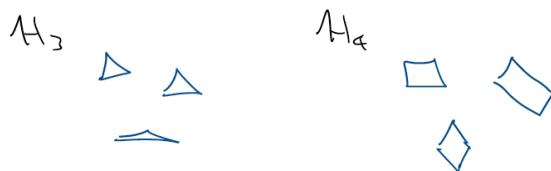
- מצד אחד אנחנו רוצים מחלקה גדולה מספיק כדי שבetta bias יהיה נמוך.
- מצד שני אנחנו רוצים מחלקה מספיק קטנה כדי שבetta variance יהיה נמוך.

התוצאה של "free lunch" סוף" שראינו, אומרת שאם לוקחים לקיצון את השיקול של הקטנת ה-bias ומרחיבים את \mathcal{H} עד שנבלול את כל היפותזות, לא משנה מה נעשה – לא נוכל לשולט ב-variance. לא נוכל לדעת $\frac{1}{16} = \frac{1}{8} = \varepsilon$. אנחנו צריכים להחיליט מה מחלקה ההי-hypothese לפניו שראויים את ה-set train – לאחר התוצאות יהיו לנו לא תקפות. **התוצאות היו ל-H קבוע ונוטן**. אם הימם לומדים היפותזה, והימם בוחרים רק אותה בתור \mathcal{H} , היינו מקבלים שגיאה נמוכה אבל בפועל זה אומר שהמחלקה שלמדו ממנה היא **מחלקת כל היפותזות**.

המחשה:

דמיין מחלקה היפותזות \mathcal{H}_{λ} כאשר λ שולט בסיבוכיות המחלקה – אם λ גדול אז \mathcal{H}_{λ} בעלת סיבוכיות מודם גדולה יותר.

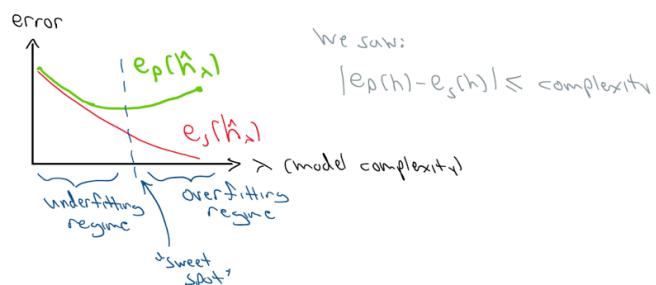
- $\mathcal{H}_{\lambda} = \{h_Q(x) = \begin{cases} 1 & x \in Q \\ 0 & x \notin Q \end{cases} : Q \text{ is a closed polygon with } \lambda \text{ edge}\}$. נגידו $\mathcal{X} = \mathbb{R}^2, \mathcal{Y} = \{0,1\}$, $\mathcal{N} = \mathbb{N}_{\geq 3}$. נשים לב כי $\mathcal{H}_3 \subseteq \mathcal{H}_4$, ככל שאנחנו מגדילים את λ אנחנו יכולים לבטא יותר דברים.



- $\lambda = \lambda$, ויש לנו קבוצת היפותזות אינסופית $\{h_1, h_2, \dots, h_{\lambda}\} = A$. נגידו $A = \{h_1, h_2, \dots, h_{\lambda}\}$ – לחתת את ה- λ הראשונים.

נניח שאנחנו חוזרים על התחילה הבא עברו λ משתנה:

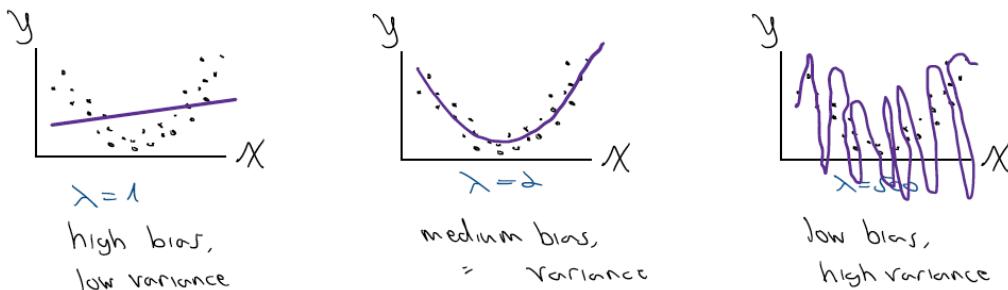
1. לומדים היפותזה מתוך \mathcal{H}_{λ} תוך שימוש במידם אימון S נתון, $\hat{h} = ERM_{\mathcal{H}_{\lambda}} \in \mathcal{H}_{\lambda}$ מtower ERM מtower המחלקה.
2. מודדים את השגיאה האמפירית (train): $e_P(\hat{h}_{\lambda})$ או $e_S(\hat{h}_{\lambda})$.
3. מודדים את השגיאה האמיתית על P: $e_P(\hat{h}_{\lambda})$.



השגיאה האמפירית יכולה רק לדעת בכל ש- λ גדול. השגיאה האמיתית בהתחילה ממשיכה לדעת כי המחלקה קטנה מדי (underfitting), ככל שבסיבוכיות הדגימה תעלה, נוכל ללמידה יותר והשגיאה האמיתית תקטן. אבל באיזשהו שלב, השגיאה האמיתית מתחילה לעלות כי למרות שיעשים fit לאטא, לא מבצעים טוב על מידע חדש, רק על האימון (overfitting). אפשרה באנטם שמאצא יישר שגיאה למצוא. סרטון נוסף שמודדים את המושגים נמצאו באנו.



דוגמה על הפולינומים: נניח כי $\mathcal{H} = \{0\} \cup \mathcal{X}$, כלומר \mathcal{H} הוא אוסף הפולינומים הדריכים. נניח שנתנו לנו training set ממשומן S. במקורה הריאלי bias הוא 0, התויג האמיתי מגע מיותר היפוטזה מהמחלקה. אם נוכל לעבוד עם מחלקה אחת בה רק היפוטזה אחת ולהישאר realizable זה הypi טוב, אבל לא מציאותי.



עבדיו נבון כיצד מתמודדים בעולם האמיטי עם ה-*tradeoff* זהה.

ידע מוקדם:

שיטת ראשונה היא להשתמש בידע שיש לנו על הבעה, כך שנוכל לצמצם את מחלוקת היפוטזות בצורה כזו שלא תפגע ב-bias, ועדין יוכל לעשות fit, שהשגיאה לא תהיה גדולה מדי:

1. **SPAAM** – נניח שנרצה להגיד האם מייל הוא ספאם או לא. נוכל לעבוד עם מחלוקת היפוטזות שמעבדות את כל המילום. אם נאמין שניתן בסיכוי גבוה לדעת האם מייל או ספאם רק **לפי הוכתרת**, יוכל **לצמצם את המחלקה בעלי לפגוע בביטויים**.

$$\begin{aligned} X &= \begin{array}{|c|c|} \hline \text{100} & \text{50} \\ \hline \text{200} & \text{100} \\ \hline \end{array} \quad \text{scale} \quad Y = \begin{array}{|c|c|} \hline \text{cat} & \text{dog} \\ \hline \end{array} \quad \text{bias: } \checkmark \quad \text{variance: } \times \\ H_1 &= \{ \text{Neural network with } 10^8 \text{ params} \} \\ \text{bias: } \checkmark & \text{variance: } \times \\ H_2 &= \{ h_w(x) = \text{sgn}(w^\top x) : w \in \mathbb{R}^{100 \times 100} \} \\ \text{bias: } \times & \text{variance: } \checkmark \\ H_3 &= \{ h_w(x) = \text{sgn}(w^\top \phi(x)) : w \in \mathbb{R}^d \} \\ \text{bias: } \text{can be small} & \text{variance: small if } d \text{ is small} \\ \text{phi: } \mathbb{R}^{100 \times 100} \rightarrow \mathbb{R}^d & - \text{representation func that someone designs} \end{aligned}$$

2. **סיווג תमונות** – נרצה לסווג תמונות להיות לבב/חתוכל.
 - nocell לעבוד עם מחלוקת היפוטזות של מסוג לנארו שמקבל פיקסלים, עושה צירוף לנארו שלהם ועשה sign – זה לא יתאפשר טוב את ההבדל בין כלבים וחתולים (אם קצת געלה את הבחרות ניזע את הפרדיקציה וזה ישפיע על התויג למחרות שתונן התמונה לא משתנה). **נקבל bias גבוהה**.
 - אם נשלח את זה לרשות מוחנים, ה-bias נזען אבל אז יש הרבה פרמטרים וסיבוכיות הדגימה גבוהה. נתען כי במקרה זה **נקבל variance גבוהה**.
 - nocell לעבוד עם מסוגים לנאראים, אבל במקום להפעיל אותם על הפיקסלים, נמצא **ממצא מודדים**. **סטטיסטיים של התמונה**: ממוצע פיקסלים למשל. אם נבנה את המודדים בצורה טובה, יוכל לקבל bias נזען.

SRM

נניח שיש לנו שרשרת של מחלוקת היפוטזות $\dots \subset \mathcal{H}_2 \subset \mathcal{H}_1$, יש לנו סט אימון S בגודל n, נרצה למדוד מתוך אחת המחלוקת הלו. כיצד נבחר את המחלוקת המתאימה? נרצה bias נזען וגם variance נזען. נניח שלכל ג' יש לנו חסם הכללה שמלל (1) $\in \delta$ ולבכל $N \in n$, בהסתברות לפחות $\delta - 1$ מתקיים:

$$e_P(ERM_{\mathcal{H}_\lambda}(S)) - e_S(ERM_{\mathcal{H}_\lambda}(S)) \leq \varepsilon(\delta, n, \lambda)$$

החסם ג' יהיה קטן יותר (ישתפרק) כאשר: ג' גדול (יותר דוגמאות אימון), δ גדול (פחות מבטיח דיק גובה יותר), λ קטן (ה- complexity קטן). נקבע את המשווה מחדר עבור k נתון, ונפעיל את זה עבור λ_k :

$$e_P(ERM_{\mathcal{H}_{\lambda_k}}(S)) \leq [e_S(ERM_{\mathcal{H}_{\lambda_k}}(S)) + [\varepsilon(\delta, n, \lambda_k)]_{\text{bias}}]_{\text{variance}}$$

נרצה לבחור \mathcal{H}_{λ_k} שմגדיר את ג' ימין של אי השוויון. הבעה היא שאי השווון תפס בהסתברות לפחות $\delta - 1$ עבור k נתון. זה לא אומר שבאוותה הסתברות זה תפס לכל ערך k ביחס. נרצה שהחסם ג' יהיה נכון עבור כל k. לשם כך, נשתמש בחסם האיחודי. ניקח משקלות $\{w_k\}$ כאשר $0 \leq w_k \leq 1$. לכל k נפעיל את החסם עם $w_k \cdot \delta$ במקומות λ_k ונקבל:

$$e_P(ERM_{\mathcal{H}_{\lambda_k}}(S)) \leq e_S(ERM_{\mathcal{H}_{\lambda_k}}(S)) + \varepsilon(\delta \cdot w_k, n, \lambda_k)$$



בהתבסירות לפחות $\omega_k - 1$ לכל k . הסיכוי שמשיסחו מה- k לא יתפוס הוא לכל היותר w_k . ההסתברות של איחוד המאורעות הראים הוא לכל היותר סכום ההסתברויות של המאורעות הראים, כי $1 = \sum w_k$. איך נבחר את המשקלות $\{\omega_k\}$? אם יש מספר סופי של היפותזות, נוכל לבחור את זה בצורה יוניפורמיות וכר לא ניתן עדיפות לפחות מחלוקת. אם יש מספר אינסופי של היפותזות נצטרך לתת משקל גדול יותר לחלק מהמשקלות כדי שהה יתכנס. נרצה לשימוש w_k גדולים בחלוקת שאנו מוכנים להן עדיפות בלבד (מסדרים את המחלוקת לפי עדיפות מסוימת).

$$\text{לרוב לוקחים } w_k = 2^{-k} \text{ או } w_k = \frac{6}{\pi^2} \cdot k^{-2} \text{ כי מתקיים } \sum \left(\frac{1}{k^2} \right) \leq \frac{\pi^2}{6} \text{ וצריך שזה יסטכם ל-1.}$$

גישת SRM (Structural Risk Minimization) ללמידה מסתובבת באופן הבא:

- קביעת משקלות $\{\omega_k\}$ בהתאם לכל המחלוקת $\{\mathcal{H}_{\lambda_k}\}$.
- בחירת δ – פרמטר confidence.
- הרעיון: חישוב של (S) ERM $_{\mathcal{H}_{\lambda_k}}$ לכל מחלוקת היפותזות \mathcal{H}_{λ_k} ומיציאת המסוווג שעבורו צד ימין באישווין מגע למינימום.

שיטות נוספות

: (Validation) Holdout

החיסרון ב-SRM הוא שצריך חסם הכללה, את ϵ , וגם דורשים שהחסם יהיה הדוק. אם אין לנו חסמים הדוקים כאלה?

1. מפעילים את סט האימון S לקבוצות זרות S_1, S_2 .
2. לכל מחלוקת היפותזות \mathcal{H}_{λ_k} משתמשים ב- S_1 כדי לומוד את \hat{h}_{λ_k} .
3. מתוך כל \hat{h}_{λ_k} מחוזרים את (h) בטור RM ERM מעל S_2 עם המחלוקת $\{\hat{h}_{\lambda_k}\}$.

אנחנו לומדים לכל מחלוקת בנפרד היפותזה על הדטא של S_1 , ולאחר מכן לוקחים את היפותזות של מדנו ועל דטא חדש שלא נגענו בו S_2 , מרכיבים ERM ולומדים. האיטרציה השנייה של ERM היא תהליך למידה חדש לפי מחלוקת היפותזות $\{\hat{h}_{\lambda_k}\}$ – ככלומר עושים ERM עם מחלוקת קטנה, ואז המרחק בין השגיאה האמפירית לשגיאה האמיתית יהיה קטן בצורה מובטחת.

זכורת: עבור \mathcal{H} סופית, לבלי P ו- $(0,1)$ $P[e_P(\text{ERM}(S)) - e_P^* > \epsilon] \leq \frac{2|\mathcal{H}|}{\epsilon^2} \ln \left(\frac{2|\mathcal{H}|}{\delta} \right)$ מתקיים שאם $\epsilon = \sqrt{\frac{1}{2n} \cdot \ln \frac{2|\mathcal{H}|}{\delta}}$ אם נחלץ את השגיאה ϵ מהמשווה נקבל $K = |\{\hat{h}_{\lambda_k}\}|$ אז בהסתברות לפחות $\delta - 1$ על פנו S_2 , מתקיים לבלי k :

$$|e_P(\hat{h}_k) - e_{S_2}(\hat{h}_k)| \leq \sqrt{\frac{1}{2|S_2|} \cdot \ln \frac{2K}{\delta}}$$

הערות:

- החסם הוא דו-ביווני, ככלומר הוא חסם הדוק.
- נקרא ל- S_2 validation set, ו- e_{S_2} validation error.
- לא ניתן להשתמש ב- S_2 לפני הצעד האחרון.
- חיסרון בשיטה זו – אנחנו מ Abedim דטא לאימון. ניתן לנקוט את $(k^*) = \arg \min_k e_{S_2}(\hat{h}_{\lambda_k})$ ולאמן מחדש את $\mathcal{H}_{\lambda_{k^*}}$.
- מההתחלתה תור שימוש בכל הדטא $S_1 \cup S_2$.

: Regularization

עד עכשיו הנהנו שיש לנו שרשרת היפותזות $\dots \subset \mathcal{H}_{\lambda_2} \subset \mathcal{H}_{\lambda_1} \subset \mathcal{H}$. לא תמיד אפשר לחשב בצורה טבעיות על שרשרת כזו של מחלוקת. במצב זה, גישה מקובלת נקראת **רגולרייזציה**. בצורה מכונה מגבלים מחלוקת היפותזות נתונה \mathcal{H}^* .

נכיה שההיפותזות ב- \mathcal{H}^* בעלות פרמטריזציה על ידי וקטורים ב- \mathbb{R}^d : $\{h_w : w \in \mathbb{R}^d\} = \mathcal{H}^*$. למשל, עבור מסוגים לגאים, ה-w יהיו שרשר של המשקלים וה-bias.

נכיה שיש לנו פונקציה $R : \mathbb{R}^d \rightarrow \mathbb{R}_{\geq 0}$: שנחשוב עליה בתור פונקציה שמקמתת מורכבות. בהינתן פונקציה כזו, אנחנו יכולים להגביל את \mathcal{H} באופן הבא: $\{\beta : \beta \in \mathcal{H}^*, R(\beta) \leq w\}$.

על ידי שינוי של β אנחנו זים בין מחלוקת היפותזה עם מורכבות שונה. ERM מעל \mathcal{H}_{β} יתן לנו:



$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \Delta(h_w(x_i), y_i), R(w) \leq \beta$$

אפשר להראות שבמקרים מסוימים בהם הרבה, למשל סוגי יינארים עם רגולרייזציה מסוימת, לפחות $\alpha > \beta$ קיים $0 < \alpha < \beta$ כך שהפתרון בעיית המינימיזציה הוא גם הפתרון בעיה הבהא:

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \Delta(h_w(x_i), y_i + \alpha \cdot R(w))$$

אם β הולך וגדל, α המתאים לו הולך וקטן, כי ככל ש- α גדול יותר זה שם יותר דגש על penalty, שזה שקול לאילוץ β יותר בבד. בפועל מה שבדר"כ עושים, זה להוסיף penalty ומשחקים עם ה- α . ככל ש- α גדול יותר עובדים עם מחלוקת קטנה יותר.

דוגמאות לפונקציות רגולרייזציה:

1. רגולרייזציה ℓ_2 : $R(w) = \|w\|_2^2 = \sum_i w_i^2$.
2. רגולרייזציה ℓ_0 : $R(w) = \|w\|_0 = |\{i : w_i \neq 0\}|$. מספר הקואורדינטות חוץ מ- x יהיו מאפסות. ברן שוטרים עם השולטים העם הגודל האפקטיבי של המחלוקת. אמנם, זה מאוד קשה חישובית לפתור את בעיית המינימיזציה אותה כי זה לא רציף.
3. רגולרייזציה ℓ_1 : $R(w) = \sum_i |w_i|$. זהו קירוב של ℓ_0 שאיתו בן ניתן לעבוד.

תרגול 4 (בחירה מודל)

בחירה מודל:

עד עבשו היוינו ב-setting של למידה בו קיבלנו מחלוקת ומודגס, והמטרה שלנו הייתה לפולוט מסווג מתוך המחלוקת שיש לו שגיאת הכללה טובה, ללמידה PAC. מה שמשמעותו אוננו זו סיבוכיות הדגימה – כמה דוגמאות אנו צרכים. הבעיה היא שלא תמיד יהיה לנו (δ, ϵ) גדול כרצוננו – להציג דאטא מתוגג הוא דבר יקר שדורש משאבים. במצב זהה, נניח שיש לנו K אופציות למחלוקות שניתנו ללמידה מהן. נראה שתי שיטות: SRM-Cross Validation. נשאל את עצמנו: איזה שגיאת הכללה אפשר להוציא בהינתן שיש לנו רק מספר דוגמאות קבוע n ? על איזה מחלוקת היפותחות נעבד?

ראינו בעבר מחלוקות היפותזה סופיות, את החסם הבא: $P \left[\sup_{h \in \mathcal{H}} |e_s(h) - e_p(h)| > \epsilon \right] \leq 2|\mathcal{H}|e^{-\frac{n\epsilon^2}{2}}$

נרצה להבין את ההתנהגות של השגיאה (ϵ), נפתח את המשווה בתלות ב- δ , ונקבל כי עבור **כמות דוגמאות קבועה n** , בהסתברות לפחות $\delta - 1$ לכל $h \in \mathcal{H}$ מתקיים: $e_s(h) + \sqrt{\frac{1}{2n} \ln \frac{2|\mathcal{H}|}{\delta}} \leq e_p(h)$.

באשר δ, n קבועים, אז ככל שהמחלקה יותר גדולה, ברן טיב הקירוב שלנו פחות טוב, כי איזה הביטוי מצד ימין גדול.

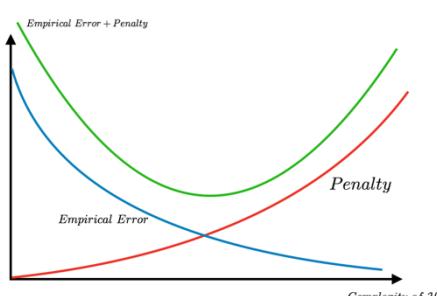
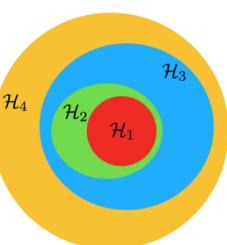
שיטת SRM: נניח שיש לנו k מחלוקות מוקנות $\mathcal{H}_k \subseteq \dots \subseteq \mathcal{H}_2 \subseteq \mathcal{H}_1$ כך שמתקיים $\mathcal{H} = \bigcup_{i=1}^k \mathcal{H}_i$ נגידר את העונש (penalty) עבור היפותזה בודדת h , להיות הגודל המשווהה הקודמת, עבור **המחלקה הקטנה ביותר** שהhypothese שיכת אליה. ככל שהמחלקה שאנו שיכים אליה יותר גדולה, ה-penalty גדול. באופן הבא:

$$\text{Penalty}_{srn}(h) = \min_{i:h \in \mathcal{H}_i} \sqrt{\frac{1}{2n} \cdot \ln \frac{2k|\mathcal{H}_i|}{\delta}}$$

אם רוצים למצאו את ההיפותזה שעשויה מינימיזציה גם לשגיאה האמפירית, אבל גם ל-**penalty** (שהוא קידוד לסיבוכיות של המחלוקת):

$$h_{srn} = \arg \min_{h \in \mathcal{H}} \{e_s(h) + \text{Penalty}_{srn}(h)\}$$

נשים לב: ככל שהמחלקה יותר גדולה, השגיאה האמפירית יורדת (בי יכול להיות התאמה יותר טובה למדגם ע"י היפותזה במחלוקת הגדולה יותר). לעומת זאת, ה-penalty עולה. אנחנו מכונים **macroscopic sweet spot** במשמעותו, אותו ישיג לנו ה-SRM.





נוכיח חסם הצללה על השגיאה האמיתית של ה-SRM:

נניח שיש לנו k מחלקות סופיות מוכנות אחת בשנייה. נסמן $h^* = \arg \min_{h \in \mathcal{H}} e_P(h)$ עברו ההיפותזה שמשיגה את השגיאה הכי טובה (הקטנה ביותר) על המחלקה הגדולה \mathcal{H} שהיא איחוד המחלקות. נסמן $i^* = \min\{1 \leq i \leq k \mid h^* \in \mathcal{H}_i\}$ עברו האינדקס של המחלקה הקטנה ביותר שהיא שייכת אליה. אנחנו רוצים להראות שבסיוכו $\delta - 1$:

$$e_P(SRM(S)) \leq e_P(h^*) + 2 \sqrt{\frac{1}{2|S|} \cdot \ln \frac{2k|\mathcal{H}_{i^*}|}{\delta}}$$

באמצעות חסם הופding union bound, לבלי $h \in \mathcal{H}_i$ עבור $i \leq k$ בסיכוי $\frac{\delta}{k} - 1$ קיבל:

$$(1) \quad |e_P(h) - e_S(h)| \leq \sqrt{\frac{1}{2|S|} \cdot \ln \frac{2k|\mathcal{H}_i|}{\delta}}$$

נסתכל על המחלקה \mathcal{H} הכוללת את כל המחלקות. לפי bound union על כל המחלקות, קיבל שהתנאי מתקיים במקביל לכל $k \leq i \leq k$ בסיכוי לפחות $\delta - 1$. נסמן $i_S = \min\{1 \leq i \leq k \mid SRM(S) \in \mathcal{H}_i\}$ עבור האינדקס של המחלקה הכי קטנה שמכילה את ההיפותזה שאנו מכבים חוזה מה-SRM. קיבל:

$$\begin{aligned} e_P(SRM(S)) &\stackrel{(1)}{\leq} e_S(SRM) + \sqrt{\frac{1}{2|S|} \cdot \ln \frac{2k|\mathcal{H}_{i_S}|}{\delta}} \underset{SRM \text{ minimizes this sum}}{\leq} e_S(h^*) + \sqrt{\frac{1}{2|S|} \cdot \ln \frac{2k|\mathcal{H}_{i^*}|}{\delta}} \\ &\stackrel{(1)}{\leq} e_P(h^*) + 2 \sqrt{\frac{1}{2|S|} \cdot \ln \frac{2k|\mathcal{H}_{i^*}|}{\delta}} \end{aligned}$$

שיטת Cross Validation: רוצים לדעת עם איזה מחלקה לעבוד. נאמן פשטוט ERM לכל מחלקה \mathcal{H}_i בנפרד, אבל בשאיר חלק מה-data set בצד שהוא יהיה ה-test set (holdout set). התהילך:

1. נחלק את הדטא ל-train, holdout עבור S_i, S_{ho} בהתאם.
2. נמצא ERM לכל \mathcal{H}_i : $G = \{g_1, g_2, \dots, g_k\}$, $g_i = \arg \min_{h \in \mathcal{H}_i} e_{S_i}(h)$.
3. נמצא ERM על ה-test set: $g_{cv} = \arg \min_{g \in G} e_{S_{ho}}(g)$.

למה אנחנו חייבים להפריד? למה לא ניתן לבצע את זה על כל ה-train set? ניצור תלויות, וונעשה overfitting על המודגם הספציפי.

2 – אלגורитמים של מידה מפוקחת

אופטימיזציה

(גרדיינטס)

נגזרת חלקית: עבור פונקציה $\mathbb{R}^n \rightarrow \mathbb{R}$: f , הנגזרת החלקית של f בנקודה $x \in \mathbb{R}^n$ ביחס למשתנה x_i היא:

$$\frac{\partial f(x)}{\partial x_i} = \lim_{h \rightarrow 0} \frac{f(x + he_i) - f(x)}{h} = \lim_{h \rightarrow 0} \frac{f(x_1, \dots, x_i + h, \dots, x_n) - f(x_1, \dots, x_i, \dots, x_n)}{h}$$

באשר e_i הוא וקטור שהכניסה ה- i -ו שלו הוא 1.

- ▶ For example, if $f(x, y) = x^2 + 2xy + x \sin(y)$, then the partial derivatives of f at (x_0, y_0) are,

$$\begin{aligned}\frac{\partial f(x_0, y_0)}{\partial x} &= 2x_0 + 2y_0 + \sin(y_0) \\ \frac{\partial f(x_0, y_0)}{\partial y} &= 2x_0 + x_0 \cos(y_0)\end{aligned}$$

- ▶ For example, if $f(x) = b^T x$ for some $b \in \mathbb{R}^n$,

$$\nabla_v f(x) = \lim_{h \rightarrow 0} \frac{b^T(x + hv) - b^T(x)}{h} = b^T v \quad \text{נגזרת כיוונית: הנגזרת הכוונית של } f \text{ בנקודה } x \text{ לאורך וקטור}\nabla_v f(x) = \lim_{h \rightarrow 0} \frac{f(x + hv) - f(x)}{h}$$

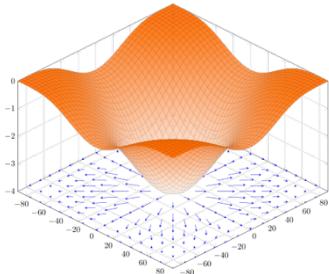
גרדיינט: הגרדיינט בנקודה x הוא הוktor של הנגזרות החלקיות בנקודה x :

- ▶ For example if $f(x) = b^T x = \sum_{i=1}^n b_i x_i$ for some $b \in \mathbb{R}^n$ then,

$$\nabla f(x) = (b_1, \dots, b_n)^T = b$$

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix}$$

ניתן להראות שבכל מקום שבו f דיפרנציאבילית מתקיים $\nabla_v f(x) = \nabla f(x)^T v$. הכוון של הוktor v שמקסם את המכפלה $v^T u$ הוא הכוון u . לכן, הכוון שבו הנגזרת הכוונית הכי גדולה הוא $\nabla f(x)$.





בעית אופטימיזציה קמורות

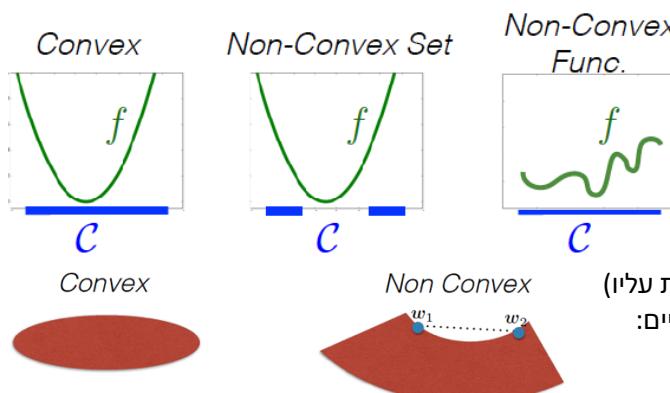
בקע: דיברנו על מידה של מסוגים לנאריים: $[w \cdot x + b = y_i]$. נזכיר בבעית ה-ERM: בהינתן n נקודות אימון (x_i, y_i) , נרצה למצוא את הפרמטרים b, w שמקטינם את השגיאה. אפשר להסתבל על בעיה אחרת – למצאו מסוג לנארוי עם שולים מקסימליים (large margin case). אפשר גם להסתבל על מסוגים לא לנאריים: $[b - (x \cdot w_1)^2 + (x \cdot w_2)^2 = sign[y]]$ עם הפרמטרים b, w_1, w_2 . כיצד נוכל לבצע ERM במקרה זה? במקרה הכללי $y = sign[g(x; w)]$.

לדוגמא, אם \mathcal{H} היא מחלקת מסוגים לנאריים מעל \mathbb{R}^{d-1} : $x \in \mathbb{R}^{d-1}, w \in \mathbb{R}^d, g(x; w) = \sum_{i=1}^{d-1} w_i x_i$. אם \mathcal{H} מתאימה לארQUITוורת של רשת נירונית, $(w; g)$ הוא הפלט של הרשת כאשר המשקלות הן w והקלט הוא x . באופן כללי אנו מנסים לבצע מינימיזציה באופן הבא:

$$\min_{w \in \mathcal{W}} f(w) = \frac{1}{n} \sum_{i=1}^n \Delta_{zo}(sign(g(x_i; w)), y_i) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}[sign(g(x_i; w)) \neq y_i]$$

הבעיה הזאת היא תיאורית NP-קשה. נתגבר על הקושי הזה באמצעות בחירת הפונקציה w ו גם הפסד (במקום 0-1).

בעית אופטימיזציה: המבנה הכללי כולל פונקציה $\mathbb{R} \rightarrow C: f$ שאנו רוצים למשער, קבוצה $\mathbb{R}^d \subseteq C$, ואנו מחפשים $(w; f(w))$ בגישה הנאיבית, אפשר לבצע brute-force ולפתור את הבעיה בזמן $O(2^d)$ על ידי בדיקת כל הערכים האפשריים. אנו רוצים לפתור את הבעיה הזאת כמו שיטור מהר.



בעית אופטימיזציה קמורות: יש מחלוקת מסוימת של בעיות שאוון כן אפשר לפתור – בעית אופטימיזציה קמורות. יש לנו פונקציית מטריה שהיא פונקציה קמורה $(w; f)$, וsett אילוצים קמור C . עבור בעיות כאלה, אין אופטימום מקומי, וניתן **למצוא את המינימום הגלובלי בזמן פולינומיי!**

קבוצה קמורה: אם לכל שני נקודות בקבוצה הכו שועור ביניהן (כל הנקודות עליה) גם נמצא בתוך הקבוצה. כלומר לכל $w_1, w_2 \in C$ ופרמטר $\alpha \in [0, 1]$ מתקיים: $\alpha w_1 + (1 - \alpha) w_2 \in C$

פונקציה קמורה: אם לכל שני נקודות הכו שמחבר ביניהם נמצא תמיד מעל הפונקציה. כלומר לכל $w_1, w_2 \in C$ ופרמטר $\alpha \in [0, 1]$ מתקיים:

$$f(\alpha w_1 + (1 - \alpha) w_2) \leq \alpha f(w_1) + (1 - \alpha) f(w_2)$$

דוגמאות:

1. פונקציות לנאריות: $f(w) = \langle x, w \rangle$.

2. פונקציות PSD ריבועיות: $f(w) = w^T A w, A \geq 0$.

3. מקסימום של קמורות: $f(w) = \max_i f_i(w)$ עבור פונקציות קמורות $f_i(w)$.

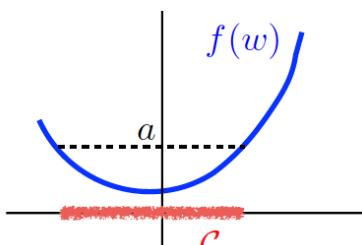
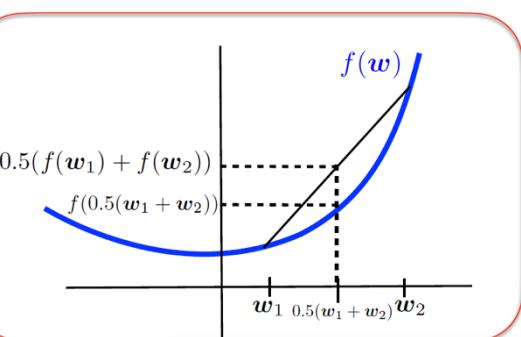
4. צירוף לנארוי של קמורות: $f(w) = \sum_i \alpha_i \cdot f_i(w)$.

5. הרכבה של קמורה מעל לנאריות: $f(w) = f_0(Aw + b)$ עבור f_0 קמורה, A מטריצה, b וקטור.

קשר נחמד בין פונקציה קמורה וקבוצה קמורה: עבור $f(w)$ קמורה ו- $a \in \mathbb{R}$.

נגידו $\{a \leq w: f(w) \leq a\} = C$. אז מתקיים **ש-C קמורה**.

שיעור מחודש על פונקציות קמורות (רק ההגדלה רלוונטי): [צבחור](#).





תכונות והגדרות של פונקציות קמורות:

מספר	תבונה	הערות
1	תהי $\mathbb{R}^d \subseteq C$ קבוצה קמורה ו- $f: C \rightarrow \mathbb{R}^d$. פונקציה קמורה f ב- C אם $\forall w \in C$ מינימום מקומי. אז w הוא גם המינימום globaal : $w \in \arg \min_{w \in C} f(w)$.	הוכחה: נניח בשילhouette שקיים $w'' \in C$ כך ש- $f(w'') < f(w')$. לפי הגדרת קמורות קיים $\lambda \in (0, 1)$ כך ש- $f(\lambda w' + (1 - \lambda)w'') \leq f(w'')$. $f(\lambda w' + (1 - \lambda)w'') < f(w')$ (שאוף לא-1 מלמטה), ונקבל סדרה לכך ש- w הוא מינימום מקומי.
2	תהי $\mathbb{R}^d \subseteq C$ קבוצה קמורה. $\exists f: C \rightarrow \mathbb{R}^d$ קמורה $\Leftrightarrow \forall u \in C \exists w \in z \in \mathbb{R}^d$ כך ש- $\forall u \in C: f(u) \geq f(w) + \langle u - w, z \rangle$	משמעות: הביטוי בצד ימין (פונקציה אפינית של u : לינארית + קבוע), צריך להיות קטן מ- $f(w)$. בالمור f לא יכולה להיות מתחת למשטח הזה, היא תמיד מעלייה, והוא משיק לפונקציה בדיקון בנקודת אחת.
3	תהי $\mathbb{R}^d \subseteq C$ קבוצה קמורה ו- $f: C \rightarrow \mathbb{R}^d$. פונקציה קמורה f ב- C , כלומר $\forall w \in C$, כל שמיים: $f(u) \geq f(w) + \langle u - w, z \rangle$ נקרא sub-gradient של f ב- w .	משמעות: אם יש לנו w , כל z שMageior משיק sub-gradient מתחת לפונקציה, נקרא sub-gradient של f ב- w . יכול להיות שיש רק 1 זהה, יכול להיות שיש לפחות 1 או יותר. קבוצת sub-differential set gradient של f ב- w נקראת ה- gradient ומסומנת $(\partial f)(w)$.
4	תהי $\mathbb{R}^d \subseteq C$ קבוצה קמורה ו- $f: C \rightarrow \mathbb{R}^d$. פונקציה גזירה ב- $w \in C$. אז מתקיים: $\{\nabla f(w)\} = \{\partial f(w)\}$.	יש רק 1 sub-gradient אחד והוא ה- gradient .
5	תהי $\mathbb{R}^d \subseteq C$ קבוצה קמורה ונניח שלכל $[k] \ni i$ גדרה $f_i: C \rightarrow \mathbb{R}$ פונקציה גזירה וקמורה. גדרה i גזירה $f_i(w) = \max_i f_i(w)$. עבור $w \in C$, לכל $j \in \arg \max_i f_i(w)$ (ניקח אינדקס של אחת מהפונקציות שבאותה נקודה היא מקסימלית) מתקיים: $(\nabla f_j(w)) \in \partial f(w)$.	הוכחה: f_j קמורה ולכן $\forall w \in C$ מתקיים (הגדרת sub-gradient): $f_j(w) \geq f_j(w') + \langle w - w', \nabla f_j(w') \rangle$ מכיוון ש- $f(w) = f_j(w)$ (לפי הגדרה j ובאופן כללי $f(w) \geq f(w') + \langle w - w', \nabla f_j(w') \rangle$) לפי הגדרה $(\nabla f_j(w)) \in \partial f_j(w)$. בשים לב כי f אינה גזירה כי היא מקסימום על פני נקודות.
6	תהי $\mathbb{R}^d \subseteq C$ קבוצה קמורה ו- $f: C \rightarrow \mathbb{R}^d$. גזירה ברציפות פעמים (נגזרות חלקיות רציפות). או f קמורה אמ"מ לכל $w \in C$ מתקיים: $\nabla^2 f(w) \geq 0$	כלומר, כאשר מטריצת הרסיאן היא PSD. ניתן לחשב את ההסיאן ולהראות שהוא אי-שלילי.



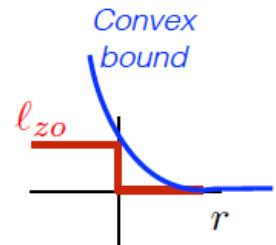
מ-ERM לאופטימיזציה קמורה

נסתכל על המסווגים ($w; sign(g(x; w)) = y$) עם הפרמטר w , בסיווג בינארי $\{0,1\} = \mathcal{Y}$. בעיית ה-ERM מביאה למינימום את:

$$e_S(w) = \frac{1}{n} \sum_{i=1}^n \Delta_{zo}(sign(g(x_i; w)), y_i)$$

אנו רוצים למצער סכום של פונקציה מהסוג $\Delta_{zo}(sign(f(x; w)), y)$ כאשר מה שמשתנה כל פעם הוא (y, x) . נחשב את $\Delta_{zo}(f(x; w), sign(y))$, להשווות ל- y ולומר אם זהה או לא – 0 או 1. נשים לב כי הפונקציה 0 \Leftrightarrow יש את אותו סימן, אחרת 1.

לכן, נפשט את הפונקציה ונכתב אותה כך: $\Delta_{zo}(sign(f(x; w)), y) = \ell_{zo}(yf(x; w))$. אם המכפלה היא שלילית (הסימן שונה) נחזיר 1, אחרת 0: $\ell_{zo}(r) = \begin{cases} 0 & r > 0 \\ 1 & r \leq 0 \end{cases}$. החסרונות בפונקציה זו: היא לא רציפה, היא קבועה, והיא לא קמורה. لكن קשה לעשות לה מינימיזציה ישירה.

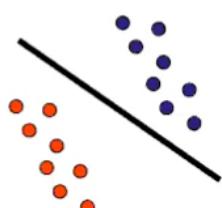


convex surrogate loss ERM: נחפש פונקציה loss חלופית. נחליף את ℓ_{zo} עם פונקציה נחמדה יותר, שהיא תחסום מלמעלה, כך שגם היא אפס, אך גם הפונקציה המקורית היא אפס. פונקציות נפוצות:

- $\ell_{hinge}(r) = \max[0, 1 - r]_+ = [1 - r]_+$:Hinge loss
- $\ell_{exp}(r) = e^{-r}$:Exponential loss
- $\ell_{log}(r) = \log_2(1 + e^{-r})$:Log loss

נסמן ב- ℓ את הפונקציה הקמורה החלופית ל- ℓ_{zo} . ותרונות:

- במקורה ש-g ליניארית, אז ה-loss הוא קמור (הרבעה). זה אומר שיש לנו סכום של פונקציות קמורות, ואפשר למצער את זה באופן יעיל.
- אם ה-loss הוא 0 אז גם ה-loss true הוא 0.



המקרה ה-e-Separable (אמיר תשפ"ג): במקרה הליינרי, אפשר להציגו לתוצאה הבאה: אם ה- Δ הוא separable, ונמצא רציף ℓ נגיע למסוג עם שגיאה 0. בניית שיס לינו דנטא Δ כולם קיימים w כך ש: $y_i = sign(w \cdot x_i)$. נחלה ב- γ ונגדיר $w^{-1} = \bar{w}$ כך $\bar{w} \cdot x_i \geq 1 \forall i$. זה שקול לכך ש- $0 \geq \gamma - x_i \cdot w \forall i$. נציב ב- γ ונגדיר $w^{-1} = \bar{w}$ כך $\bar{w} \cdot x_i - y_i \bar{w} = 0 \forall i$. במקרה זה \bar{w} שולץ ומדובר בהיפוך. בפועל ש- Δ אינו מושג תמיד גדול מ-1, ולכן \bar{w} שולץ ומדובר בהיפוך. במקרה שהוא מושג תמיד מ-1, אז \bar{w} שולץ ומדובר בהיפוך.

לכן **כל w_{min} שמצער את hinge loss יהיה עם שגיאה 0** (כי קיים \bar{w} שמשיג את זה). אם יש לנו מינימום hinge loss שווה 0 אז $1 \geq w_{min} \cdot x_i \forall i$. מכך $w_{min} \cdot x_i \geq 1 \forall i$ ו**מזהיג נכון!**

לסיכום, אנו רוצים ללמד מסוג LINEAR ($x \cdot w = sign(f(x))$). בהינתן נתונים (x_i, y_i) נגדיר ERM באופן הבא:

$$\min_w \sum_i \Delta_{zo}(sign(w \cdot x_i), y_i)$$

ניתן להיעזר בשתי פונקציות loss חלופיות:

- באמצעות hinge loss: $\min_w \sum_i \max[1 - y_i w \cdot x_i, 0]$
- באמצעות log loss: $\min_w \sum_i \log(1 + e^{-y_i w \cdot x_i})$

אלגוריתמים לאופטימיזציה

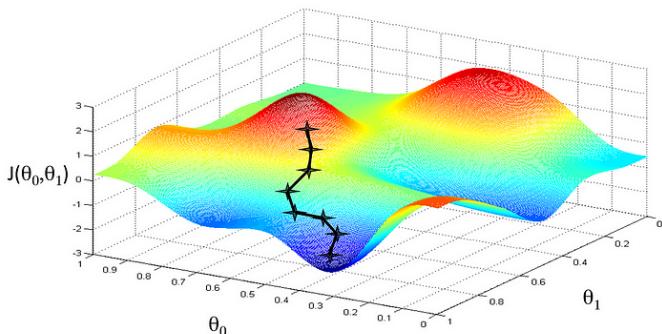
אלגוריתמי אופטימיזציה מנסים למצער objective בצורה איטרטיבית – מנסים למצוא פתרון עבור מזער של $(w; f)$ על פני w . ניתן לסייע להם לפי האינפורמציה שהם דורשים. בימידת מכונה מודרנית יש d גדול מאוד ולרוב משתמשים ב- 1 -order:

1. 0 -order: מניחים שיש גישה רק לערכים (w, f) , קוראים לפונקציה ומבצעים עליה **evaluation**.
2. 1 -order: מניחים שיש גישה לערכי (w, f) וגם $\nabla f(w)$.
3. 2 -order: גישה לערכים הקודמים וגם לערכי $\nabla^2 f(w)$.



אלגוריתם (GD) Gradient Descent:

אינטואיציה: ניעזר בעובדה שగודיאנט בנקודה מסוימת מצביע על הכוון שבו הפונקציה עולה הכי מהר. הכוון ההופך מצביע על הכוון שבו הירידה הכי מהירה. נרצה להגיע למינימום, ובן-蹊וח כל פעם צעד בכיוון שבו הירידה הכי גדולה עד שנגיע למינימום. מה שסבטיות לנו להגיע לנקודה זו, היא העובדה שהפונקציה קמורה.



פרמטרים:

- $w \in \mathbb{R}^d$ – ערך ההתחלתי.
- $T \in \mathbb{N}$ – מספר איטרציות.
- $\eta_t > 0 \}_{t=1}^T$ – גודל הצעדים, קצב הלמידה.

האלגוריתם:

1. נתחל $w' = w_1$.
2. עבור $T, \dots, t = 1$
- a. נבצע $w_{t+1} = w_t - \eta_t \nabla f(w_t)$.

משפט: נרצה להראות שהוא מתכנס למינימום גלובלי באשר בעיית האופטימיזציה קמורה. תהי $f: \mathbb{R}^d \rightarrow \mathbb{R}$: קמורה וגדירה. נניח כי $w^* \in \arg \min_w f(w)$ ו- $G \leq \|w^*\|_2 \leq \|G\|_2$. נניח כי $B \subseteq \mathbb{R}^d$ לכל $w \in B$. נסמן $w_t = \bar{w}$ (ممוצע בכל הוקטורים w). $f(\bar{w}) - f(w^*) \leq \varepsilon$ מתקיים: $T = \frac{\varepsilon^2 G^2}{\varepsilon^2} = \frac{B^2 G^2}{\varepsilon^2}$ עבור $\eta_t = \frac{\varepsilon}{G^2} > 0$. ($\eta_t = \frac{\varepsilon}{G^2}$ מתקבל מעריכת האלגוריתם).

הערות:

- יש מקרים בהם האופטימום הוא 0.
- ניתן לפרש את התוצאה באופן הבא: אם נróż מספיק איטרציות האלגוריתם יתכנס בהינתן הפרמטרים המתאימים.
- ניתן להוסיף פונקציות רגולרייזציה וכן לקבל הבטחה של התכנסות (למשל נורמת L_2).

(SGD) Stochastic Gradient Descent

נניח שלפונקציה f יש את הצורה הבאה: $(w)_i = \frac{1}{n} \sum_{i=1}^n f_i(w)$. במקרה שלנו loss הוא ממוצע של loss ולבן זה המצב במקרה שלם (ERM), כאשר זה שווה למספר הדוגמאות. כדי לחשב את הגודיאנט של f צריך לעבור על כל הפונקציות ולחשב את הגודיאנט של כלם. בשיש הרבה דוגמאות החישוב הנ"ל הוא כביד – בכל איטרציה צריכים לעבור על כלם. השתמש ב-SGD.

אינטואיציה: בחורמים אינדקס של objective אחד, מנסים למצער טופוגרפיה של הרבה טופוגרפיות, דוגום אחת מהן ומזו לפיה. כל איטרציה תהיה בעלת סיבוכיות יותר נמוכה. מצד שני, לא ברור לאן האלגוריתם הולך.

פרמטרים: כמו ב-GD.

האלגוריתם:

1. נתחל $w' = w_1$.
2. עבור $T, \dots, t = 1$
- a. נבחר $i \in [n]$ באופן רנדומלי אחיד ונבצע: $w_{t+1} = w_t - \eta_t \nabla f_{i_t}(w_t)$.

הערות:

• GD ו-SGD מייצגים מקרים קיצוניים: GD עובר על כל ה- $f_i(w)$ (high cost, low var) ו-SGD עובר רק על 1 בכל איטרציה (low cost, high var). Batch SGD הוא דרך המedium ביןיהם: לוקחים m פונקציות בכל איטרציה ומח산ים את הגודיאנט באמצעותם.

בaan, הבטחת ההתכנסות תהיה **בתוחלת**, זה המחיר שנשלם.

אפשר להפעיל את SGD גם אם $\{f_i\}$ לא גדיות אבל קמורות. במקרה זה, לוקחים gradients **sub-gradients** במקום gradients. לדוגמה: $\max\{0, 1 - \lambda \langle w, X_i \rangle\} + \lambda \langle w, X_i \rangle = \max\{0, 1 - \lambda \langle w, X_i \rangle\}$. במקרה זה, אם נרצה לקחת sg יש מקסימום על שתי פונקציות לינאריות, $-\lambda X_i$ ו- λw קמור כי ההסיאן שלו הוא מטריצה יחידה, ולכן כל הביטוי הוא קמור.

○ אם צדקנו $-\lambda \geq \langle w, X_i \rangle$, אז הוא הערך שמקסם את מתקיים: $(w) \in \partial f_i(w)$ ואז העדכון של SGD יהיה:

$$w_{t+1} = w_t - \eta_t \cdot 2\lambda w_t = w_t(1 - 2\eta_t \lambda)$$

○ אם עשינו שגיאה $-\lambda < \langle w, X_i \rangle$ אז העדכון של SGD יהיה:

$$w_{t+1} = w_t - \eta_t \cdot (-Y_i X_i + 2\lambda w_t) = w_t(1 - 2\eta_t \lambda) + \eta_t Y_i X_i$$



תרגול 5 (אופטימיזציה קמורה)

אופטימיזציה קמורה - GD:

בקורס עד כה, עבדנו עם מודגמ, מחלוקת היפותזות, ופתרנו ERM. אמנם, באופן כללי זו בעיה קשה חישובית, ואופטימיזציה הוא התחום שמאפשר לנו להתעסק עם מינימיזציה לפונקציות, ובפרט ERM. למצוא מינימום לפונקציה זה גם קשה חישובית, אבל תחת הנחות מסוימות (כמו קמיות), אפשר למצוא אופטימום באופן יעיל. לבן, נבחר פונקציית loss שהיא קמורה, ומחלוקת ההיפותזות צריכה להיות בזו שלא הורשת את הקמיות של loss. נגיד את הבעה שלנו (w) $\min_{w \in C} f(w)$ כאשר:

- C קמורה (לכל שתי נקודות, המיתר שעובר ביניהן נמצא גם בקבוצה C): $\lambda w_1 + (1-\lambda)w_2 \in C$
- f קמורה (לכל שתי נקודות הפעלה של f על הקומבינציה הקמורה קטנה או שווה לקומבינציה הקמורה כאשר קודם נפעיל את f על הנקודות): $(\lambda f(w_2) + (1-\lambda)f(w_1)) \leq \lambda f(w_2) + (1-\lambda)f(w_1)$.
- אם f דיפרנציאבילית, הגדרה שcola היא: $(f(u) - f(w) + \nabla f(w) \cdot (u - w)) \geq 0$. זה אומר שהמשור המשיק f -בנקודה ש נמצא מתחתיה.

אלגוריתם GD:

- נאחל $w_1 = w$ ולכל $T = 1, \dots, T$, נקבע $w_t = w - \eta \nabla f(w_t)$. את גודל הצעד שלנו η צריך לבחור בקפידה. אם הצעד יהיה קטן מדי, נתקדם לאט מדי. אם הצעד יהיה גדול מדי, נוביל לפחות ולפספס את הנקודה.
- אנחנו יודעים להוכיח התכונות של GD תחת הנחות מסוימות: f דיפרנציאבילית וקמורה. נסמן V_t את הנקודה שambilיה אותה למינימום. נניח כי $\|V_t\|_2 \leq \|w^*\|_2$. בנוסף, $G \leq \|\nabla f(w^*)\|_2$ (תנאי ליפшиז). הפלט של האלגוריתם שלנו בנקודות זמן T בה נוצר, יהיה \bar{w} ממוצע הנקודות שחישבנו עד T .
- משפט: יהי $0 > \epsilon$ (הדיוק שאנו שואפים אליו). עבור $\frac{\epsilon}{G^2} = \eta$ בגודל צעד $\frac{B^2 G^2}{\epsilon^2} = T$ צעדים: $\epsilon \leq |f(\bar{w}) - f(w^*)|$

הוכחת המשפט עבור GD (תשפ"ד):

נסמן $V_t = \nabla f(w_t)$. ראשית, נרצה לחסום את $|w^* - w_t|$. נסתכל על הביטוי הבא:

$$\begin{aligned} \|w_{t+1} - w^*\|_2^2 &= \|w_t - \eta V_t - w^*\|_2^2 = \|(w_t - w^*) - \eta V_t\|_2^2 = \|w_t - w^*\|_2^2 - 2\eta V_t(w_t - w^*) + \eta^2 \|V_t\|_2^2 \\ \Leftrightarrow V_t(w_t - w^*) &= \frac{\|w_t - w^*\|_2^2 - \|w_{t+1} - w^*\|_2^2}{2\eta} + \frac{\eta}{2} \|V_t\|_2^2 \end{aligned}$$

לפי א-שוויון ינסן לפונקציות קמורות: $f(\bar{w}) = f\left(\frac{1}{T} \sum_{t=1}^T w_t\right) \leq \frac{1}{T} \sum_{t=1}^T f(w_t)$. בפרט, לפי קמיות נקבל:

$$f(\bar{w}) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^T f(w_t) - f(w^*) = \frac{1}{T} \sum_{t=1}^T (f(w_t) - f(w^*)) \leq \frac{1}{T} \sum_{t=1}^T (w_t - w^*) \nabla f(w_t)$$

בעת הגיעו להסם:

$$\begin{aligned} f(\bar{w}) - f(w^*) &\leq \frac{1}{T} \sum_{t=1}^T (w_t - w^*) \nabla f(w_t) = \frac{1}{T} \sum_{t=1}^T \frac{\|w_t - w^*\|_2^2 - \|w_{t+1} - w^*\|_2^2}{2\eta} + \frac{\eta}{2} \|V_t\|_2^2 = \\ &\frac{1}{T} \sum_{t=1}^T \frac{\|w_t - w^*\|_2^2 - \|w_{t+1} - w^*\|_2^2}{2\eta} + \frac{1}{T} \sum_{t=1}^T \frac{\eta}{2} \|V_t\|_2^2 \leq \frac{\|w_1 - w^*\|_2^2}{2\eta T} + \frac{\eta G^2}{2} \underset{w_1=0}{\leq} \frac{B^2}{2\eta T} + \frac{\eta G^2}{2} = \epsilon \end{aligned}$$

אלגוריתם SGD:

- פונקציית האופטימיזציה מייצגת את השגיאה האמפירית שלנו: $(w, f) = \frac{1}{n} \sum_{i=1}^n f_i(w)$, ממוצע על n פונקציות אחרות. לכן עליה לנו $O(n)$ לחשב לה גרדיאנט. נמצא אומד בלתי מוטה שלו – נבחר η באופן אחיד ורנדומלי ונשתמש בו כדי לחשב את הגרדיאנט $\nabla f_i(w)$.
- אלגוריתם: לכל $T = 1, \dots, T$, נבחר $[n] \in I_t$ בהסתגלות אחידה, נחשב $V_t = \nabla f_{I_t}(w_t)$ ובעת $V_t \cdot \eta = w_{t+1} - w_t$.
- הנחות: דומה להנחות הקודמות, רק שכאן כל f_i דיפרנציאבילית וקמורה, $\|f_i\|_2 \leq G$ (האומד בנקודות שאנו צריכים).
- משפט: יהי $0 > \epsilon$. עבור $\frac{\epsilon}{G^2} = \eta$ ו- $T = \frac{B^2 G^2}{\epsilon^2}$ מתקיים: $\epsilon \leq |f(\bar{w}) - f(w^*)|$. עבשו יש לנו **תוחלת** כי האלגוריתם רנדומרי. בכל סיבוב t הוא משתנה מקרי כי הגרדיאנט נבחר באקראיות.

הוכחת המשפט עבור SGD (תשפ"ג):

ראשית, נרצה לחסום את $(w_t - w^*) = \frac{\|w_t - w^*\|_2^2 - \|w_{t+1} - w^*\|_2^2}{2\eta} + \frac{\eta}{2} \|V_t\|_2^2$. כמו בהוכחה הקודמת: $V_t = f'(w_t)$. לפיכך $f(\bar{w}) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^T (w_t - w^*) \nabla f(w_t)$. בעת גי"ע לחסם: $\mathbb{E}[f(\bar{w}) - f(w^*)] \leq \frac{1}{T} \sum_{t=1}^T \mathbb{E}[(w_t - w^*) \nabla f(w_t)] = \frac{1}{T} \sum_{t=1}^T \mathbb{E}[(w_t - w^*) \mathbb{E}[V_t | w_t]] = \frac{1}{T} \sum_{t=1}^T \mathbb{E}[\mathbb{E}[V_t \cdot (w_t - w^*) | w_t]]$

$$\begin{aligned} \mathbb{E}[f(\bar{w}) - f(w^*)] &\leq \frac{1}{T} \sum_{t=1}^T \mathbb{E}[(w_t - w^*) \nabla f(w_t)] = \frac{1}{T} \sum_{t=1}^T \mathbb{E}[(w_t - w^*) \mathbb{E}[V_t | w_t]] = \frac{1}{T} \sum_{t=1}^T \mathbb{E}[\mathbb{E}[V_t \cdot (w_t - w^*) | w_t]] \\ &= \frac{1}{T} \sum_{t=1}^T \mathbb{E}[V_t \cdot (w_t - w^*)] \leq \frac{1}{T} \sum_{t=1}^T \mathbb{E}\left[\frac{\|w_t - w^*\|_2^2 - \|w_{t+1} - w^*\|_2^2}{2\eta} + \frac{\eta}{2} \|V_t\|_2^2\right] = \text{סכום תחלות} \\ &\frac{1}{T} \sum_{t=1}^T \mathbb{E}\left[\frac{\|w_t - w^*\|_2^2 - \|w_{t+1} - w^*\|_2^2}{2\eta}\right] + \frac{1}{T} \sum_{t=1}^T \mathbb{E}\left[\frac{\eta}{2} \|V_t\|_2^2\right] \leq \frac{\|w_1 - w^*\|_2^2}{2\eta T} + \frac{\eta G^2}{2} \stackrel{w_1=0}{\leq} \frac{B^2}{2\eta T} + \frac{\eta G^2}{2} = \varepsilon \end{aligned}$$

 שאלה מבחן – תרגול תשפ"ג (תשפ"א א' – מועד א):

Consider the function

$$F(\mathbf{w}) = \|X\mathbf{w} - \mathbf{y}\|_2^2,$$

where $X \in \mathbb{R}^{n \times n}$ is invertible. Let UDU^T be the eigen decomposition of $X^T X$, where U is an orthogonal matrix, $D = \text{diag}(\lambda_1, \dots, \lambda_n)$ and $\lambda_1 \geq \dots \geq \lambda_n > 0$.

1. Show that F can be written as,

$$F(\mathbf{w}) = (U^T(\mathbf{w} - \mathbf{w}^*))^T D (U^T(\mathbf{w} - \mathbf{w}^*)) + c,$$

where \mathbf{w}^* is the optimal solution and c is some constant.

2. Assume that we run GD on the function F . Let $q_t = U^T(\mathbf{w}_t - \mathbf{w}^*)$. Show that $q_{t+1} = q_t - 2\eta D q_t$.
3. Show that for any initialization q_1 , if $\eta < \frac{1}{\lambda_1}$ then $\lim_{t \rightarrow \infty} q_t = 0$.

סעיף 1: מבקשים מאיינו לייצג את הפונקציה הריבועית, בתבנית ריבועית שתלויה בפירוק שלה לערכים עצמיים ובפתרון האופטימלי. ראיינו איך לחשב גודיאנט לפונקציה מהצורה זו: $\|\mathbf{Ax} - \mathbf{b}\|_2^2 = 2(A^T \mathbf{Ax} - A^T \mathbf{b})$. לכן נשווה לאפס:

$$\nabla F(\mathbf{w}) = 2(X^T \mathbf{X} \mathbf{w} - X^T \mathbf{y}) = 0 \Leftrightarrow X^T \mathbf{X} \mathbf{w}^* = X^T \mathbf{y} \Leftrightarrow \mathbf{y}^T \mathbf{X} = X^T \mathbf{X} \mathbf{w}^*$$

נבטא את F באמצעות \mathbf{w} ונציב:

$$\begin{aligned} F(\mathbf{w}) &= (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y}) = (\mathbf{w}^T \mathbf{X}^T - \mathbf{y}^T)(\mathbf{X}\mathbf{w} - \mathbf{y}) = \mathbf{w}^T \mathbf{X}^T \mathbf{X}\mathbf{w} - \mathbf{w}^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X}\mathbf{w} + \mathbf{y}^T \mathbf{y} = \\ &= \mathbf{w}^T \mathbf{X}^T \mathbf{X}\mathbf{w} - 2\mathbf{y}^T \mathbf{X}\mathbf{w} + \|\mathbf{y}\|_2^2 \stackrel{\mathbf{y}^T \mathbf{X} = \mathbf{X}^T \mathbf{X}\mathbf{w}^*}{=} \mathbf{w}^T \mathbf{X}^T \mathbf{X}\mathbf{w} - 2\mathbf{w}^T \mathbf{X}^T \mathbf{X}\mathbf{w}^* + \|\mathbf{y}\|_2^2 = \\ &= \mathbf{w}^T \mathbf{X}^T \mathbf{X}\mathbf{w} - 2\mathbf{w}^T \mathbf{X}^T \mathbf{X}\mathbf{w}^* + \mathbf{w}^{*T} \mathbf{X}^T \mathbf{X}\mathbf{w}^* - [\mathbf{w}^{*T} \mathbf{X}^T \mathbf{X}\mathbf{w}^* + \|\mathbf{y}\|_2^2] = \\ &= (\mathbf{w} - \mathbf{w}^*)^T \mathbf{X}^T \mathbf{X}(\mathbf{w} - \mathbf{w}^*) + c = (\mathbf{w} - \mathbf{w}^*)^T UDU^T(\mathbf{w} - \mathbf{w}^*) + c = \\ &= (U^T(\mathbf{w} - \mathbf{w}^*))^T D (U^T(\mathbf{w} - \mathbf{w}^*)) + c \end{aligned}$$

סעיף 2: בסעיף הקודם, הוכחנו כי מתקיים:

$$\begin{aligned} F(\mathbf{w}) &= (U^T(\mathbf{w} - \mathbf{w}^*))^T D (U^T(\mathbf{w} - \mathbf{w}^*)) + c = (\mathbf{w} - \mathbf{w}^*)^T UDU^T(\mathbf{w} - \mathbf{w}^*) + c \\ &= \mathbf{w}^T UDU^T \mathbf{w} - 2\mathbf{w}^{*T} UDU^T \mathbf{w} + c \end{aligned}$$



נחשב את הגרדיאנט:

$$\nabla F(w) = 2UDU^T w - 2UDU^T w^* = 2UDU^T(w - w^*)$$

נשתמש במבנה של w_t לפי GD:

$$\begin{aligned} w_{t+1} &= w_t - \eta \nabla F(w) = w_t - 2\eta UDU^T(w_t - w^*) = w_t - 2\eta UDq_t \Leftrightarrow \\ w_{t+1} - w^* &= w_t - w^* - 2\eta UDq_t \Leftrightarrow \\ U^T(w_{t+1} - w^*) &= U^T(w_t - w^*) - 2\eta U^T U D q_t \xrightarrow[U^T U = I]{} q_{t+1} = q_t - 2\eta D q_t \end{aligned}$$

סעיף 3: נמשיך את המשוואה שקיבלנו ונפתח אותה לאחר מכן באינדוקציה:

$$q_{t+1} = q_t - 2\eta D q_t = q_t(I - 2\eta D) = \dots = q_1(I - 2\eta D)^t$$

כדי להוכיח את הגבול, מספיק להוכיח בכל קואורדינטה בנפרד. נסתכל על קואורדינטה כללית i . דהיינו מטריצה שהאלביסון שלו מכיל את הערכים העצמיים בסדר יורד. לכן נקבל: $(q_{t+1})_i = (q_1)_i(1 - 2\eta \lambda_i)^t$

אם $\eta < 2\lambda_i$ אז $2\lambda_i < 1$ ולכן $|1 - 2\lambda_i| < 1$ וכך נקבל שאנו שוארים לפחות ב- t מرات. אם $\eta > 2\lambda_i$ אז $1 - 2\lambda_i < 0$ ולכן $|1 - 2\lambda_i|^t \rightarrow 0$

$$|(q_{t+1})_i| < |(q_1)_i| \cdot |(1 - 2\lambda_i)^t| \xrightarrow[t \rightarrow \infty]{} 0$$

שאלה מבחנים – תרגול תשפ"ד (תשפ"ג א' – מועד ב')

Consider the quadratic function

$$f(w_1, w_2) = \frac{a_1}{2}w_1^2 + \frac{a_2}{2}w_2^2$$

for some constants $a_1, a_2 \geq 0$, not both are zero. Note that f is convex ([why?](#)). Assume we run GD on the function f with step size $\eta > 0$, initialized at the point $w^{(1)} = (\frac{1}{2}, \frac{1}{2})$.

1. Prove that the step size which gives a minimum value for f after a single step of GD is $\eta^* = \frac{a_1^2 + a_2^2}{a_1^2 + a_2^2}$.
2. Assume $a_1 = 1$ and we run GD with the step size η^* . Give an intuitive explanation for the fact that the values for a_2 which yield a minimum value for f after a single step of GD are $a_2 = 0$ and $a_2 = 1$.

סעיף 1: תחילת נחשב את הגרדיאנט של f : $\nabla f(w_1, w_2) = (a_1 w_1, a_2 w_2) \Rightarrow \nabla f(w^{(1)}) = \left(\frac{a_1}{2}, \frac{a_2}{2}\right)$

לכן מתקיים לפי הצעד הראשון של GD: $w^{(2)} = w^{(1)} - \eta \nabla f(w^{(1)}) = \left(\frac{1 - \eta a_1}{2}, \frac{1 - \eta a_2}{2}\right)$

$$\begin{aligned} f(w^{(2)}) &= \frac{a_1}{2} \left(\frac{1 - \eta a_1}{2}\right)^2 + \frac{a_2}{2} \left(\frac{1 - \eta a_2}{2}\right)^2 = \frac{1}{8}(a_1^3 \eta^2 - 2a_1^2 \eta + a_1 + a_2^3 - 2a_2^2 \eta + a_2) = \\ &= \frac{1}{8}((a_1^3 + a_2^3)\eta^2 - 2(a_1^2 + a_2^2)\eta) + const \end{aligned}$$

קיבלנו פונקציה של η שהיא קמורה. כדי למצוא את η^* נחשב את הנגזרת לפי η ונשווה לאפס:

$$2(a_1^3 + a_2^3)\eta^* - 2(a_1^2 + a_2^2) = 0 \Leftrightarrow \eta^* = \frac{a_1^2 + a_2^2}{a_1^3 + a_2^3}$$

סעיף 2: נציב $a_1 = 1$ ונקבל: $f(w_1, w_2) = \frac{1}{2}w_1^2 + \frac{a_2}{2}w_2^2$

- באשר $a_2 = 0$ קיבל $f(w_1, w_2) = \frac{1}{2}w_1^2$ והפונקציה כלל לא תלויות- w_2 . מתחילה מ- $(\frac{1}{2}, \frac{1}{2})$, ונרצה להגיע למינימום. אנחנו יודעים כי $(0, 0)$ הוא מינימום. במקרה זה, כל נקודה שבה w_1 הוא אפס מהוות מינימום. לכן נרצה שהגרדיאנט יהיה אוטנו בכיוון $0 = w_1$. לכן, $(0, -\frac{1}{2}) = -\nabla f(w^{(1)})$ – הולך שמאליה ויביא אותנו בצעד אחד ל- $(\frac{1}{2}, 0)$ כי בחרנו η אופטימלי.
- באשר $a_2 = 1$ קיבל $f(w_1, w_2) = \frac{1}{2}w_1^2 + \frac{1}{2}w_2^2$ ובעת רק הראשית היא מינימום, לשם אנחנו רוצים להגיע. הגרדיאנט ציר להיות בכיוון של $(1, 1)$ כדי שנלך מהנקודת $(\frac{1}{2}, \frac{1}{2})$ לדראשת. אכן מתקיים $-\nabla f(w^{(1)}) = (-\frac{1}{2}, -\frac{1}{2})$, שתי הקואורדינטות זהות.



SVM

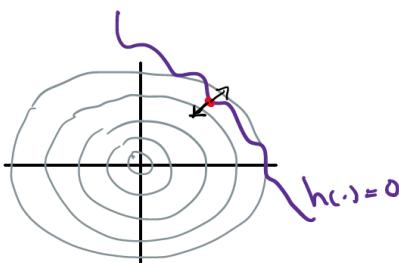
אופטימיזציה עם אילוצים

נתיחס למקורה בו הסת הפייזייל של הפרמטרים הוא לא המרחב בולו. נציג כלים תאורטיים שעוזרים לנו להתמודד עם בעיות מהסוג זהה, ויכולם לתת לנו גם אלגוריתמים לפתרון בעיות כאלה. נראה את זה בא ידי ביוטו בקשר למודל הקלאסיפיקציה הראשון שונראה בשם SVM. בעיית אופטימיזציה בלילית שהיא constrained optimization (תיאור של משוואות ואי-שוויונות כדי שיתיה למנו נוח):

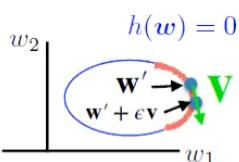
$$\min_w f(w) \text{ s.t. } \begin{cases} h_i(w) = 0 & i = 1, \dots, p \\ r_j(w) \leq 0 & j = 1, \dots, k \end{cases}$$

פונקציות לגראמט:

במצב הפשוט: $\min_w f(w)$ כאשר $0 = f(w) \cdot h$. אם לא היה את האילוץ היחיד h , אז תנאי הכרחי לאופטימליות היה $0 = f'(w)$.

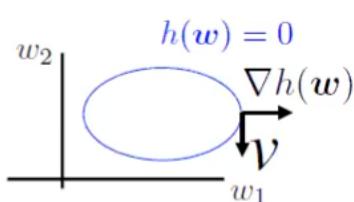
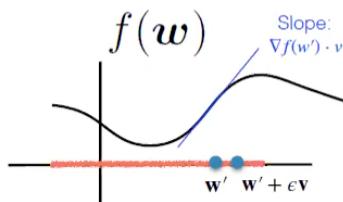


- נכיה שבנקודה האדומה יש לנו את המינימום המקומי. נטען שהגראינט שם חייב להציבו במאונך לירעה באותו נקודה, כי אחרת היינו מקבלים הטלה חיובית על כיוון התנועה על פני הירעה והפונקציה הייתה עולה/ירודת. למעשה זו הבללה של כופלי לגראנט.
- אפשר להגיד אותו הדבר על הגראינט של h – אם הוא היה מציבע לכיוון אחר, היינו יכולים לזרז על פני הירעה ולהגדיל ולהקטין את h , אך הירעה מוגדרת ע"י $0 = h$. בסוף התנאי ההכרחי שנתקבל על מנת שהאופטימליות תתקבל ב- $h = 0$ הוא: $0 = \nabla f(w) + \lambda \nabla h(w)$ ע"י. כלומר, שני וקטורי הגראינט ת"ל.



(אימר): עלינו ללבת לאורך הקוו הכלול, ולבדוק את הגובה של כל נקודה (w, f) . הנקודות היחידות שמשמעותן אותן הן הנקודות שבהן הגראינט של f מתאפסים מבוטן. נניח שהיא בנקודת w' , ונעה בכיוון v המשיק לאילוץ: נגיע לנקודת $w' + \epsilon v$. אנו יודעים כי עברו אפסון: $\frac{f(w' + \epsilon v) - f(w')}{\epsilon} \approx v \cdot \nabla f(w')$ וכאן $f(w' + \epsilon v) \approx f(w') + \epsilon v \cdot \nabla f(w')$.

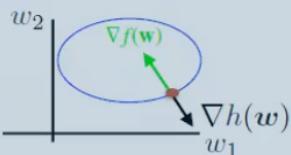
מסקנה 1: הווקטור המשיק מאונך לגראינט של w .



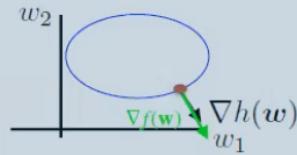
נסתכל על $0 = h(w)$, הפונקציה h קבועה על העקום הזה. אנחנו יודעים שקווי גובה של פונקציה הם מאונכים לגראינט של הפונקציה. לכן בכל נקודה w על העקום, המשיק v מקיים $0 = v \cdot \nabla h(w)$.

מסקנה 2: הווקטור המשיק מאונך לגראינט של h .

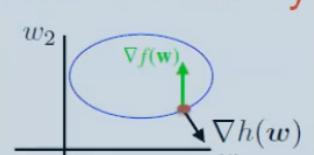
Stationary Point!



Stationary Point!



Not stationary



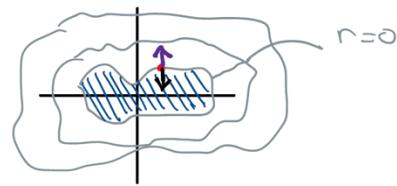
72

ניתן להגדיר פונקציית לגראנט: $L(w, \lambda) = f(w) + \lambda h(w)$. האופטימום ציריך לספק:

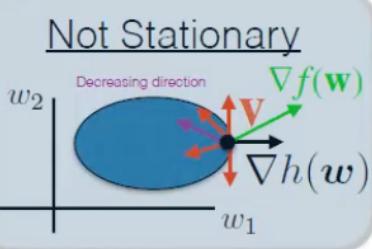
- $\nabla_w L(w, \lambda) = 0$: זה נותן את התנאי של הגראינטים $\nabla f(w) + \lambda \nabla h(w) = 0$.
- $\nabla_\lambda L(w, \lambda) = 0$: זה נותן את התנאי של האילוץ $h(w) = 0$.

במה אילוצי שווין: $\min_w f(w)$ כאשר $0 = h_i(w)$ עבור $i = 1, \dots, p$.

- עכשו התנאי ההכרחי הוא: $0 = \nabla f(w) + \sum_{i=1}^p \lambda_i \nabla h_i(w)$.
- עדין יש תלות לינארית. נגדיר פונקציית לגראנט: $L(w, \lambda) = f(w) + \sum_{i=1}^p \lambda_i h_i(w)$. זו פונקציה שבל אחד מהמשתנים שלה הוא וקטור: λ .
- נדרש גם שיטקיום: $0 = \nabla_\lambda L(w, \lambda) \Leftrightarrow \forall i. h_i(w) = 0$.



אי-שוויון בודד: $\min_w f(w) \leq r^*$ כאשר $w^* \in \mathbb{R}^n$. קיימות שתי אפשרויות:



- האילוץ מתקיים בצורה חזקה: $w^* \in \text{int}(S)$ והוא פנימית ונמצאת בתוך השטח שמתוחת לו.
- האילוץ מתקיים בצורה חלשה: $w^* \in S$ נמצאת על השפה של הסט הפיזיבי, בולם $r(w^*) = 0$. הגרדיינט של r חייב להיות בדיק החוצה, מאונך לשפה. אחרת יש כיוונים שאנו חסנו יכולים לזרז בהם וهم יוצאו מהסט הפיזיבי אבל r לא יגדל, וזה לא יכול להיות מהגדלת r .

- בולם $\nabla r(w^*) = \alpha \nabla f(w^*) - \nabla h(w^*) \geq 0$. כאן הגרדיינט של r צריך להצביע בדיק לפיוון ההרוף מהגדיאנט של f . זה בינויד במקרה הקודם שהמ הוא ת"ל, כאן הם הפוכי ביוון.

- סה"כ נקבל שלושת הדברים הבאים מתקיימים, דרך שוקלה לכתוב את זה:
 - $\nabla f(w^*) = -\alpha \nabla r(w^*)$
 - אחד מהגורמים שווה לאפס $\alpha r(w^*) = 0$
 - $\alpha \geq 0, r(w^*) \leq 0$

אילוצי שוין ואי-שוויון: נראה תוצאה כללית לאילוצים שהכרחיים לאופטימליות. עברו בעיה אופטימיזציה כללית:

$$\min_w f(w) \text{ s.t. } \begin{cases} h_i(w) = 0 & i = 1, \dots, p \\ r_j(w) \leq 0 & j = 1, \dots, k \end{cases}$$

נגידר את פונקציית לגראנץ בצורה הבאה, עברו $w \in \mathbb{R}^n, \lambda \in \mathbb{R}^p, \alpha \in \mathbb{R}^k$

$$L(w, \lambda, \alpha) = f(w) + \sum_{i=1}^p \lambda_i h_i(w) + \sum_{j=1}^k \alpha_j r_j(w)$$

תנאי KKT:

התנאים הללו הכרחיים עבור אופטימיזציה. אם הבעיה קמורה: f ו- r_j קמורות, $-h_i$ אפוניות, אז תנאי ה-KKT גם מספקים לאופטימיזציה. תהי נקודת אופטימלית w^* . קיימים $\lambda^* \in \mathbb{R}^p, \alpha^* \in \mathbb{R}^k$ כך שמתקיים התנאים הבאים:

$$\begin{aligned} h_i(w^*) = 0 & \quad i = 1, \dots, p \\ r_j(w^*) \leq 0 & \quad j = 1, \dots, k \end{aligned} \quad \text{:Primal feasibility .1}$$

$$\alpha_j \geq 0, j = 1, \dots, k \quad \text{:Dual feasibility .2}$$

$$\nabla_w L(w^*, \lambda^*, \alpha^*) = 0 \quad \text{:Stationarity .3}$$

$$\alpha_j r_j(w^*) = 0, j = 1, \dots, p \quad \text{:Complementary slackness .4}$$

דואליות:

מעבר בעיה אופטימיזציה כללית: $\min_w \max_{\lambda \in \mathbb{R}^p, \alpha \in \mathbb{R}^k} L(w, \lambda, \alpha) \text{ s.t. } \begin{cases} h_i(w) = 0 & i = 1, \dots, p \\ r_j(w) \leq 0 & j = 1, \dots, k \end{cases}$ זה שקול ל: $\max_{\lambda \in \mathbb{R}^p, \alpha \in \mathbb{R}^k} \min_w L(w, \lambda, \alpha)$ לפי הגדרה מתקיים: $L(w, \lambda, \alpha) = f(w) + \sum_{i=1}^p \lambda_i h_i(w) + \sum_{j=1}^k \alpha_j r_j(w)$. מספיק שאחד מה- h_i שונה מ-0, יוכל לעיל ידי הקואורדינטה המתאימה להשאיפ את זה לאינסוף. אם קיימים j שונה מ-0, יוכל גם באותו אופן להשאיף לאינסוף. לכן:

$$\max_{\lambda \in \mathbb{R}^p, \alpha \in \mathbb{R}^k} L(w, \lambda, \alpha) = \begin{cases} \infty & \exists i. h_i(w) \neq 0 \vee \exists j. r_j(w) > 0 \\ f(w) & \text{otherwise} \end{cases}$$

למזור את הבעיה זו, זו בדיק הבעיה שלנו.

נניח שאנו קובעים w^*, λ^*, α^* . מתקיים: $L(w^*, \lambda^*, \alpha^*) \geq \min_w L(w, \lambda, \alpha)$. כי אם נמזור את w את נשאר את השאר אותו דבר יוכל רק למדער. נשאר את w קבועה, ונמוקם על פניו α, λ . נקבע:

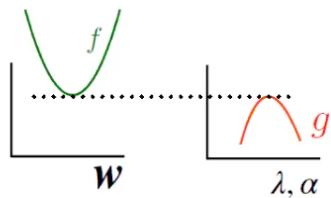
$$\max_{\lambda \in \mathbb{R}^p, \alpha \in \mathbb{R}^k} L(w^*, \lambda, \alpha) \geq \max_{\lambda \in \mathbb{R}^p, \alpha \in \mathbb{R}^k} \min_w L(w, \lambda, \alpha)$$

זה מתקיים לכל w ולבן גם על המינימלי מבין כל $-w$, שכן:

$$\left[\min_w \max_{\lambda \in \mathbb{R}^p, \alpha \in \mathbb{R}^k} L(w, \lambda, \alpha) \right] \geq \left[\max_{\lambda \in \mathbb{R}^p, \alpha \in \mathbb{R}^k} \min_w L(w, \lambda, \alpha) \right]$$



- Define: $g(\lambda, \alpha) = \min_w \mathcal{L}(w, \lambda, \alpha)$
- Dual problem: $\max_{\lambda, \alpha \geq 0} g(\lambda, \alpha)$



את הבעה הפרימלית נסמן ב-(P), ואת הדואלית ב-(D). הערך (D) – (P) נקרא **duality gap**. כלומר, אם נפתרו את הבעה הדואלית, הפתרון שלה הוא חסם תחתון לפתרון הבעה שלנו.

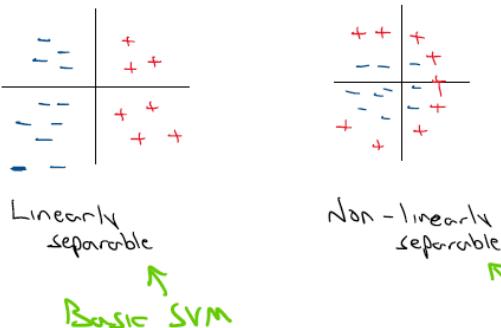
דואליות חזקה: אם הבעה הפרימלית היא קמורה, אז תחת תנאים כלים: $0 = \text{duality gap}$. כלומר, ניתן לפתור את הבעה הפרימלית על ידי מעבר לבעה הדואלית, לפתור אותה, ואז הפתרון שלhn זהה.

- אם w^* אופטימלי ל-(P) ו- α^* אופטימליים ל-(D) וגם מתקיימת דואליות חזקה, אז $(\alpha^*, \lambda^*, w^*)$ מקיימים את תנאי ה-KKT.
- לא תמיד יותר קל לעבוד עם הדואלית, אבל לעיתים זה יכול לעזור לנו להבין את הפרימלית טוב יותר.

Basic SVM

רקע:

מודל בסיסי לקלסיפיקציה לינארית. נטרץ במרקחה זהה בклסיפיקציה ביןארית, כאשר $\{y_i\}_{i=1}^n \in \mathbb{R}^d$. נניח כי $\mathcal{X} = \mathbb{R}^d$ וכתוון לנו סט אימון בגודל n : $\{(x_i, y_i)\}_{i=1}^n$. נרצה להוציא לפולט מסווג $\mathbb{R} \rightarrow \mathcal{X}: h$. כך ש:



kernel (non-linear) SVM

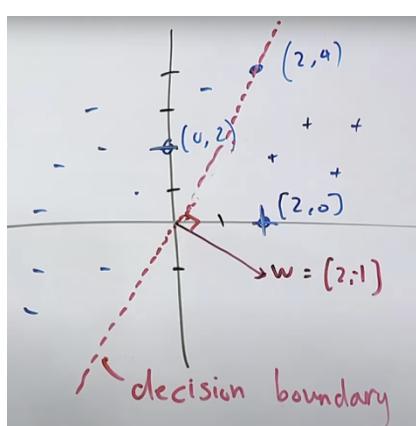
- אם $y_i = 1$ נקבל $0 \geq h(x_i)$.
- אם $y_i = -1$ נקבל $0 < h(x_i)$.

נפוך את זה להיפותזה באמצעות הפעלת פונקציית margins. יכולות להיות התפלגויות שונות לדואט. במרקחה הבסיסי, SVM מותן מעבה לאטא שהיא יתירה. ניתן להרחב אותה גם למקרה ה-non-linear (נקרא kernel svm).

למושוג לינארי יש את הצורה הבאה: $b + w \cdot x = h(x)$ כאשר המשקלות הן $w \in \mathbb{R}^d$ והן $b \in \mathbb{R}$. אם נתבונן במרקחה שהדאטה מתפלג במרקחה הבסיסי, ויש 2 מפרדים לינאריים עליון – נצטרכן להגדיר מיהם יותר טוב. נראה הגיוני להגדיר שהירוק יותר טוב: הוא יותר רובייטי, אם בכינס עוד נקודות הסגול יפגע לפני הירוק, נאמר שלירוק יש margin יותר גדול. נשאף למצוא את המושוג הלינארי עם המargin הכי גדול.

מושוג לינארי $b + w \cdot x = h(x)$ בעל גבול החלטה (decision boundary) שהוא על מישור מסווג (hyperplane) שכינצ'ב ל-w. זהו הגבול שמכריע בין נקודות שיסווגו + או -.

טענה (margin): המרחק האוקלידי בין נקודה $x \in \mathbb{R}^d$ וヶ月ור מפריד המוגדר על ידי h הוא: $\frac{|w \cdot x + b|}{\|w\|}$.



- זהו ממד גאומטרי שאיןנו רגיש לגודל, רק הכוון משתנה.

המרחק של המקור $0 = x$ הוא פשוט $\frac{|b|}{\|w\|}$, זהה דרך גאומטרית לפרש את ה-bias. הוא אומר באיזה מרחק w צריך להסיט מהריאשטי את hyperplane המפריד לפי הכוון w .

שיעור בנושא SVM.

דעתא שניתן להפרדה לנארית (Hard SVM):

במצב בו הדעתא נתית להפרדה לנארית (realizable), ניתן לעשות fit מושלם לדעתא עם מסוג לנארי, ואז נרצה למצוא את הישר שמן פריד בעל ה- margin הבי גודל. נגידור את ה- margin המרחק המינימלי של נקודות מהישר המפריד:

$$d(\mathbf{w}, \mathbf{b}) = \min_i \frac{|\mathbf{w} \cdot \mathbf{x}_i + b|}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|} \min_i |\mathbf{w} \cdot \mathbf{x}_i + b|$$

נרצה לפטור: $\max_{\mathbf{w}, b}$ בך שלכל i יתקיים $0 \geq (\mathbf{b} + \mathbf{w}) \cdot \mathbf{x}_i$. בולם למקסם את ה- margin בתנאי שהפרדי קציה נכונה. האילוצים הם אילוצי אי-שוויון אבל הביטוי של $(\mathbf{b} + \mathbf{w}) \cdot \mathbf{x}_i$ מסובך ולא ברור אם הוא קמור. נביא אותו לצורה שהיא לנו יותר נוח לעבוד איתה.

נשים לב שאם $(\mathbf{b}^*, \mathbf{w}^*)$ הוא פתרון לבעה, אז גם $(\mathbf{c}\mathbf{w}^*, \mathbf{c}\mathbf{b}^*)$ הוא פתרון לכל $0 > c$ (קבוע חיובי). יש לנו כאן דרגת חופש, ואנחנו יכולים לנормל על ידי זה שמדרווש: $1 = |\mathbf{b}^* + \mathbf{w}^* \cdot \mathbf{x}_i|$.

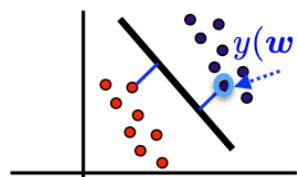
המצב היחיד בו לא נוכל לעשות את זה הוא אם קיימת נקודה שיוושבת על הישר המפריד, ועבורה זה שווה 0. מכיוון שככל המטרה של בעיית האופטימיזציה זה למקסם את ה- margin , כל עוד נתן להציג margin חיובי לא נהיה בסיטואציה זו (מקרה קצה). נקבל מחלוקת שיקולות ועל ידי הנרמול קיבענו נציג בכל מחלוקת שיקולות כזו. נקבל את הבעה הבאה (בי המונח שווה בעת ל-1):

$$\max_{\mathbf{w}, b} \frac{1}{\|\mathbf{w}\|} \text{ s.t. } \begin{cases} \forall i. y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 0 \\ \min_i |\mathbf{w} \cdot \mathbf{x}_i + b| = 1 \end{cases}$$

- נשים לב שהאליעזר הראשון אומר כי $|\mathbf{b}^* + \mathbf{w}^* \cdot \mathbf{x}_i| = 1$ והוא יכול להיות $\{-1, 1\}$ – ולכן בערך מוחלט זה יהיה 1. וכן נוכל לבתוב את בעיית האופטימיזציה שלנו באופן הבא: $1 = \min_i |\mathbf{w} \cdot \mathbf{x}_i + b|$.
- זה יהיה האופטימום גם של הדבר הבא: $1 \geq \min_i |\mathbf{w} \cdot \mathbf{x}_i + b|$. על פניו החלשנו את האילוצים וככה ניתן לקבל משאו ורק יותר טוב: $\|\mathbf{w}\|$ גדולה יותר בך שנתקבל שקטן יותר. אבל זה לא יכול להיות – לא נוכל לקבל משאו יותר טוב: האופטימום של הבעה החדשה יכול רק לגדול משל הבעה הקודמת, וכך אם אי השוויון מתקיים בזירה חזקה, המינימום ממש גדול מ-1. במקרה הזה ניתן לעשות scaling down ולהקטין את \mathbf{b} , \mathbf{w} , לשפר אותו, וזה עדין יתקיים בזירה חלהה.
- הצעד האחרון שנעשה – למקסם את $\|\mathbf{w}\|$ שווה להביא למינימום את $\frac{1}{2} \|\mathbf{w}\|^2$ וכן נקבל את הבעה הקמורה:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \text{ s.t. } \forall i. y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$$

מקובל לשים $\frac{1}{2}$ כדי שלאחר גזירה של ביטוי שהוא בחזקת 2 המקדם יעלם.



Support Vectors (SV): הנקודות (\mathbf{x}_i, y_i) שעבורן מתקיים: $1 = y_i(\mathbf{w} \cdot \mathbf{x}_i + b)$, בולם המרחק שלהם מהישר המפריד הוא מינימלי, הנקודות שהכ密切 קרובות למשור.

אופטימיזציה (Hard SVM) – פיתוח מהשיעור:

הבעיה הפרימילית: נרצה לפטור את הבעה הבאה:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \text{ s.t. } \forall i. y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$$

זהו בעה ריבועית (quadratic), בעיה קמורה כאשר objective הוא ריבועי והailוצים הם אפיניים. נעבור לבעיה הדואלית של ה-SVM ונראה איך פותרים.

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \Leftrightarrow r_i(\mathbf{w}) := 1 - y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \leq 0$$

פונקציית לוגראם' היא: $(\mathbf{1} - y_i(\mathbf{w} \cdot \mathbf{x}_i + b))^2 = \frac{1}{2} \|\mathbf{w}\|^2 + \sum_i \alpha_i (\mathbf{1} - y_i(\mathbf{w} \cdot \mathbf{x}_i + b))$.

זהו בעה קמורה, הפתרון של הפרימילית שווה לפתרון של הדואלית. אם יש לנו פתרון \mathbf{w}^* , b^* , α^* לבעה שלנו, אז קיימם פתרון לבעה הדואלית α^* בך ש- $(\alpha^*, \mathbf{w}^*, b^*)$ מקיימים את תנאי KKT, כלומר:

- סטציונריות: $\sum_i \alpha_i^* y_i \mathbf{x}_i = 0 \Rightarrow \mathbf{w}^* = \sum_i \alpha_i^* y_i \mathbf{x}_i$.
- slackness: $\alpha_i^*(1 - y_i(\mathbf{w}^* \cdot \mathbf{x}_i + b^*)) = 0 \Rightarrow 1 - y_i(\mathbf{w}^* \cdot \mathbf{x}_i + b^*) = 0 \Rightarrow \alpha_i^* = 0$.
- לא SV לפוי אין שהגדכנו אותו, ונכתב: $\sum_{i:(x_i, y_i) \text{ is a SV}} \alpha_i^* y_i \mathbf{x}_i = \mathbf{w}^*$, בולם צ"ל של ה-SV.



הבעיה הדואלית: הבעיה הדואלית שלנו היא: $\min_{w,b} L(w,b,\alpha)$. מתקיים:

$$\begin{aligned} g(\alpha) &= \min_{w,b} L(w,b,\alpha) = \min_{w,b} 0.5\|w\|^2 + \sum_i \alpha_i(1 - y_i(w \cdot x_i + b)) \\ &= \min_{w,b} 0.5\|w\|^2 + \sum_i \alpha_i(1 - y_i(w \cdot x_i)) - \left(\sum_i \alpha_i y_i\right)b \end{aligned}$$

- אם α מקיים $0 \neq \sum_i \alpha_i y_i \neq \infty$ אז $= (\alpha)g$, כי אם נפתח סוגרים נראה שהתלות ב- b לינארית, ועל ידי משחק עם b אפשר לקחת את הכלול למינוס אינסופי.

- אחרת, אז נקבל: $(\alpha)g = \min_{w,b} 0.5\|w\|^2 + \sum_i \alpha_i(1 - y_i(w \cdot x_i))$. עבור w^* מינימלי מתקיים כי הערך המינימלי הוא $\sum_i \alpha_i - 0.5 \cdot \sum_{ij} \alpha_i \alpha_j y_i y_j x_i \cdot x_j$.

נכתב את הבעיה הדואלית בצורה הבאה:

$$\max_{\alpha} \sum_i \alpha_i - 0.5 \cdot \sum_{ij} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \text{ s.t. } \begin{cases} \forall i. \alpha_i \geq 0 \\ \sum_i \alpha_i y_i = 0 \end{cases}$$

זו בעיה קמורה. ניתן לכתבו בצורה מטריצונית: $\sum_i \alpha_i x_i - 0.5 \alpha^T G G^T \alpha$ כאשר G היא מטריצה שבשורה ה- i שלה ישוב x_i .

- לבעיה הדואלית יש פרמטרים שהם מספר הדוגמאות שאימנו עליה. הבעיה הפרימילית היו d פרמטרים. זה יכול להיות יתרון גדול בשחה- d ממש גדול. אפשר לפתור SVM עם מימד DATA עצום (אינסופי גם) ואז לעבור לבעיה הדואלית שהמייד שהזה מספר הדוגמאות.
- אם d גדול מאוד ויש לנו דרך לחשב את $x_i \cdot x_j$ ביעילות, אפשר לעבור לבעיה הדואלית לפתור את SVM ללא עובודה ב- \mathbb{R}^d . זה מה שנשותטמש בו ב-SVM kernel. נוכל לעבור למרחבים שבהם נוכל לחשב מכפלות פנימיות ביעילות, וכך להשתמש גם במסוגים לא לינאריים.

מעבר מהפתרון של הדואלית לפרימילית: ראיינו שבහינתן פתרון לבעיה הדואלית α נקבע $\sum_i \alpha_i^* y_i x_i = w^*$. מיידית נוכל לחשב את המשקלים. על מנת לקבל את ה-bias האופטימלי b מהתמונה של slackness $= 0 = \sum_i \alpha_i^* (1 - y_i(w^* \cdot x_i + b^*))$. אם $\alpha_i^* \neq 0$ עבור i בלבד, אז $y_i(x_i, w^*)$ הוא SV. לכן, אנחנו יודעים כי $0 = \sum_i \alpha_i^* (1 - y_i(w^* \cdot x_i + b^*))$ ואז אפשר לחשב: $y_i = b^* - w^* \cdot x_i$

חסמי הכללה (Hard SVM):

חסם margin-based: החסם הזה מניח שהדאטא כן ניתן להפרדה לינארית (בלומר realizable). נניח שמתקיים התפוגות שניתנת להפרדה לינארית כך ש- $R \leq \|x\|$ בהסתברות 1. ניתן להראות כי:

$$\text{SVM true error of learned predictor} \leq \frac{C_1}{n} \cdot \frac{R^2}{\gamma^2} + \frac{C_2}{n} \cdot \log \frac{n}{\delta}$$

בהתבסירות גדולה שווה $M - \delta - 1$ באשר M מסpter הדוגמאות, γ זה ה- $\text{h}\text{-margin}$ קבועים כלשהם. ככל שה- $\text{h}\text{-margin}$ יותר גדול נתקבל הבטחה יותר טובה. נשים לב שאין תלות במימד d ולכן ניתן להפיע את החסם הזה גם אם המימד הזה מאד גדול.

חסם LOO (Leave One Out): בהינתן סט אימון S_n , נגדיר את $e_{LOO}(S_n)$ להיות הסט שמתפרק לאחר מחיקת הדוגמה ה- i . נעשה ניסוי אחדנו מורידים את הדוגמה ה- i , מאמנים Hard SVM ובודקים אם טעינו. נחזיר על זה לכל הדוגמאות ונחשב את השגיאה הממוצעת:

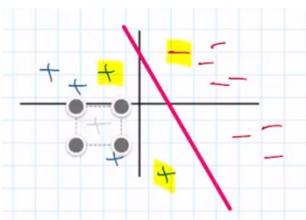
$$e_{LOO}(S_n) = \frac{1}{n} \sum_{i=1}^n \mathbb{I} [h_{S_n^{-i}}(x_i) \neq y_i]$$

זה כמו cross validation – כל פעם משתמשים בדוגמה אחרת כדי לאמוד את השגיאה.

ניתן להראות כי: $\mathbb{E}_{S_n} [e_{LOO}(S_n)] = \mathbb{E}_{S_{n-1}} [\mathbb{E}_{S_n} [e_P(h_{S_n^{-i}}(x_i) \neq y_i)]]$ – אם נגריל סט בגודל n , השגיאה הממוצעת שווה לשגיאה הממוצעת של הגרלת סט קטן ב-1 ולקיים השגיאה הממוצעת.

בבירור מתקיים: $\mathbb{E}_{S_{n-1}} [e_P(h_{S_n^{-i}}(x_i) \neq y_i)] \leq \frac{\#SV}{n}$, כי אם היה ממשו שלא היה SV, ונאמן בReLU, נגיע לאוותנו הפתרון. הוא לא ישפייע על הפתרון אם נשלוף אותו החוצה וכן נוכל לטועות רק על ballo שחיי SV.

$$\mathbb{E}_{S_{n-1}} [e_P(h_{S_n^{-i}}(x_i) \neq y_i)] \leq \frac{1}{n} \mathbb{E}_{S_n} [\#SV]$$





מספר ה-SV מאוד תלוי בדעת, לפחות 2. לא משנה באיזה מימד נעבד, ה-SV זה הנקודות הכי קרובות. מצד שני, יכול להיות מקרה מאוד רע בו כולם SV. ניתן להראות במקורה הכללי, שאם נגריל דआטא מספר ה-SV יהיה בערך המימד של המרחב, אבל אז נקבל משהו שתלוי ב-VCdim.

דאטא שאיןו ניתן להפדה לינארית (Soft SVM) – פיתוח מלא בתרגול 6:

נעביר בעת למקורה הנפוץ יותר, שאינו realizable. במקרה זה, אנחנו מוסיפים משתני slack. במקום לדרוש עמידה באילוץ באופן כללי, נאפשר אי-עמידה באילוץ: $y_i(w \cdot x_i + b) \geq 1 - \xi_i$ וונסה למצער את החירגה שאפשרנו:

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \quad s.t. \quad \forall i. \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

באשר C קבוע חיובי שנקבע מראש.

- נשים לב שהה **תמיד פיזיביל** – נוכל למצאו C, b, w שמקיימים את האילוצים, על ידי כך שנגדיל את ξ מספיק.
- הפרמטר C** קבוע כמה נתיחס בחומרה להפרות – אם הוא מאוד קטן, זה לא נראה שיש הפרעות. נתרץ במצער של $\|w\|$, אחרת יש לכל הפרה משקל גדול יותר. יש כאן trade off ולכן זה קשור לרגוליזציה – ככל ש- λ גדול יותר נגדיל את מחלוקת ההיפוטזות. התפקיד של C דומה אבל הפוך – ככל שהוא קטן יותר נוכל להפר יותר, הוא **פרמטר רגוליזציה**. אפשר להשתמש ב-cross validation כדי למצוא את ה-C המתאים.
- אפשר לקבל את **הבעיה הדואלית**, וההבדל היחיד הוא שמוסיפים את האילוצים $\leq \alpha$ (בתרגול 6).

נפתח את הבעיה שהציגנו, כי לכל ξ יש שני אי-שוויונות שצרכי לקיים: $\{(x \cdot w + b) + \xi_i \leq 1 - y_i(w \cdot x_i + b)\}$

נגדיר $\lambda = \frac{1}{2C}$ וכן: $\min_{w,b} \lambda \|w\|^2 + [\sum_{i=1}^n \max\{0, 1 - y_i(w \cdot x_i + b)\}]$. לכן **נסיק כי Soft SVM זה מינימיזציה של hinge loss עם רגוליזציה של ℓ_2** .

תרגול 6 (Soft SVM)

דיברנו על GD ואופטימיזציה בשានן לנו אילוצים על המשקלים שלנו. בפרט, הבעיות שמעניינות אותנו הן SVM: בהינתן סט של נקודות במרחב, למצוא מריד לינארי שמפריד אותן טוב ועם גודל margin גבוה – המרחק שלו מהנקודות הוא הכי גדול שאפשר, וזה אינטואיטיבית ייתן לנו הבהלה.

אופטימיזציה עם אילוצים:

אנחנו רוצים למצער פונקציה $\mathbb{R} \rightarrow \mathbb{R}^d$: כאשר אנחנו נתונים ל-N אילוצי אי-שוויון $0 \leq r_i(x) \cdot h_j(x) \leq 1 - M$ אילוצי שווין $0 = r_i(x) \cdot h_j(x)$. נקרא **הבעיה הפרימילית**. הבעיה היא קמורה, אם f קמורה והאילוצים מגדרים קבוצה קמורה.

$$\min_{x \in \mathbb{R}^d} f(x) \quad s.t. \quad \begin{cases} r_j(x) \leq 0 & i = 1, \dots, N \\ h_i(x) = 0 & j = 1, \dots, M \end{cases}$$

הfonקציה העיקרית שאנו עושים אותה היא הלגראנז'אן \mathcal{L} , שמשלבת בין הfonקציה לבין האילוצים:

$$\mathcal{L}(x, \alpha, \beta) = f(x) + \sum_{i=1}^N \alpha_i r_i(x) + \sum_{j=1}^M \beta_j h_j(x)$$

באשר המקדים β_j, α_i הם כפולי לגראנט.

- אפשר לתאר את הבעיה שלנו בתור אופטימיזציה על \mathcal{L} : $\min_x \max_{\alpha \geq 0, \beta} \mathcal{L}(x, \alpha, \beta)$. קודם אנחנו מפרקמים את \mathcal{L} על α, β , עבור x קבוע, ואז מוצאים את ה- x שמצויר את זה.
- הבעיה הדואלית היא: $\mathcal{L}(x, \alpha, \beta) = \min_x \max_{\alpha, \beta} g(\alpha, \beta)$. קודם אנחנו מסתכלים על ה- x שמצויר את \mathcal{L} עבור β, α קבועים ואז מפרקמים את זה על β, α .

דויאליות: בהינתן x לפרימילית $-(*, \beta^*, \alpha^*)$ פתרון לדואלית.
תמיד מתקיים: $\min_x \mathcal{L}(x, \alpha^*, \beta^*) \leq f(x^*)$. **עבור בעיות קמורות מתקיים weak duality**: $\min_x \mathcal{L}(x, \alpha^*, \beta^*) = \min_x g(\alpha^*, \beta^*)$. **עבור בעיות קמורות מתקיים strong duality**: $f(x^*) = g(\alpha^*, \beta^*)$. כלומר, $f(x^*) = g(\alpha^*, \beta^*)$.



באמצעות תנאי KKT, נוכל לחלץ את x^* בהינתן הפתרון (α^*, β^*) :

1. Primal feasibility $\forall i, j. r_i(x^*) \leq 0, h_i(x^*) = 0$
2. Dual feasibility $\forall i. \alpha_i^* \geq 0$ (קיים אילוצי דואלי: $\sum_i \alpha_i^* r_i(x^*) = 0$)
3. Stationarity (הגדריאנט של הלגראנט אין מתאפס במנזער x^*): $\nabla_{x^*} \mathcal{L}(x^*, \alpha^*, \beta^*) = \nabla f(x^*) + \sum_{i=1}^N \alpha_i^* \nabla r_i(x^*) + \sum_{j=1}^M \beta_j^* \nabla h_j(x^*) = 0$
4. Complementary slackness ($\forall i. \alpha_i^* r_i(x^*) = 0$) (תמיד המכפלת באילוצי אי-שוויון היא אפס)

:Soft SVM

ראינו את בעיית-hard SVM, על דاطה שהוא linearly separable. מבחן ביעילות הלמידה, תמיד אפשר לקחת בעיה עם bias ולהפוך אותה לבעיה חדשה בלי bias (локחים את כל ה- α , משרשים להם עוד 1 בסוף ומגדלים את המינימום של הבעיה ב-1, עבשו, (feature), bias, ועוד) (תמיד המכפלת הפנימית שלהם עם ש' יהיה עוד ערך נלמד בסוף והוא ייקח את התפקיד של bias, הוספנו עוד לבני מודולר).

$$\min_w \frac{1}{2} \|w\|^2 \text{ s.t. } \forall i. y_i(w \cdot x_i) \geq 1$$

מה נעשה אם הדאטא הוא לא linearly separable? לקבל מפריד שמנצער את zero-one loss זו בעיה NP-קשה. או אפשר לציפות שלגורייתם ועל יפותו את הבעיה הזאת. ננסה **לקבר את הפתרון**.

נסיף משתנים לבעיה שיאפשרו לנו לשנות את השגיאה הזאת: ξ , נסיף את ξ , לנדרוש שניהה גדולים שווים מ-1, אלא מ- $\xi - 1$. אם במקרה לא יוכל להפריד נקודה ולהגיע ממש ל-1, אז יש לנו טוח שגיאה ולא נסוג מושלם את הדאטא.

איפה אנחנו נגענים? ב-objective function. ה- ξ מכמתים לנו כמה רוחקים אנחנו מהטוויג המושלים, ואנו מוסיפים את הסכום שלהם עם קבוע בולשוו C .

$$\begin{aligned} \min_{w, \xi} & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \\ \text{s.t. } & \forall i. y_i(w \cdot x_i) \geq 1 - \xi_i, \quad \xi_i \geq 0 \end{aligned}$$

נקבל את פונקציית לגראנט' הבאה:

$$\mathcal{L}(w, \xi, \alpha, \beta) = \left[\frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \right] + \sum_{i=1}^n \alpha_i [1 - \xi_i - r_i(w, \xi)] + \sum_{i=1}^n \beta_i [-\xi_i]$$

הבעיה הפרימלית:

המשתנים הפרימלים הם ξ, w , והדוואלים הם β, α כמו קודם. כל האילוצים שלנו הם אילוצי אי-שוויון (יש לנו שניים). הפונקציה הדואלית מוגדרת כרך: $(\alpha, \beta) = \min_{w, \xi} \mathcal{L}(w, \xi, \alpha, \beta)$. נבחן כי \mathcal{L} עבר β, α , w, ξ קבועים היא פונקציה קמורה של ξ, w . لكن כדי למנוע אותה, נגזר אותה לפי ξ, w ולהשוו ל-0, ואז נבין כיצד נראה ההפונקציה (α, β) .

- נגזר לפי w : נורמה בריבוע, הנגזרת שלו זה $2w$ ועם הפקטור 0.5 זה נניה w . לאחר מכן יש לנו פונקציה לינארית ונישאר עם המקדים של הוקטור שעושים לו מכפלת עם w .

$$\nabla_w \mathcal{L} = w - \sum_{i=1}^n \alpha_i y_i x_i = 0 \Leftrightarrow w = \sum_{i=1}^n \alpha_i y_i x_i \quad (1)$$

- נגזר לפי ξ : מהסכום הראשון נישאר עם C . מהסכום השני נישאר עם α_i , ומהשלישי עם $-\beta_i$.

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \Leftrightarrow \alpha_i = C - \beta_i \quad (2)$$

נציב הכל בחזרה בתרור \mathcal{L} ונרצה להישאר עם β שהיא פונקציה רק של α . תחיליה נשבכט קצת את הסדר:

$$\mathcal{L}(w, \xi, \alpha, \beta) = \frac{1}{2} w \cdot w + \left[\sum_{i=1}^n \alpha_i [1 - \xi_i - y_i(w \cdot x_i)] \right] + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \beta_i \xi_i$$

(*)



באשר:

$$(*) = \sum_{i=1}^n \alpha_i [1 - \xi_i - y_i(w \cdot x_i)] = \sum_{i=1}^n \alpha_i - \alpha_i \xi_i - w \alpha_i y_i x_i = -w \left(\sum_{i=1}^n \alpha_i y_i x_i \right) + \left(\sum_{i=1}^n \alpha_i \right) - \left(\sum_{i=1}^n \alpha_i \xi_i \right)$$

כעת נוכל להציב את (1) ו-(2) ונקבל את g הבא:

$$\begin{aligned} g(\alpha, \beta) &= -\frac{1}{2} \left[\sum_{i=1}^n \alpha_i y_i x_i \right] \cdot \left[\sum_{i=1}^n \alpha_i y_i x_i \right] + \left(\sum_{i=1}^n \alpha_i \right) - \sum_{i=1}^n \alpha_i \xi_i + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \beta_i \xi_i \\ &= -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i \cdot x_j + \left(\sum_{i=1}^n \alpha_i \right) (C - \sum_{i=0}^n \beta_i) \sum_{i=1}^n \xi_i = -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i \cdot x_j + \sum_{i=1}^n \alpha_i \end{aligned}$$

אכן קיבלנו ש- g לא מכילה תלויות במשתנים הפרימיטיבים – כך תמיד אמרו לצאת לנו עבור פונקציה דואלית.הבעיה הדואלית:

במקרה $0 \leq \alpha_i \leq C - \beta_i$ ו- $\beta_i = C - \alpha_i$ לכל $i \in [n]$ נוכל לבתוב במילויים אחרים $C \leq \alpha_i \leq \beta_i$ לכל $i \in [n]$. ניעזר בעובדה כי: $\max_{\alpha, \beta} -g(\alpha, \beta) = \max_{\alpha, \beta} \min_{w, \xi} \mathcal{L}(w, \xi, \alpha, \beta) = \min_{\alpha, \beta} -g(\alpha, \beta) = \min_{\alpha, \beta} g(\alpha, \beta)$

$$\begin{aligned} \min_{\alpha, \beta} -g(\alpha, \beta) &= \min_{\alpha, \beta} \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i \cdot x_j - \sum_{i=1}^n \alpha_i \\ \text{s.t. } &\forall i. C \geq \alpha_i \geq 0 \end{aligned}$$

מה שונה בין זה לבין בעיית ה-SVM? לא היה לנו את הפתרון הדואלי כאן הוא ש- α חסומים ע"י C .

הערות:

- בעיה הפרימיטיבית היה לנו את $\mathbb{R}^d \in w$, מימד שיכל להיות גדול מאוד. בעיה הדואלית α הוא במימד מספר הדוגמאות, מה שמקל علينا לפטור.
- בהינתן שפתרו את הבעיה הדואלית ומצאו לנו את α^* , נוכל לחשב $x_i \cdot x_j$.
- אפשר להיעזר ב-slackness כדי למצוא את ξ^* .

Kernel SVM

דברינו על המקירה בין יש ואין הפרדה ביןארית, עבור מסוגים ביןאריים. הרבה פעמים יש להם approximation error גדול, כי טוב מחלוקת ההיפוטזות מושפעת מיכולת הכללה ויצוג. גישה אחת שראינו (בהקשר ל- bias-variance tradeoff – לשנות את הייצוג, במקרה להפעיל על תמונה נוכל לבנות להן ייצוג מיוחד, ואז נקטין את השגיאה ותהייה לנו הכללה טובה. גישה אחרת היא להשתמש ביצוגים מוכרים, אחת מהן היא Kernel SVM: מיפויים גנריים שימושים את הייצוג, ולמרות שימושו החדש המסוג הוא לינארי, מעל הייצוג הישן הוא לא יהיה לינארי.

:Hard SVM (QP) Quadratic Program

$$\max_{\alpha \in \mathbb{R}^n} \sum_i \alpha_i - 0.5 \cdot \sum_{ij} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \quad \text{s.t. } \sum_i \alpha_i y_i = 0, \alpha_i \geq 0$$

$$\begin{aligned} \min_{\alpha} \quad & 0.5 \alpha^T \underbrace{\begin{bmatrix} y_1 y_1 x_1 \cdot x_1 & y_1 y_2 x_1 \cdot x_2 & \dots & y_1 y_n x_1 \cdot x_n \\ y_2 y_1 x_2 \cdot x_1 & y_2 y_2 x_2 \cdot x_2 & \dots & y_2 y_n x_2 \cdot x_n \\ \vdots & \vdots & \vdots & \vdots \\ y_n y_1 x_n \cdot x_1 & y_n y_2 x_n \cdot x_2 & \dots & y_n y_n x_n \cdot x_n \end{bmatrix}}_M \alpha - \mathbf{1}^T \alpha \\ \text{s.t.} \quad & y^T \alpha = 0 \\ & \alpha \geq 0 \end{aligned}$$

ניתן לבתוב אותה בצורה מטריצונית (ນבצע מינימום על מינוס המשוואה הקודמת). נשים לב כי אפשר לבתוב $P P^T = M$ באשר:

$$P = \begin{bmatrix} & & & \\ - & y_i x_i & - & \\ & & & \end{bmatrix}$$



ולכן M היא מטריצת PSD, $0 \leq M$.
בالمור ה-*quadratic objective* הוא קמור והאלוצים把他 לינאריים. נניח שקיים פתרון את זה ביעילות ע"י אלגוריתם שהוא קופסה שחורה, ונשתמש בה כדי לישם SVM לא לינארי.

SVM לא לינארי באמצעות קרנלים:

מסוגים לינאריים מעל \mathbb{R} הם בסופו של דבר thresholds, וכאן תהיה שגיאה גדולה בדעתה. טכניקה שנייה להשתמש בה היא **מייפוי של הנקודות ל-*feature space* ממימד גבוה יותר**. נניח שכל נקודה x הופכת להיות (x^2, x) , ואז נקבל שהדעתה כן ניתן להפרדה. כמובן, על ידי המרה של ייצוגים של מופעים אנחנו מקיים להוריד את ה-*approximation error* של **מסוגים לינאריים**.

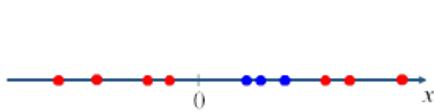


Figure 6.11: An interval of blue between two red areas

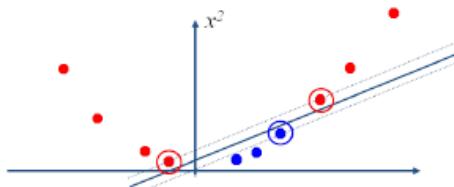


Figure 6.12: Lifting the data to 2D allows linear separation.

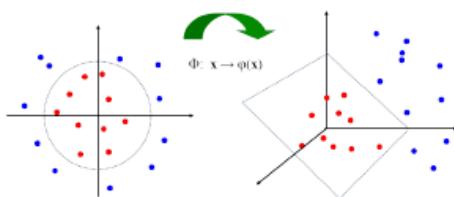


Figure 6.13: Mapping from 2D to 3D

נניח שיש לנו מייפוי $\Phi: \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ של מופעים לרפchnציות (feature mapping). נפעיל Hard SVM לרפchnציות. נקבל את הבעה הבאה, בה במקום x יהיה לנו את $\Phi(x_i)$:

$$\min_{w,b} \frac{1}{2} \|w\|^2 \text{ s.t. } \forall i. y_i(w \cdot \Phi(x_i) + b) \geq 1$$

נרצה למצוא את הבעה הדואלית:

$$\min_{\alpha} 0.5 \alpha^T \begin{bmatrix} - & y_i y_j \Phi(x_i) \Phi(x_j) & - \\ | & | & | \\ - & y^T \alpha & s.t. y^T \alpha = 0, \alpha \geq 0 \end{bmatrix} \alpha$$

הערות:

- נשים לב שחו"ז מאשר Φ אין שום דבר שיושב ב- $\mathbb{R}^{d'}$. מספר המשתנים זה כמו מספר הדוגמאות T .
- אם $(\Phi \cdot \Phi)(x) = 0$ ניתן ליחסוב ביעילות, לכל x , ניתן לפתור את הבעה לא משנה כמה גדול המימד של הרפchnציה.
- נניח ש- α^* הוא פתרון, אז נוכל לקבל פתרון לפורימלית כמו שהראנו על Hard SVM רגיל: רגיל: $w^* = \sum_{i=1}^n \alpha_i^* y_i \Phi(x_i)$.
 - אם (x_i, y_i) הוא לא SV אז $\alpha_i^* = 0$.
 - אם (x_s, y_s) הוא SV אז $\alpha_s^* > 0$.
 - אם (x_s, y_s) הוא SV ו- $\alpha_s^* < 0$ אז בגלל תנאי KKT מתקיים: $b^* = y_s - w^* \Phi(x_s)$.

נקבל שפונקציית הסיווג היא:

$$h(x) = sign(w^* \cdot \Phi(x) + b^*) = sign \left(\sum_i \alpha_i^* y_i \Phi(x_i) \cdot \Phi(x) + y_s - \sum_i \alpha_i^* y_i \Phi(x_i) \cdot \Phi(x_s) \right)$$

בالمור אנו יכולים לעשות גם אימון וגם inference (קלסיפיקציה) בלי לעבוד אף פעם ב-*feature space* אם נוכל לחשב ביעילות את המכפלות ולעבור עם הkernel. **בהתנות מייפוי $\Phi: \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$** : **נגדיר את kernel המתאים**:

$$K(x', x'') = \Phi(x') \cdot \Phi(x'')$$

דוגמאות:

1) נסתבל על מיפוי של ישר למישור: $\Phi(x) = (x, x^2)$. במקרה זה אם יש לנו $\langle (x', x'^2), (x'', x''^2) \rangle = K(x, x')$ נקבל: $x'x'' + x'^2x''^2 = K(x, x')$. זה לא יתן לנו הרבה, אבל יש מקרים בהם זה יהיה מינימום הרבה יותר גבוה.

2) נסתבל על הפונקציה: $K_2: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$: Kernel quadratic ($K_2(x, x') = (1 + x \cdot x')^2$). המינימום של ה- d שצורך כדי למשם את זה הוא בערך d^2 . נראה את זה עבור feature space

$$\begin{aligned} K_2(x, x') &= (1 + x \cdot x')^2 = (1 + x_1 x'_1 + x_2 x'_2)^2 \\ &= 1 + x_1^2(x'_1)^2 + x_2^2(x'_2)^2 + 2x_1 x_2 x'_1 x'_2 + 2x_1 x'_1 + 2x_2 x'_2 \\ &= \underbrace{\begin{pmatrix} 1 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ x_1^2 \\ x_2^2 \\ \sqrt{2}x_1 x_2 \end{pmatrix}}_{\phi(x)} \cdot \underbrace{\begin{pmatrix} 1 \\ \sqrt{2}x'_1 \\ \sqrt{2}x'_2 \\ x'_1^2 \\ x'_2^2 \\ \sqrt{2}x'_1 x'_2 \end{pmatrix}}_{\phi(x')} \end{aligned}$$

אפשר להראות באופן דומה עבור d כלשהו שקיים $\Phi \in \mathbb{R}^d \rightarrow \mathbb{R}^{O(d^2)}$ שemmמש את K_2 .

3) נסתבל על הפונקציה: Degree r polynomial kernel ($K_r: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$): הנ吐נה על ידי $(x' \cdot x)^r$. בآن $K_r(x, x') = (1 + x \cdot x')^r$. בآن $\Phi \in \mathbb{R}^d \rightarrow \mathbb{R}^{O(d^r)}$ שemmמש את K_r . בآن ב-feature space נבודעם מינימם גדול אם לא היינו משתמשים ב-kernel.

4) Radial basis function kernel (RBF): kernel space ($K_g(x, x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}}$) בآن $\Phi \in \mathbb{R}^d \rightarrow \mathbb{R}^\infty$ (סדרות שהסכום שלהם בריבוע מתכנס) שemmמש את K_g . יותר מכך, לא קיים מיפוי למינימם ∞ ה- d שemmמש את K_g . הוא איןסופי ולא ניתן למשם שום דבר בצורה ישירה, אבל על ידי kernel space אפשר.

אפיון של קרנליים:

מצומם: בהינתן $K: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ וקבוצה $C = \{x_1, \dots, x_m\} \subseteq \mathbb{R}^d$, נגדיר על ידי K_C את המצומם של K על C:

$$(K_C)_{ij} = K(x_i, x_j)$$

משפט (Mercer's theorem): K הוא קרナル, כלומר יכול להיות ממומש על ידי Φ בלבד, אם ומתקיימים התנאים הבאים:

1. K הוא סימטרי: $K(x, x') = K(x', x)$.
2. לכל קבוצה C מתקיים $0 \leq K_C \leq I$, כלומר המטריצה K_C היא PSD.

הכללה של קרנליים חדשים: אם $\mathbb{R} \rightarrow \mathbb{R}^d \times \mathbb{R}^d \rightarrow K_1, K_2$ הם קרנליים, אז גם הדברים הבאים הם קרנליים:

- חיבור: $K_3(x, x') = K_1(x, x') + K_2(x, x')$
- כפל: $K_3(x, x') = K_1(x, x') \cdot K_2(x, x')$

למה 1: תהא $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}$ מטריצת הדטא. אם יש ל- X דרגה מ- d אז הדטא ניתן להפרדה לינארית.

למה 2: אם מטריצת הקרナル $(K_S)_{ij} = K(x_i, x_j)$ בעלת דרגה מ- d , אז שגיאת האימון של Hard SVM תהיה 0.



Multi-Class Learning

מה קורה באשר $\infty < |y| < 2$? כל המושגים שדיברנו עליהם היו כוכבים עבור סיווג **בינארי**. בעת נדבר על המצב שבו יש יותר מ-2 תיוגים אפשריים. אפשרות פחות מקובלת היא לעשויות דזוקציה לסיווג ביןארי. האפשרות שבחר בה היא באופן ישיר לשלב את מרחב התיאוגים בבעיה. נctrיך להגדיר loss מותאים בשבייל זה.

נניח שיש לנו פונקציות score לכל מחלוקת, כאשר המודל שלנו מקבל את x בתור קלט ויש לו פרמטר w : $(w; x) \rightarrow r_1(x; w), \dots, r_{|y|}(x; w)$.
נעשה קלסיפיקציה על ידי הציג המקסימלי שיתקיים על ידי המודל: $(w; x) \rightarrow \hat{y} = \arg \max_y r_y(x; w)$. עבור ERM עם zero-one loss

$(\hat{y}_i, y_i) = \min_w e_S(w) = \frac{1}{n} \sum_{i=1}^n \Delta_{zo}(\hat{y}_i, y_i)$. ראיינו שבסיווג ביןארי למצער את zero-one loss ERM עם loss ℓ ($r_1(x; w), \dots, r_{|y|}(x; w)$) המקיימת את התנאים הבאים:

1. היא קמורה במשתני scores השונים: $(w; x) \rightarrow r_1(x; w), \dots, r_{|y|}(x; w)$.
2. הפונקציה חוסמת מלמעלה את Δ_{zo} של הפרדיקציה: $(y, y') \rightarrow \Delta_{zo}(\arg \max_{y'} r_{y'}(x; w))$.
3. ה-loss שואף ל-0 כאשר $(w; x) \rightarrow r_y$ של המחלוקת הנכונה יהיה גדול ביחס לאחרים $\{r'_y(x; w)\}_{y' \neq y}$.

Multi-class Hinge loss

$$\begin{aligned} \ell_{hinge}(r_1(x; w), \dots, r_L(x; w), y) &= \max_{y'} \{r_{y'}(x; w) - r_y(x; w) + \Delta_{zo}(y', y)\} \\ &= \max \left\{ 0, \max_{y' \neq y} \{r_{y'}(x; w) - r_y(x; w) + 1\} \right\} \end{aligned}$$

- אם Δ_{zo} של המחלוקת הנכונה הרבה יותר גדול מכל שאר scores, הפער בין שני scores השני היבן גדול הוא יותר מ-1, אך הביטוי בתוך המקסימום השני יהיה שלילי לפחות $y' \neq y$ (כל המחלוקות הללו נוכנות), עושים אחר כך מינימיזציה עם 0 ולבן לא משלמים כלום.
- אם זה לא המצב והפער קטן מ-1, נתחייב לשלם לפיה hinge-score גבוהה של מחלוקת כלשהי שהיא לא נכונה.

טענה: ℓ_{hinge} מקיים את 3 הדרישות שצינו.

Cross entropy loss

$$P[y|x; w] = \frac{e^{r_y(x; w)}}{\sum_{y'} e^{r_{y'}(x; w)}}$$

לפעולה זו קוראים בדרך כלל softmax. נשים לב שלעשות קלסיפיקציה על ידי score מקסימלי זה אותו דבר כמו על ידי הסתברות מקסימלית. hinge-loss, שנקרא גם log-loss, מבוסס על הרעיון של מינימום ההסתברות של הקלאס הנכון. בצורה שקולה, למצער את מינום הלוג של ההסתברות שחלוקת נבונה.

$$\begin{aligned} l_{log}(r_1(x; w), r_2(x; w), \dots, r_{|y|}(x; w), y) &= -\log P[y|x; w] = -\log \left(\frac{\exp(r_y(x; w))}{\sum_y \exp(r_{y'}(x; w))} \right) = \\ &= -r_y + \log \left(\sum_{y'} e^{r_{y'}} \right) \end{aligned}$$

דוגמאות לפונקציות score: למודל שלנו יש פרמטרים w , הוא מקבל הקלט את x ומוחזיר scores $r_1(x; w), \dots, r_{|y|}(x; w)$. אופן הבניה של הפונקציות האלו הנו הגדרת המודל.

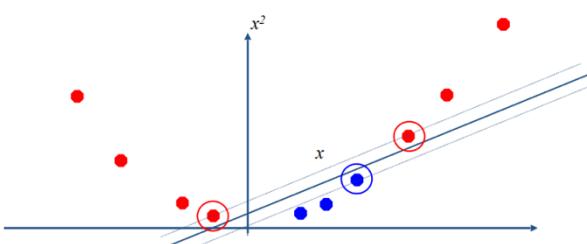
- הפליטים יכולים להיות מוצא של רשת נוירונים.
- פרדיקציה ליניארית: $r_y = b_y + \Phi(x) \cdot w$. כלומר יש r_y לכל מחלוקת y עם הפרמטרים שלה. במקרה זהה מתקיים $\Phi(w_1, b_1, w_2, b_2, \dots, w_{|y|}, b_{|y|}) = \Phi(w_1, b_1, w_2, b_2, \dots, w_{|y|-1}, b_{|y|-1})$. אם cross-loss קמור ביצויים, אז בשנכניס את הצינום שיוצאות ממודול לפונקציית loss הוא יהיה קמור ב- Φ . זה אומר שהשגיאת האימון גם תהיה קמורה.

הערות:

- ניתן להראות ש- ℓ_{log} הוא קמור ביצויים $(w; x) \rightarrow r_1(x; w), \dots, r_{|y|}(x; w)$. ניתן להוכיח על ידי שימוש בנגזרות שניתנות.
- אם $r_y \gg r_{y'} \gg 0$ עבור $y \neq y'$ מתקיים $0 \approx r_y + r_{y'} - \ell_{log} = r_y - r_{y'}$. ניתן להוכיח על ידי גבולות, $\infty \rightarrow r_{y'}$.
- עד כדי הכפלה בפקטור של e^{\log_2} , מתקיים כי ℓ_{log} חוסם מלמעלה את Δ_{zo} . אם הפרדיקציה לא נבונה, יוכל לדעת ש- ℓ_{log} חסם מלמעלה של ביטוי מסוים, ואם נכפיאו אותו בקבוע (NORMAL) הוא חסם מלמעלה את Δ_{zo} .

תרגול 7 (Kernel SVM)

:Non-Linear SVM



Slide credit: Andrew Moore

דיברנו על SVM בטור הסיטואציה המכונה אונחן חיצים לשוג דאטא בצורה לינארית. במקרה שהדאטא אינו להפרדה לינארית, אז SVM נותן לנו מפער בין m_{margin} ו- m_{cost} (נוטן למ הכללה). בפועל, הרבה פעמים הדאטא לא מסודר בצורה שנייה להפרדה לינארית. אפילו במקרה חד-מימדי עם דוגמאות חיוביות בשני הצדדים ושליליות באמצע, לא יוכל לשוג בצורה לינארית.

לכן, נעשה מניפולציה שבה אנחנו מופיעים את הדאטא למיד יותר גובה, עם התוכנה שבמימד החדש הוא יהיה ניתן להפרדה לינארית או לפחות קרוב לה. אם לוקחים את x ומושיפים את x^2 נוכל לשוג במימד 2. המפער במימד 2 מיתרגם למפער לא-לינארי במימד המקורי. בפועל, למדן מפער לא-לינארי. פונקציית המיפוי היא $\phi(x)$ שמעבירה למיד גובה (אפילו איןסוף).

למה שנוכל לעשות זאת באופן יעיל? אם מופיעים למיד ענק צריך להתמודד עם דאטא במיד מאד גבוהה. בעובדה עם Kernel מטיביות חישוב כפונקציה של המיד הגובה, ונחשב דברים במימד הנומוק. מה שמאפיין את הבעה של ה-SVM, הוא שהפתרון שלה תלוי לא ביצוג המפורש של ϕ אלא רק במכפלות פנימיות בין $\phi(x_i)$ ו- $\phi(x_j)$.

טריק הקרNEL: נגידו את פונקציית הקרNEL: $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ שמחשבת לנו את המכפלת הפנימית שמשמעותה אונחן. שפונקציות K מיוחדות שאפשר לחשב אותן בצורה שטוליה רק ב-d המוקורי, ומרות שבפועל המיפוי הוא למיד הרבה יותר גודל. זה שימושי כאשר הבעה שלנו תלויות ורק במכפלות הפנימיות האלה. דוגמאות לKERNELים שימושיים:

- קRNEL פולינומילי מדרגה d.
- הומוגני - $K(x, x') = (x \cdot x')$
- לא הומוגני - $K(x, x') = (1 + x \cdot x')$. בתוספת 1. כשפותחים את הסוגרים יהיו גורמים בדרגות נמוכות מ-k.
- קRNEL גאוסיאני/RBF: $K(x, x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}}$

:RBF

למה 1: תה $X = \begin{pmatrix} x_1^T \\ \dots \\ x_n^T \end{pmatrix}$ מטריצה הדאטא מסודרת בשורות, כל שורה ממימד p (אחרי המיפוי) ויש n שורות. אם יש ל-X דרגה n אז הדאטא ניתן להפרדה לינארית.

הוכחה: נזכיר כי דרגה של מטריצה היא כמה העמודות/השורות הבת'ל במטריצה. הדרגה לא יכולה לעבור את מספר השורות/העמודות. הכיו טוב שוכב לקוות זה שהדרגה היא n (כאן $d < n$ כי ניקח את d לאינסוף).

המימד של מרחב העמודות של X הוא n. בפרט, כל וקטור $\in \mathbb{R}^n$ אפשר לקבל בצל' של העמודות. נעשה את זה על ידי כפל המטריצה X בוקטור w (כפל מטריצה בעמודה הוא בצל' של עמודות המטריצה) ולכן קיים w כך ש- $x^T w = Xw$ כי העמודות פורשות את כל המרחב ובפרט את וקטור התוצאות. נסתכל על הסיגנ שזה מגדי: $y_i = sign(w \cdot x_i) = \pm 1$.

למה 2: אם מטריצת הKERNEL $(K_{ij}) = K(x_i, x_j)$ בעלת דרגה n, אז שגיאת האימון של Hard SVM תהיה 0 (במילים אחרות הדאטא אחרי הטרנספורמציה ניתן להפרדה לינארית).

הוכחה: תה $\phi_s = \begin{pmatrix} \phi(x_1)^T \\ \dots \\ \phi(x_n)^T \end{pmatrix}$. מתקיים $\phi_s^T \phi_s = K_s$, כך נקבל את המכפלות הפנימיות המתאימות. מכפלת של מטריצות יכולה רק להקטין את הדרגה (מבצעים עד צ'ל של העמודות) וכן:

$$n = rank(K_s) \leq rank(\phi_s) \leq \min(n, d) \leq n$$

בשני האגפים קיבלנו את n, מתקיים $n = rank(\phi_s)$ ולכן נסיק מלה 1 שהדאטא ניתן להפרדה.



טענה: עבור קרגל RBF, קיים מיפוי ϕ שסמן מרחב מממד אינסופי ומקיים $(\phi(x) \cdot \phi(x')) = K(x, x')$.

הוכחה: נניח שיש לנו דאטה חד-מימדי $\{(x_i, y_i)\}_{i=1}^n$. ניעזר בעובדה שאם a_1, \dots, a_n שונים זה מזה ו- b_1, \dots, b_n גם שונים זה מזה אז המטריצה שמודדרת $A_{i,j} = e^{a_i b_j}$ היא מדרגה מלאה. בפרט נדע כי $A_{i,j} = e^{a_i b_j}$ מדרגה מלאה. נבדוק כדי להביא אותה לצורה של ה-RBF. הפעולות שנעשה עליה לא ישנו את הדרגה.

לכל שורה i נכפיל אותה ב- $e^{-\frac{x_i^2}{2}}$. קיבל עמודה j נכפיל אותה ב- $e^{-\frac{x_j^2}{2}}$. נקבל את המטריצה הבאה שהיא מדרגה מלאה:

$$B_{i,j} = e^{-\frac{x_i^2 + x_j^2 - 2x_i x_j}{2}} = e^{-\frac{(x_i - x_j)^2}{2}} = K(x_i, x_j)$$

מלמה 2, שגיאת האימון של SVM Hard תהיה 0.

Kernel SVM עם SGD

נזכר בבעיית h-SVM. רأינו שהוא אפשר למצורע loss hinge עם גבוליזציה ℓ_2 – בעיה שהיא בכלל לא מאולצת!

$$\min_{w \in \mathbb{R}^d} C \sum_{i=1}^n \max\{0, 1 - y_i(w \cdot x_i)\} + \frac{1}{2} \|w\|^2$$

זו בעיית אופטימיזציה שאפשר לפתור עם SGD. נניח שעשינו כבר טרנספורמציה על ה- x -ים למינד גבוה d . באמצעות הקרגל, נישאר בסיבוכיות של d לכל אורך הריצה ולא d . רأינו דרך לפתור Soft SVM עם הפטרון של הדואלית, בעת נראה את ה-SGD:

$$\min_{w \in \mathbb{R}^d} C \sum_{i=1}^n \max\{0, 1 - y_i(w \cdot \phi(x_i))\} + \frac{1}{2} \|w\|^2$$

נתחל $w_0 = 0$. רأינו בשיעור על SGD במקרה של hinge-loss:

- אם $y_i(\phi(x_i) \cdot w_t) < 1$ אז העדכון יהיה $y_i(\phi(x_i) \cdot w_t) + \eta C$.
- אחרת העדכון יהיה $y_i(\phi(x_i) \cdot w_t) - \eta C$.

الطريق הבא להבין ש- w_t הוא איזשהו צ"ל של $\phi(x_i)$. בהתאם ($\phi(x_i)$ – צירוף טריוויאלי) בכל צעד עדכון אנחנו כופלים אותו בסקלר ($\eta - 1$) ומוסיפים לו איזשהו $y_i(\phi(x_i) \cdot w_t)$. אפשר לעקוב אחרי המקדים בצל זהה. ב모ות המקדים היא לכל היתר α (במונחים הדוגמאות שלמן). אז נראה שאפשר למשה בפועל את ה-SGD על ידי כך ששנחזיק את המקדים שמייצגים את w .

נסמן את המקדים של הצל w כ- $\alpha^{(t)}$, כלומר $\alpha_1^{(t)}, \dots, \alpha_n^{(t)}$. נעקוב אחרי המקדים וכיידם משתנים:

- אם $y_i(\phi(x_i) \cdot \alpha^{(t)}) < 1$ אז מתקיים $y_i(\phi(x_i) \cdot \alpha^{(t+1)}) = y_i(\phi(x_i) \cdot \alpha^{(t)} + \eta C)$.
- במקרה $y_i(\phi(x_i) \cdot \alpha^{(t)}) \geq 1$ אז מתקיים $y_i(\phi(x_i) \cdot \alpha^{(t+1)}) = y_i(\phi(x_i) \cdot \alpha^{(t)}) - \eta C$.

אם אפשר לחשב את הקרגל בזמן $O(dn)$ אז סיבוכיות כל איטרציה היא $O(dn)$.



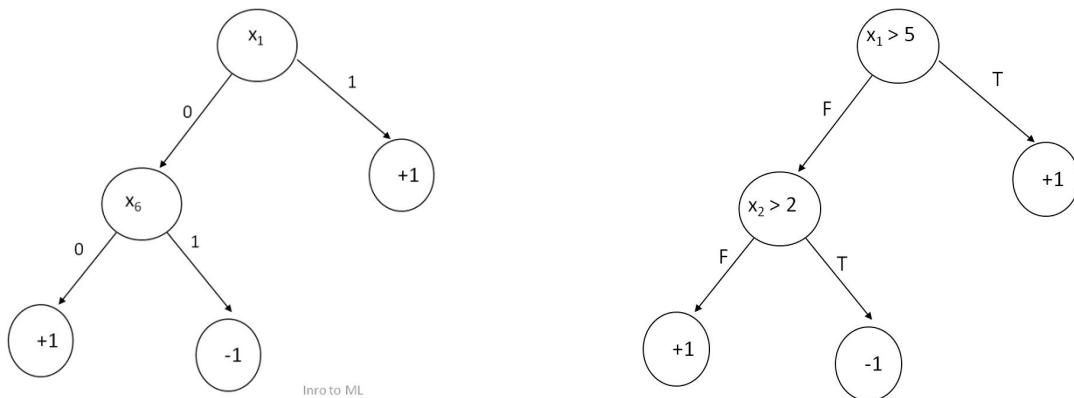
Decision Trees

עצים החלטה

בקיצור: עץ החלטה (decision tree) הוא מודל מיידא מפוקחת, שבבניות מסוימות הוא גם היום נותן תוצאות טובות, והיתרונו הענק בו על פני מודלים אחרים הוא שקל מאוד להבין את החלטות שמתבצעות בו – עובד בצורה יותר קרובה לאיך שאנו חושבים.

- תהילך הקלסיפיקציה הוא ע"י מסלול בעץ מהשורש עד לאחד העלים.
- העלה שבו נוחותים מגדרו מההתוווג שאנו חווים.
- ב策מתים הפנימיים יש **predicates** –策מת ההחלטה. אלו הם תנאים בוליאניים שתוצאתם אמת/שקר.
 - אמת (1) – נלך ימינה.
 - שקר (0) – נלך שמאלה.

במקרה שלנו, נניח $\{0,1\}^d = \mathcal{X}$ ו- $\{+1, -1\} = \mathcal{Y}$. נתעסק במקרים של **עצים ביןאריים מלאים** – כל策מת שאינה עלה הוא בעל שני בנים. הדטאא לא חייב להיות ביןארי, והוא יכול להיות גם רציף. במקרה זה מקבל ע"ז דומה, אבל התנאים צריכים להשתנות כי הוקטוריהם בברא לא ביןאריים.



התנהגות בלית:

- עץ ההחלטה מבוסס על קבוצת הפרדיקטיבים האפשריים H .
- הקולט לתהילך הלמידה של DT הוא סט אימון $\{(x_i, y_i)\}_{i=1}^n = S$, והפלט הוא עץ ההחלטה בו כל策מת פנימי מתאים לפדריקט $h \in H$, וכל עלה מתאים לתוווג.
- אנחנו רוצים שייהי לנו עץ ההחלטה שמתאים לדטאא S , אבל באותו זמן שיהיה קטן. אם אפשר לעצמנו לעבוד עם כל העצים, לפחות במקרה בוליאני ניתן למשם את כל הפונקציות, ואם נקבע את גודל העץ ניתן לשוט על גודל המחלקה.

VCdim של עץ החלטה: נחושב על המקרה בו $\{[d]\} \in i : \{x\} = H$ – כלומר הפרדיקטיבים מסתכלים על קוורדינטה. אם הגודל של העץ לא מוגבל, כל פונקציה $\mathcal{U} \rightarrow \mathcal{X}$ ניתן לנתח את קבוצת כל הקטליטים האפשריים, כי נוכל למשם כל תיוג אפשרי. לכן $VCdim = 2^d$ **VCdim** זה בעצם הגודל של \mathcal{X} .

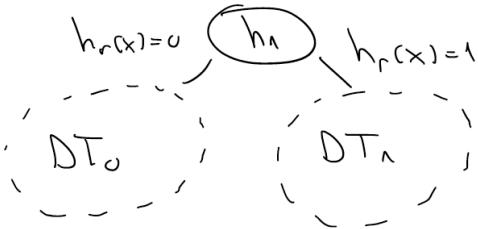
בעת נניח שאנו מגבילים את מספר העלים להיות $N \in S$. מחלוקת ההיפותזות שלנו, בוללת את כל העצים שיש להם לפחות S עלים. נסמן את המינימד בתור $VCdim(S, d)$.

טענה: $VCdim(S, d) \geq S$

רעיון הוכחה: אם יש לנו קבוצה של S נקודות, נוכל לבנות עץ עם S עלים שנכתב כל נקודה לעלה אחר. ניתן לעשות זאת בצורה קונסטרוקטיבית – בהינתן S נקודות נבחר תנאי שמאפשר לחלק לפחות אחת מהן ומחליק לפחות לקבוצות שונות, שלא שולח את כלן לאותו ביוון. אם אין לכך כל הנקודות אותו דבר. בכלל פעם נקבע את הקבוצה בלבד 1 עד שנפריד את כל הנקודות. אם ניתן לבנות עץ זהה, נוכל לנתח את הקבוצה זו ולבחור את התוווגים איך שנרצה.

בשביל חסם עליון, נשתמש בתוצאה ידועה: מספר העצים הביןאריים עם S עלים הוא $4^S \leq \binom{2^S}{S+1}$. מספר הפרדיקטיבים $-1 - S$ 策מתים הפנימיים (מספר策מתים הפנימיים קטן ב-1 ממספר העלים) שווה $-1 - 2^S \cdot d$. מספר המשומות של תיוגים לעלים הוא 2^S . סה"כ נקבל $(8d)^S \leq 2^S \cdot 2^{S-1} \cdot 2^S \cdot d^S = 8d^S \leq log_2(8d^S) = Slog_2(8d) = S$.

אפשר להשתמש ב- S (מספר העלים) כדי לשוטט בסיבוכיות המחלקה. המטריה שלנו היא לשותות fit-training set נתון, וכך לבודד עם מחלוקת היפותזות קטנה כדי שנוכל להכפיל. כאמור, אם נשלוט על מספר העלים, חסמי הבדיקה שראינו בתחילת הקורס יהיה תקפים. באופן כללי, מציאת DT עם מספר עלים קטן שווה מ-S ומידoor השגיאה האמפירית היא בעיה קשה. לכן נשתמש באלגוריתמים שימושערכיים את ה-ERM.

**בנייה העץ****תהליכי הבנייה:**

1. בהינתן סט אימון S , נבחר פרדייקט כלשהו r לשורש (לב האלגוריתם הוא איך לבחור אותו).
2. נקבע את S_0 להיות דוגמאות האימון שעבורן $h_r(x) = 0$, ואת S_1 להיות דוגמאות האימון שעבורן $h_r(x) = 1$.
3. באופן רקורסיבי, עז החלטה DT_0 נבנה על ידי S_0 , ו- DT_1 נבנה על ידי S_1 .
4. לחבר לשורש את DT_0 בתת-העץ השמאלי, ואת DT_1 בתת-העץ הימני.

כדי שייהי לנו מימוש של הסכימה זו, נצטרך להחליט איך לבחור את r – הפרדייקט לצאתו השורש. יוגדר בצורה רקורסיבית, ولكن נתמקד בבחירה הפרדייקט לשורש.

בחירה פרדייקטים:**פונקציית פוטנציאלי val :** נגדיר אותה בצורה הבאה:

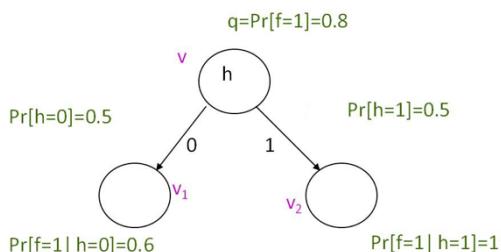
- עבור **עליה** τ שבו $\tau \in [0,1]$ אחד הדוגמאות החשובות שבנ התוצאות שוים ל-1+ נגיד: $val(q_v) = val(\tau)$.
- למשל נוכל להגיד $q_v = \min\{q_{v_1}, 1 - q_{v_2}\} = val(\tau)$. למשל אם 30% מהדוגמאות שmagiyot ל-v הן חשובות, עדיף לנו להגיד את התיאוג כ-1- (נניח צודקים 70% מהזמן) ונשלם מחיר של 0.30%. אם $q_v = 0.6 = val(q_v)$ נשלם מחיר של 0.4. זה אחד הטעות שנשלמים עבור התיאוג הכி טוב.
- עבור **צומת פנימי** τ עם פרדייקט h ושני **צאצאים עליים** w ו- s מקבל דוגמאות שעבורן $h(x) = w$ מקבל דוגמאות שעבורן $h(x) = s$, נגדיר: $val(q_w) = val(q_s) + (1 - p) \cdot val(q_w) + p \cdot val(q_s)$.
- ק' זה אחד הדוגמאות שmagiyot ל-w (לפי תוצאה הפרדייקט).
- q_w זה אחד הדוגמאות החשובות שבנ התוצאות שוים ל-1+ שmagiyot ל-w.
- q_s זה אחד הדוגמאות החשובות שבנ התוצאות שוים ל-1+ שmagiyot ל-w.
- עבור כל העץ נגדיר $val(q_v) = \sum_{v \in Leaves} p_v \cdot val(T_v)$.

פונקציית הפוטנציאלי תלולה גם בעץ וגם ב-set. עבור עליה זה תלוי רק באחיזה הדוגמאות החשובות (היחס) שmagiyot אליו. עבור צומת פנימי זה צירוף ממושקל של הפוטנציאלי של העלים, לפי אחיזה הדוגמאות החשובות שmagiyot לכל עליה. אפשר להמשיך ולהגדיר בכיה את הפוטנציאלי של צומת שהצאצאים שלו לא בהכרח שנייהם עליים.

בחירה פרדייקט לשורש: כאשר אנחנו בוחרים פרדייקט לשורש, אנחנו נחשב על כל האפשרויות שנוצרות מכל הפרדייקטים האפשריים. כמובן, אם שורש מקבל אוסף של דוגמאות, ויש X פרדייקטים אפשריים. אם נבחר את הראשון זה היה משורה החלקה כלשהי של training set לדוגמאות שהפרדייקט נותן להם 0 ולאלו שהוא נותן 1. נחשב את הפוטנציאלים של העלים האלו, זה משורה פוטנציאלי על השורש, ונרשום את זה בצד.

- נמשיך **לכל הפרדייקטים האפשריים**. נקבל X פוטנציאלים ל-X תסրיטים אפשריים ונבחר את זה שייתן לנו את **הפוטנציאלי הנמור ביוור**.
- אם אף פרדייקט לא הוריד פוטנציאלי ביחס לשמרות השורש בעלה, אז פשוט לשמור את השורש בעלה. זה מבטא את העובדה שנרצה את העץ הכி קטן שיסוג את training set ה-כוי טוב, אם זה לא מוריד פוטנציאלי אין צורך לשמור סתם עוד צמתים.

נשאר לנו להגיד את $val(q) = \text{עbor } [0,1] \in q$. אנחנו רוצים שהפוטנציאיל ישקוף את הטועות. כמובן כאשר ה- error prediction נמוך, והיפך. בחירה טבעית היא $q = \min\{q, 1 - q\} = val(q)$. במקרה זה, $val(q) = 0.5$. במקורה ה- $val(q) = 0.5$ שווה לטעות תחת התיאוג ה-כוי טוב. פוטנציאלי של שורש עם נשיעים הולך לייצג את classification error. תחת הבחירה הזו באופן חמדני, נקבל דוגמאות ונבחר פרדייקט כך שלצמת בעל 2 עליים יהיה classification error ה-כוי נמוך.



הבחירה הזו בעייתית:

- נניח שעבור השורש v , מתקיים $Pr[h=0] = q_v = 0.8$, כלומר 80% מהדוגמאות שmagiyot להן חשובות, לעומת 20% שmagiyot לא חשובות. נניח $Pr[h=1] = 1 - q_v = 0.2$ (מתייחס למעבר דוגמאות בפייל, כמו הולכות ימינה וכמה שמאליה, אחיזה הדוגמאות שmagiyot 1 = $h(x)$ לפ- הפרדייקט h).
- לפניו הפייטול: $Pr[v_1] = 0.2 = \min\{0.8, 0.2\} = val(q_v)$.
- אחרי הפייטול: $Pr[v_2] = 0.2 = \min\{1, 0.2\} = 0.5 \min\{1, 0\} + 0.5 \min\{0.6, 0.4\} = val(q_v)$. הפוטנציאיל מצביע על בר שאין התקדמות, אם הימנו בתהליכי שלבחירה פרדייקט לשורש, הימנו חושבים שהוא לא טוב כי הוא לא מוריד את הפוטנציאיל ביחס למצב שבו נשאיר רק את השורש.



מצד שני, **איןטואטיבית** לחת את h כנ עוזר, כי חצי מהדוגמאות מוגנות לעלה ויש לבצע אותו תיו. בולמר במכה אחת, **חצית הדוגמאות יכולות להיות מסווגות באופן מושלם**. לכן, דזוקא כנ היינו רוצים פרדיקט מהסוג זהה. אם בכל שלב הפרדיקט מסוגל לסוגו אותן נכון, נטפל בכך השני וכל פעם נוכל לנפות חצי מהן. לאחר מכן א' עצדים הדוגמאות שנשארו לנו הן 2^k . בקצב אקספוננציאלי מהר נוכל לטפל בכל ה-*training set*. לכן נרצה שפונקציית הפטונציאלית תיתן תגמול לפיצול זהה.

נדרוש שפונקציית הפטונציאלי **תקן תחת כל פיצול לא טריוויאלי**. בולמר אם $q_{v_1} \neq q_{v_2}$ שונים מ- q . בולמר, תמהיל הדוגמאות החשובות השתנה. אם הוא לא השתנה, נוכל ליגהע לה על ידי חלוקה רנדומית של הדוגמאות. אם כן, נרצה לשקר את זה. באופן q הוא תמיד צירוף קמור של q_{v_1}, q_{v_2} , לכן $q = p q_{v_1} + (1-p) q_{v_2} \in [0,1]$. הכ' גראוע שיבול להיות זה 0.5 בשנייהם. נרצה להיות ורchosים מ-0.5 לפחות 0 או 1. נרצה **שיטקיים** (פונקציה קעורה, לא קמורה):

$$\begin{aligned} val(q_v) &> p \cdot val(q_{v_1}) + (1-p) \cdot val(q_{v_2}) \\ val(q_v) &= val(p \cdot q_{v_1} + (1-p) q_{v_2}) \end{aligned}$$

נרצה שפונקציית הפטונציאלי תקיים את התנאים הבאים:

$$1. val(q) \geq 0$$

$$2. \text{סימטרית סביב } 0.5: \text{בולם } (1-q) = val(q) = val(q)$$

$$3. \{1, 0\} = q \Leftrightarrow val(q) = 0$$

$$4. \text{נורמליזציה: מתקיים } 0.5 = val(0.5), \text{ אינוריאנטיות להכפלה בקבוע חיובי.}$$

$$5. \text{כל פיצול לא טריוויאלי מפחית את הפטונציאלי.}$$

$$\forall p \in (0,1), q_{v_1} \neq q_{v_2} \in [0,1]:$$

$$val(p \cdot q_{v_1} + (1-p) q_{v_2}) > p \cdot val(q_{v_1}) + (1-p) \cdot val(q_{v_2})$$

נשים לב שהפונקציה שראינו עד כה מקיימת את 4 התנאים הראשונים אבל לא את התנאי ה-5. היא מקיימת את זה עם אי-שוויון חלש, לאו דזוקא עם אי-שוויון חזק. למשל אם ניקח את q_1, q_2 להיות נקודות על הפונקציה, זה מתלבך אותה והאי-שוויון מתקיים באופן חלש.

פונקציות פוטונציאליות: דוגמאות לפונקציות פוטונציאלי נפוצות:

$$1. val(q) = 2q(1-q) : \text{Gini index}$$

$$2. val(q) = \frac{1}{2} \left[q \cdot \log_2 \frac{1}{q} + (1-q) \log_2 \frac{1}{1-q} \right] : \text{Entropy}$$

$$3. val(q) = \sqrt{q(1-q)} : \text{Standard Deviation}$$

כל הפונקציות הללו מקיימות את 5 התנאים שלנו, פרט לפונקציה $\{q, 1-q\}$ mind כפי שראינו.

בעת נציג אלגוריתם שבונה עץ החלטה עבור סט אימון S , מחלוקת פרדיקטים H , ופונקציית פוטונציאלי val . הוא לא מגביל את גודל העץ.

עבור פוטונציאלי val הוא נקרא **ID3**.

בהתחלת נסובל על המקרה שבו כל התוצאות הם 0 או שכולם 1.

אחרת לכל $h \in H$:

- נחשב את S_h עבור הדוגמאות שהלכו ימינה, ואת הפרופורציה הזאת מתוך כלל הדוגמאות p_h .

- מכון נחשב את $S_{h,1}$, הדוגמאות שהלכו ימינה עם תיוג חיובי.

- בהתאם $S_{h,0}$, הדוגמאות שהלכו שמאליה עם תיוג חיובי.

נגידור את h , הפטונציאלי לכל פרדיקט על הפיצול שהוא ישרה.

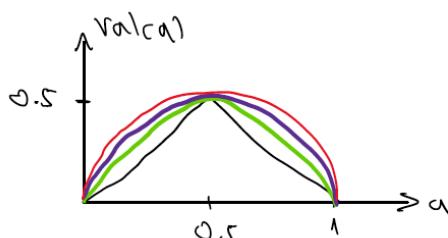
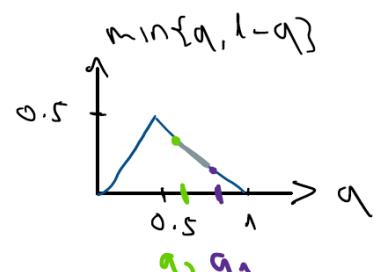
נבחר את הפרדיקט h' ש**שמגדיר את הפטונציאלי**. נגידור בהתאם את S_h .

עבור הדוגמאות שהלכו ימינה עם h' .

נחשב את תת-העצים רקורסיבית.

לבסוף, נחזיר DT כאשר השורש מתיוג h , תת-העץ הימני הוא T_1

והשמאלי הוא T_0 .



If $\forall (x, y) \in S$ we have $y = 0$ then

Create a leaf with label 0 and return.

Let $S_1 = \{(x, y) \in S | y = 1\}$

Let $q_1 = |S_1| / |S|$

Let $v = val(q_1)$

For each $h \in H$

Let $S_h = \{(x, y) \in S | h(x) = 1\}$.

Let $p_h = |S_h| / |S|$

Let $S_{h,1} = \{(x, y) \in S_h | y = 1\}$

Let $q_{h,1} = |S_{h,1}| / |S_h|$.

Let $S_{h,0} = \{(x, y) \in S - S_h | y = 1\}$

Let $q_{h,0} = |S_{h,0}| / |S - S_h|$.

Let $v_h = p_h \cdot val(q_{h,1}) + (1-p_h) \cdot val(q_{h,0})$,

Let $h' \in \arg \min_{h \in H} v_h$.

If $v_{h'} \geq v$ then

If $q_1 \geq 0.5$ then

Create a leaf with label 1 and return.

else

Create a leaf with label 0 and return.

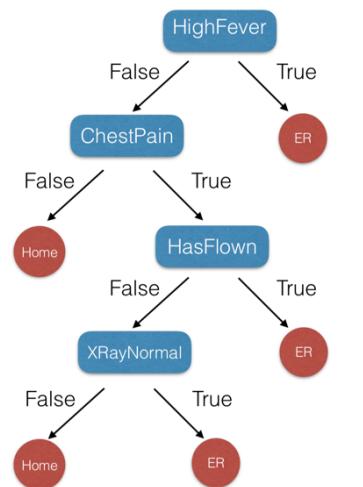
Let $S_{h'} = \{(x, y) \in S | h'(x) = 1\}$.

Call $DT(S_{h'}, H, val)$ and receive T_1 .

Call $DT(S - S_{h'}, H, val)$ and receive T_0 .

תרגול 8 (עצי החלטה וAINFORMCITY)

עצי החלטה:



המודול עצמו מודול פשוט: המוטיבציה – עובדים לא רעים בשיש מעט DATA (לא תמיד אפשרי לאוסף הרבה). תבונה יותר חשובה היא שההיפותזה שבסוף נפלטה היא זו שאנוחנו בטור בני אדם יכולים להסתכל עליה ולבהיר איך היא עובדת (Human interpretability). אנחנו רואים את התהליך המחשבתי.

- מזכיר בעץ עם צמתים (פרדיקט – שאלת, האם האינפורט מקיים תנאי כלשהו?), כאשר העלים הם התוצאות (labels).
- יכולים להיות סוגים שונים של פרדיקטים: תבונה בזאת $1 = X_i > 4$, סף מסויים X_i בזאת ושאלות מישוריות, רשות נירוחים וכו'. נרצה להסתכל על קוורדינטה (feature) בזאת ושאלות הוא גדול מערך כלשהו.
- הדרך שבה נעשה אבולוציה – מה ה-label המתאים עבור x כלשהו – עוקבים אחריו המסלול עד עלה כלשהו וזה התוצאה שנתקבל.

נתמך במקרה של $\mathcal{X} = \{0,1\}^d$ והפרדיקטים הם X_1, \dots, X_d .

- ראינו כי $2^d = VCdim(\mathcal{H}_{Full})$ וזה יתן לנו overfit עם גודל עם מחלוקת כל העצים האפשריים. לכן, נגביל את המחלוקת באמצעות שיר – עד S עליים בעץ: $VCdim(\mathcal{H}_S) \leq Slog8d$. חסמי ההכללה שראינו בקורס נכונים עברו ERM.
- הבעיה שמדובר בבעיה חישובית קשה לחשב ERM במחלוקת S . לכן, נשאף לעשות את הטוב ביותר שראינו יכולם לקבל פתרונות שיש להם הגון פנימי, ובמקרים פרקטיים מושגים לנו עצים לא רעים – היררכיות. מדובר באלגוריתמים חמדניים (כמו 3D) שאוינו לנו הבטחות לגביום. הם בעליים, אבל אין לנו אמירה של חסם על השגיאה.
- 3D הוא אלגוריתם חמדני, עושה רקורסיות ובכל שלב בודק על איזה פיצ'ר להסתכל כדי למצסם פונקציית Gain (מה שנותן הכח הרבה מידע). הוא חמדני כי הוא מסתכל לוקאלית על הפיצול שמקסם את ה-Gain, ולא מתקנים אחרות טעויות שנעשו. נסימן באשר כל הדוגמאות מתוצאות עם אותו label או שנשארנו בדיאגרמות.
- מובטח שאם נתונים לו לחוץ עד הסוף, תהיה לו שגיאה 0 על הדאטא – הוא נותן עץ שמתציג את הדאטא בצורה מושלמת. זה עץ שברוב המקרים יהיה גדול מדי. עצים ענקים לא נתונים לנו הכללה טובה, אולי יהיו לנו טעויות ונחיה עם זה.
- תתחילה לוח, ברגע שמייצרים יותר מידיעלים תעבור. עץ בגודל סביר, אולי יהיה לנו פוגעת משמעותית בביטויים של העץ.

בנייה העץ:

נתון לנו סט אימון. ניתן לתאר את האלגוריתם בקצרה באופן הבא, על ידי פונקציית gain:

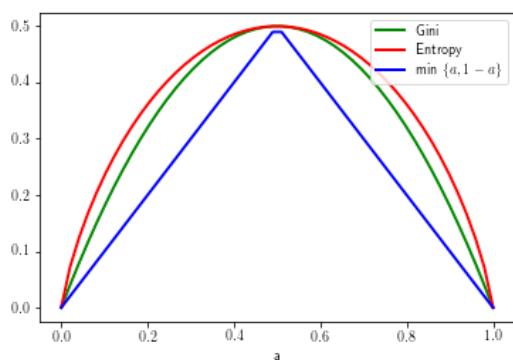
$$Gain(S, i) = [C[P_S[Y=1]] - P_S[X_i=1]C(P_S[Y=1|X_i=1] \underset{\text{before split}}{+} P_S[X_i=0]C(P_S[Y=1|X_i=0] \underset{\text{after split}}{)}$$

יש לנו שגיאה לפני הפיצול, ושגיאה אחריו הפיצול. אינטואיטיבית, הפרש אמור לנו כמה אינפורמציה יש לנו על ה-label, בהינתן שהסתכלתי על הפיצ'ר ה- i (X_i), בהשוואה לכמה מראש אינפורמציה הייתה לנו על ה-label בז' שהסתכלתי על הפיצ'ר הזה. C. הינה פונקציית פוטנציאלי, נדרש עליה שתיה קעורה ממש.

Gain is non-negative

פונקציה קעורה ממש: אפשר לחשב על כך בתור מינוס של פונקציה קמורה. נהפוך את אי השוויון של פונקציה קמורה. נדרוש strictly concave ממש – כלומר אם ניקח שתי נקודות שונות ונסתכל על המיתר שמחבר אותן, הערך על המיתר קטן ממש מהערך של הפונקציה. פורמלית עבור פונקציה $\mathbb{R}^d \rightarrow \mathbb{R}$: f אם עבור $y \neq x$ $f(y) < f(x)$ ג' מתקיים:

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



טענה: תהיו C פונקציה קעורה ממש המוגדרת על $0 \leq p \leq 1$ ב- $C(0) = C(1) = 0$ מינימלי – בין יש 100 אחוז דואות, ו- $C(0.5) = 0.5$ מקסימלי – זאת אי-ודאות מקסימלית. מתקיים $0 \geq i(S) \geq f$ לכל פיצול i (בלומר שאפשר להוסיף אינפורמציה בכל פיצול, או שזה לא ישנה אבל אין אינפורמציה שלילית).

מתקיים שיש שוויון אם: $P_S[X_i=1|X_i=1] = P_S[Y=1|X_i=1]$ או $P_S[X_i=0|X_i=0] = P_S[Y=0|X_i=0]$.

נחשב על הפונקציות העגולות, לא הכלולה. נראה שאם ניקח פונקציה כזו, ה- Gain הוא תמיד אי-שלילי, אף פעם לא נפסיד ממנו אם נפצל לפוי פיצ'ר כלשהו. לא נוכל אף פעם לקבל אינפורמציה שלילית אם נסתכל על עד מושהו מהדאטא.



- התנאי הראשון אומר שהזאות של X_i אם הוא 1 או 0 לא משפיע בכלל על התפלגות מעלה התיאוג Z , לא נתנת לנו אינפורמציה על Z .
- התנאי השני אומר שלכל פיצ'ר הסיבוי שהוא 0 הוא או 0 או 1. כל הערכות של הפיצ'ר הם אותו דבר. אם נחליט לפצל לפי פיצ'ר כזה, צד אחד יהיה ריק – לא באמת נפצל זה סוג של פיצול מכוון. הגיוני לומר שה-Gain בה הוא 0.

הוכחה:

$$\begin{aligned} \text{נחשב ראשית את השגיאה לפני הפיצול (בנעד בנסיבות אשר } [0] = P_S[X_i = 0] \\ C(P_S[Y = 1]) &= C(P_S[Y = 1|X_i = 1]P_S[X_i = 1] + P_S[Y = 1|X_i = 0]P_S[X_i = 0]) \\ &\geq_{\substack{\text{הסתברות שלמה} \\ \text{קשורה}}} P_S[X_i = 1]C(P_S[Y = 1|X_i = 1]) + P_S[X_i = 0]C(P_S[Y = 1|X_i = 0]) \end{aligned}$$

קיבלנו את הביטוי עבור השגיאה לאחר הפיצול ולבן קיילמו 0 $\geq G(S, i) = pre_{error} - post_{error}$ שכן $P_S[X_i = 0] \in \{0,1\}$ או $P_S[Y = 1|X_i = 0] = P_S[Y = 1|X_i = 1] \Leftrightarrow \lambda \in \{0,1\}$, או $y = x$.

Mutual information and KL divergence

נרצה לבחור C ספציפית שתיה קשורה לאינפורמציה. נבחר פונקציית אנטרופיה (Entropy).

$$H(X) = - \sum_x P[X = x] \log P[X = x]$$

הfonקצייה זו מכמתת לנו **במה אי-וודאות (חוסר אינפורמציה) יש לנו במשתנה מקרי**. התפלגות שמקסימה אנטרופיה למשל היא התפלגות אחידה, אין לנו שום מידע מוקדם על מה יצא הערך, זה יכול לצאת כל ערך בהסתברות שווה. אם ניקח התפלגות נורמלית, יש אנטרופיה יותר נמוכה, כי ערכייהם קרובים לערך אחד וודאות שנתקבלו אותן מאשר ערכים שקרובים למרცב.

נגדיר גם אנטרופיה מותנית, במה אי-וודאות יש לנו ב- Y אחרי שאמרו לנו את X :

$$H(Y|X) = - \sum_{x,y} P[X = x, Y = y] \log P[Y = y|X = x]$$

post-split error =

$$\begin{aligned} &= -P_S[X_i = 1]\mathbb{P}_S[Y = 1 | X_i = 1] \log \mathbb{P}_S[Y = 1 | X_i = 1] \\ &- P_S[X_i = 1]\mathbb{P}_S[Y = 0 | X_i = 1] \log \mathbb{P}_S[Y = 0 | X_i = 1] \\ &- P_S[X_i = 0]\mathbb{P}_S[Y = 1 | X_i = 0] \log \mathbb{P}_S[Y = 1 | X_i = 0] \\ &- P_S[X_i = 0]\mathbb{P}_S[Y = 0 | X_i = 0] \log \mathbb{P}_S[Y = 0 | X_i = 0] \\ &= H(Y | X_i) \end{aligned}$$

בetta באמצעות האנטרופיה את $H(Y|X_i)$ שראינו קודם, כלומר נבחר **את C להיות פונקציית האנטרופיה ונקבל:**

$$C(P_S[Y = 1]) = -P_S[Y = 1] \log P_S[Y = 1] - P_S[Y = 0] \log P_S[Y = 0] = H(Y)$$

עבור השגיאה לאחר הפיצול נקבל $H(Y|X_i)$.
סה"כ ($G(S, i) = H(Y) - H(Y|X_i) = I(Y; X_i)$) **במה אי-וודאות היה לנו על Y מראש, פחות במה יש אי-וודאות יש לנו אחריו שראינו את X . קוראים להפרש זהה האינפורמציה המשותפת.**

טענה: הקשר בין אינפורמציה משותפת D_{KL} הוא:

$$I(Y; X) = H(Y) - H(Y|X) = \sum_{x,y} P(x, y) \left(\log \frac{P(x, y)}{P(y)P(x)} \right) = D_{KL}(P_{X,Y} || P_{X \otimes Y})$$

כלומר אינפורמציה משותפת היא סימטרית ואי-שלילית: $0 \leq I(X; Y) = I(Y; X) \leq 0$ אם X, Y הם ב"ת.

Instability of ID3

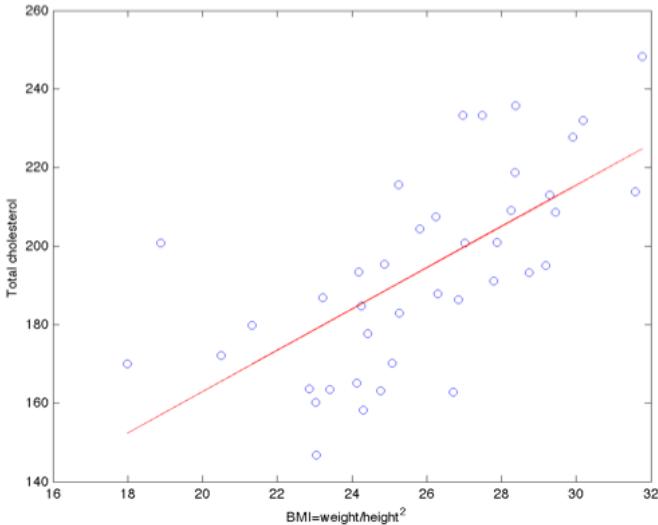
תמונה בעייתית ב-ID3 היא חוסר יציבות. יכולם להיות מקרים שאנו מתחילה סט ולומדים עליו עז. מישחו בא ומיד נקודה בצורה מינימלית, ונלמד עליו עז מחדש. יכול לקרות שניים העצים שייתקבלו יהיו שונים בצורה קיצונית אחד מהשני. הרבה פעמים בלמידה, יציבות של אלגוריתם היא מזדד ליכולת ההכללה שלו.

כנich שיש לנו דאטא של שני פיצ'רים x_1, x_2 והעץ מפצל לפי סף (threshold). אם נזיז נקודת אחת טיפה, הפיצול שעשינו בשורש משתנה. פיצלנו במקור לפי x_1 ובעשיינו נרצה לפצל לפי x_2 .

Regression

רגression

עד עבשינו התרכזנו בקלסיפיקציה: \hat{y} היא קבוצה ללא מבנה מטרי, כלומר אין משמעות למרחק בין תוצאות – או שהטיפול נIRON או שמדובר לא נIRON. אם הוא טועה, זה לא משנה מה הטעות, מושלמים אותו הדבר. ברגression זה לא המצב. **תוצאות יש מבנה מטרי**, ובדרך כלל **המ-ב- \mathbb{R}^k** .



לשם הפשטות, נחשב על המקורה שבו $\mathbb{R} = \mathbb{Y}$. המטרה שלנו היא לא להיות צודקים בתוצאות – לא למצוא את התוצאות הנכונות. התוצאות הAO רציף, ולכן לא ניתן ליצג בצורה נומירית כל אחד מהמספרים שלו (למשל כטוטם π). המטרה שלנו היא לחוץ תוצאותיהם שהם קרובים ל- \mathbb{Y} . **h-loss** \rightarrow **ground truth**. יש לרגression המון שימושים, למשל:

1. חיזוי רמת כולסטרול בדם על סמך מאפיינים פיזיולוגיים כמו BMI.
2. לחוץות בזמן תיכון פעולה מסוימת, למשל בקונטקט של חישובי: חיזוי בזמן t_0 ייחוך בשרת מסויים, על מנת להיעזר ולהקצות משאים בהתאם.

רגression לינארית

נניח כי $\mathbb{R} = \mathbb{Y} = \{(x_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \mathbb{R}$ – $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}^d$. ניקח \mathcal{H} מהצורה הבאה:

$$\mathcal{H} = \{x \rightarrow \langle a, x \rangle : a \in \mathbb{R}^d\}$$

- נשים לב שעלי ידי הוספת קוואורדינטה dummy שווה ל-1 לכל המופעים, נקבל bias בקוואורדינטה الأخيرة.
- בדרך כלל מדברים על regression לינארית עם loss ℓ_2 .

נניח שאנו לומדים על ידי מדוור loss square ℓ_2 : $\hat{a} = \arg \min_a \sum_{i=1}^n (y_i - a \cdot x)^2$ כאשר $\langle x, a \rangle = a \cdot x$.

- האילוץ \hat{a} הוא קמור, מכיוון שהוא סכום של ריבועים והרכבה של פונקציה לינארית.
- $x \cdot a - y_i =: r_i$ נקראים residuals של regression (הטעות שלנו על הדוגמה i -ה).

פתרונות בעיית regression:

נסמן $\mathbb{R} \in \mathbb{R}^{n,d}$ וקטור התוצאות, $\mathcal{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^{n,d}$ מטריצה שמחזיקה את הדוגמאות בטור שורות. נכתוב גם בתור עמודות.

ונכל לבתוב את \hat{a} כך X זה מטריצה a זה וקטור עמודה, נכפיל ונקבל וקטור. נחסר את זה מ- \mathcal{Y} ונקבל וקטור. נעשה נורמה בריבוע זה סכום הריבועים של הקואורדינטות. הוקטור הזה מייצג את residuals.

$$\hat{a} = \min_a \|Y - Xa\|^2 = (Y - Xa)(Y - Xa)^T = \dots = [\|Y\|^2 - 2Y^T Xa + a^T X^T Xa]_{\ell(a)}$$

טענה: a הוא מינימום גlobלי של $\ell(a)$ אם $\nabla \ell(a) = 0$

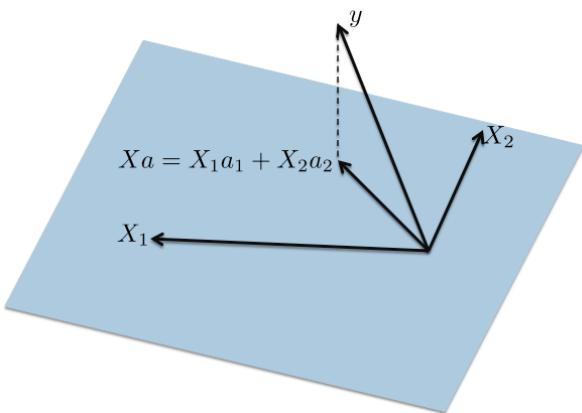
נחשב את הגרדיינט ונקבל: $\nabla \ell(a) = 2X^T Xa - 2X^T Y$. נשים לב כי אם נעביר אגפים נקבל $Y^T Xa = X^T Xa$. אם $X^T X \in \mathbb{R}^{d,d}$ היא מטריצה לא סינגולרית (כלומר מטריצה שהיא הפיכה), אז נקבל ביטוי סגור: $a^* = (X^T X)^{-1} X^T Y$

פרשנות גאומטרית של הפתרון:באשר $Xa - Y := r$ מתקיים:

$$\begin{aligned} X^T X a^* = X^T Y &\Leftrightarrow X^T (Y - Xa) = 0 \Leftrightarrow X^T r = 0 \\ &\Leftrightarrow r \perp v_1, \dots, v_d \in \mathbb{R}^n \end{aligned}$$

כלומר הפתרון לרגסיה לינארית, הוא זה שהשברות residuals ניצבים לכל הפיצרים. v_i הוא הפיצ'ר ה- i -י של כל a . נוכל לחוש על זה בצורה הגאומטרית הבאה:

מזה מספר הדוגמאות. העמודות v_1, \dots, v_d פורשות תחת-מרחבי. הפתרון הוא כזה שהshareית $Xa - Y$ ניצבת לכל הוקטורים. זה מה שקרה בשטחים נקודה על תחת-מרחבי. הפרדייקציות Xa הן הקירוב הטוב ביותר של Y בתת המרחב שנפרש על ידי $\text{span}\{v_1, \dots, v_d\}$.

סינגולריות:

מטריצה רנדומית שהכニסות שלה מוגעות מהתפלגות איחידה רציפה, תהיה הפיכה בסביבות גבוהה. לעומת זאת, במקרה של חוסר הפיכות הוא מקרה מנון – אותו נבדוק בעת התנאי לאופטימליות שכבר ראיינו הוא:

$$X^T X a = X^T Y$$

טענה: אם $X^T X$ סינגולרית, אז $Y = X^T X a$ בעלת אינסוף פתרונות.

מקרה מנון שمدגים את זה – נניח כי $[a_1, \dots, a_d] = X_i$, כל הדוגמאות יושבות ב- \mathbb{R}^d אבל מכילות 0 בכל הקואורדינטות מלבד הראשונה. בפועל $\mathbb{R}^d = X$ ולא \mathbb{R}^n . לעומת זאת, העמודות הראשונה a_1 נתונה על ידי האופטימום עבור $\mathbb{R}^n = X$ והמקדמים a_2, \dots, a_d חסרי משמעות – יכולות להיות כל דבר.

נניח ש- $X^T X$ סינגולרית. נגיד ש- a^* הוא פתרון של $Y = X^T X a$. מערכת המשוואות שקיבלו היא לינארית, והמטריצה של המקדמים היא סינגולרית. אנחנו יודעים שיכול להיות שלא יהיה אף פתרון, או שהוא אינסוף פתרונות. איך יודעים שיש בכלל פתרון? ניתן להראות על ידי SVD. נניח שיש פתרון $-(X^T X)^{-1} u \neq 0$ בلومר הגען של המטריצה לא טרוויאלי, $0 = u^T X a$. אז:

$$X^T X(a^* + u) = [X^T X a^*] + [X^T X u] = X^T Y \Rightarrow a^* + u \text{ is optimal}$$

טענה: $X^T X$ סינגולרית בשדרגה של $d < n$. לעומת זאת, שווה ש- X לא מלהיר ($\text{rank}(X^T X) = \text{rank}(X)$). אפשרות להראות ע"י SVD: $\text{rank}(X^T X) = \text{rank}(X)$.

מתי הדרגה לא מלאה?

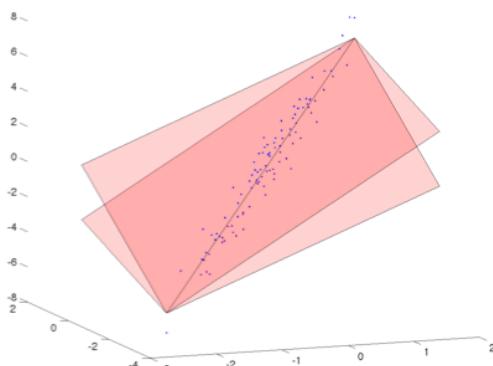
1. באשר $d < n$: מספר הדוגמאות קטן ממש ממספר הקלט.

2. כאשר העמודות של X תל"י: כי יש ל- X d עמודות והדרגה של X הקבועה הכיו גדולת על העמודות תל"י, ואם $d < n$ יש עמודות שען כנ"ל. למשל ניתן לחשב על מקרה בו v_i (העמודה ה- i -ה) שווה ממש לעמודה v_j . אז אם a^* הוא אופטימום, גם מה שנקבל מהיחס $\beta = -a^* + \lambda v_i$ ולהויר $\beta = -a^* + \lambda v_j$ לכל $\lambda \in \mathbb{R}$.

הערה: בפועל, אם $d > n$ אז X כמעט תמיד אפס פעם לא הפיך. באופן כללי, חסור הפיכות לא קורית באמת אף פעם, וזה לא מקרה מעניין. אבל כמעט תמיד תופיע שוקריות במקורה הפיך, המטריצה יכולה להיות כמעט סינגולרית ויהי את אותן בעיות במעט באזות מידה. לעומת זאת, במקורה לא תהיה סינגולריות אלא **במעט סינגולריות**. היחס בין הערך הסינגולרי (ערך עצמי במקורה ש- X ריבועית) בכ"ג גדול והערך הסינגולרי בכ"ג קטן, הוא יחס גדול.

Ridge Regression

נסתכל על מצב שבו הדאטא הוא \mathbb{R}^2 ואנו רוצים לבנות רגסיה לינארית. היא תהיה מישור, כי הוא פונקציה $\mathbb{R} \rightarrow \mathbb{R}^2$. הנקודות יושבות על מישור, ויש הרבה מישורים שנוכל להתאים להן. אפשר לסובב את המישור ועדיין לעשות fit על הנקודות והן לא ישבו בצורה מושלמת על הישר, כי כל רוש יוכל להציג אותן. למרות שבפועל יש מישור שיתאים בצורה בכ"ג טובה, יש מישורים שכמעט מתאימים לדאטא, וזה דבר לא נכון.



ניתן לעשות במקורה זה רגוליזציה ולהכניס bias. כאשר $X^T X$ סינגולרי, הרבה פתרונות מתאימים לדאטא. הראן מצב דומה ב-SVM. הדאטא היה ניתן להפרדה לינארית והוא הרבה פתרונות שהפרידו את הנקודות ואז הימ צריכים לבחור ולהכניס تعدוף, ובחרנו את המפריד עם המargin ה- m בכ"ג גדול שווה שקול למפריד עם הנורמה בכ"ג קטן.



גישה דומה ברגression מובילה ל-Ridge regression שבו נחפש \hat{a} באופן הבא:

$$\hat{a} = \arg \min_a (Y - Xa)^T (Y - Xa) + \lambda \|a\|^2$$

באשר λ קבוע רגולרייזציה אי-שלילי. הפתרון להה ייתן לנו:

$$\hat{a} = (X^T X + \lambda I)^{-1} X^T Y$$

המטריצה שעצבי צריך להפוך $I + X^T X$ תמיד תהיה PD (positive definite), שזה משמש תמיד איננו סינגולרי. בכלל ש- λ חזק יותר, האלמנטים על האלבוסון דומיננטיים יותר. אם λ שואף לאינסוף אז \hat{a} יהיה מטריצת יחידה שמוכפלת בסקלר קטן, ולכן נקבל נורמה נמוכה, שזה יגע ב-fit לדאטא.

הרחבות של רגסיה:

רגסיה לא לינארית: הסתכלנו על מקרה שבו $f(x) = \hat{y}$ כאשר f לינארית. אפשר גם לחשב על f לא לינארית. למשל f שמנומנת על ידי רשות נוירונים. יוכל גם למשל לחתך $(x) \phi \cdot a = (x) f$, כאשר ϕ היא מיפוי של פיצרים. ניתן להראות שבמקרה זה אם $(x) \phi \cdot a$ ניתן לחישוב ביעילות, אפשר להשתמש ב-kernel trick בילי לעבד במרחב הפיצרים.

מקרה של sparse regression: כאן אנחנו מוחפשים פתרון שיש לו קצת קוודינטות שונות מ-0. ניתן לפתח חסמי הכללה שתלויים במספר הקואורדינטות שונות מ-0 (סוג של רגולרייזציה). בנוסף, ניתן להשתמש זהה על מנת להוריד חישוב. לעיתים הרגסיה מבוססת לא על מידע גולמי, אלא על איזה מהם פיצרים שלפעמים גם מהנדסים אותם. בעיקר ב-maindom שאין בהם הרבה נתונים. לעיתים זה בודד לחשב את הפיצרים, ובכל מקום שיש 0 אין צורך לחשב את הפיצ'ר הזה ולפניהם זה יכול לחסוך זמן ריצה. זה סוג של feature selection.

תרגול 9 (רגסיה)

ברגסיה, בשונה ממלטי-פוקצייה, יש לנו דוגמאות מתINGTON אבל התוצאות יכולים להציג מעולם רציף (ממשיים למשל), אנחנו רוצים ללמידה פרדיktור שיבנה תיוגים באופן רציף. loss zero-one לא עובד כאן, אך לנו מסתכנים לרוב על loss square. נסתכל בעת על מקרה של רגסיה לא לינארית. איך ניתן לתקן באמצעות קרנל? יש לנו מיפוי של הדטאumar מימד d למימד d' ופונקציית קרナル שמאפשרת לנו לחשב מכפלות פנימיות ביעילות.

שאלה מבחון (תשפ"א ב' – מועד א):

Let $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ and let K be a kernel function that corresponds to ϕ . We want to solve the following optimization problem:

$$\min_{w \in \mathbb{R}^{d'}} \frac{1}{2} \sum_{i=1}^n (\mathbf{w} \cdot \phi(\mathbf{x}_i) - y_i)^2 + \frac{1}{2} \lambda \|\mathbf{w}\|^2,$$

where $\lambda > 0$.

ה-loss הבסיסי הוא ההפרש בין הפרסום לתיאוג האמתי, בריבוע. בשביל שהכול יהיה נחמד נוסיף גם רגולרייזציה ℓ_2 . מובטח שלבעה יש פתרון יחיד ויש נוסחה סגורה לחישוב של הפתרון. במקרה זה לא נרצה לעשות חישובים שתלויים ב- d' (להחזיק את w). לכן, נציג את w בתווך צ"ל של הנקודות האחרות – כמו ב-SVM (שם עברנו לדואלית). כאן נוכיח את זה בצורה ישירה.

a. Show that there exist scalars α_i such that

$$w = \sum_{i=1}^n \alpha_i \phi(\mathbf{x}_i).$$

באופן כללי, ניתן לפרק את w לריבוב ניצב ורכיב לא ניצב למרחב שנפרש ע"י $\{\phi(\mathbf{x}_i)\}$: $w = w^\perp + w^\parallel$. נציב:

$$\begin{aligned} & \frac{1}{2} \sum_{i=1}^n ((w^\perp + w^\parallel) \cdot \phi(\mathbf{x}_i) - y_i)^2 + \frac{1}{2} \lambda \|w^\perp + w^\parallel\|^2 \\ & \stackrel{w^\perp \cdot w^\parallel = 0}{=} \frac{1}{2} \sum_{i=1}^n (w^\parallel \cdot \phi(\mathbf{x}_i) - y_i)^2 + \frac{1}{2} \lambda \|w^\parallel\|^2 + \frac{1}{2} \lambda \|w^\perp\|^2 \\ & > \frac{1}{2} \sum_{i=1}^n (w^\parallel \cdot \phi(\mathbf{x}_i) - y_i)^2 + \frac{1}{2} \lambda \|w^\parallel\|^2 \end{aligned}$$

כלומר אם $0 \neq w^\perp$, התחלנו מ- w אופטימלי ומצאנו פתרון יותר טוב ממנו. אך $w^\perp = 0$ ונקבל $\{w^\parallel\} = \text{span}\{\phi(\mathbf{x}_i)\}$.



- b. Substitute $w = \sum_i \alpha_i \phi(x_i)$ in the optimization problem to show that finding w is equivalent to solving

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2} \|y - K\alpha\|^2 + \frac{1}{2} \lambda \alpha^T K \alpha,$$

where K is the kernel matrix.

נתרגם את בעיית האופטימיזציה של הרגרסיה, לבעה שתוליה בוקטור $\mathbb{R}^n \in \alpha$. את הבעה זו נדע לפתור ביעילות כי היא בעיה במינימום, קמורה ב- α . נכתוב את הבעה שלנו $\min_w \frac{1}{2} \sum_{i=1}^n (w \cdot \phi(x_i) - y_i)^2 + \frac{1}{2} \lambda \|w\|^2$ בכתב מטריציוני: נסתכל על מטריצה

$$\Phi = \begin{pmatrix} \phi(x_1) & \cdots & \phi(x_n) \\ \vdots & \ddots & \vdots \\ \phi(x_n) & \cdots & \phi(x_1) \end{pmatrix}_{n \times d},$$

שיש בה את $(x_i) \phi$ בשורות: $\Phi w - y = (\phi^T)_{d' \times n} (\alpha)_{n \times 1}$. יש לנו סכום ריבועים, אז נתרגם את זה להיות נורמה בריבוע: $\|\Phi w - y\|^2$.

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2} \|y - \Phi w\|^2 + \frac{1}{2} \lambda \|w\|^2 = \min_{\alpha \in \mathbb{R}^n} \frac{1}{2} \|y - \Phi \Phi^T \alpha\|^2 + \frac{1}{2} \lambda \|\Phi^T \alpha\|^2 = \min_{\alpha \in \mathbb{R}^n} \frac{1}{2} \|y - K \alpha\|^2 + \frac{1}{2} \lambda \alpha^T K \alpha$$

- נשים לב כי מתקיים $K = \Phi \Phi^T$ כי $K_{ij} = \phi(x_i) \cdot \phi(x_j)$
- $K^T = K \Rightarrow K^T K = K^2$

$$\|\Phi^T \alpha\|^2 = (\Phi^T \alpha)^T \Phi^T \alpha = \alpha^T \Phi \Phi^T \alpha = \alpha^T K \alpha$$

- c. Show that one can find the coefficients α with complexity that does not depend on d' (assume that you can compute $K(x, x')$ in $O(1)$).

נניח שהפונקציה K ניתנת לחישוב בזמן יעיל (O). איך נמצער את הפונקציה? בשפות חיסום מסוימים נקבל גורמים לינאריים ועוד גורם $\alpha^T K K \alpha$, נזכיר כי K מטריצה היא PSD ולכן הפונקציה קמורה (הע"ג אי-שליליות). נגזר ונשווה לאפס.

$$\begin{aligned} f(\alpha) &= \frac{1}{2} \|y - K \alpha\|^2 + \frac{1}{2} \lambda \alpha^T K \alpha = \frac{1}{2} (y - K \alpha)^T (y - K \alpha) + \frac{1}{2} \lambda \alpha^T K = \\ &= \frac{1}{2} (y^T y - y^T K \alpha - \alpha^T K^T y - \alpha^T K^T K \alpha) + \frac{1}{2} \lambda \alpha^T K \alpha = \frac{1}{2} (y^T y - 2y^T K \alpha - \alpha^T K^2 \alpha) + \frac{1}{2} \lambda \alpha^T K \alpha = \\ f'(\alpha) &= \frac{1}{2} (2K^2 \alpha - 2Ky) + \frac{1}{2} (2\lambda K \alpha) = K^2 \alpha - Ky + \lambda K \alpha = K^2 \alpha + \lambda K \alpha - Ky \\ &= K(K + \lambda I) \alpha - Ky = 0 \Rightarrow \alpha(K + \lambda I)^{-1} y \end{aligned}$$

נשים לב כי $I + K$ הפינה: מספיק להראות שהע"ג חיוביים ממש (אם יש ע"ג 0 זה הופך את המטריצה ללא הפינה). אנחנו יודעים שהם אי-שליליים כי K היא PSD. נכתוב $K = UDU^T = UDU^T$ (פרק ספקטורי). נכתוב גם $U^T = U$. אם נחבר נקבל את D בתוספת ע"ג חיוביים ממש על האלכסון, ונקבל הע"ג של הכל יוצאים חיוביים ממש.

וככל לסוג נקודת חדשה על ידי:

$$w \cdot \phi(x) = \sum_{i=1}^n \alpha_i \cdot \phi(x_i) \cdot \phi(x) = \sum_{i=1}^n \alpha_i \cdot K(x_i, x)$$

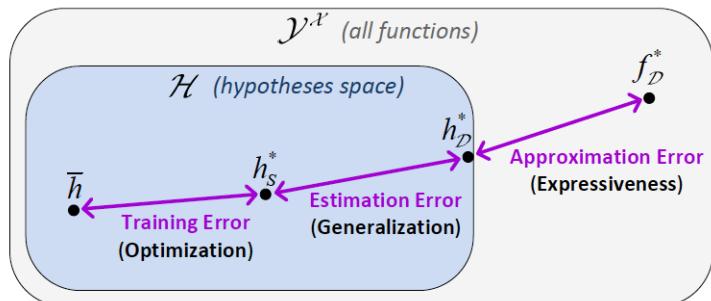
3 – למידה לא מפוקחת (Unsupervised)

שיטת נוספת

Deep Learning

למידה סטטיסטית (הבסיס של מדנו):

אנחנו עוסדים עם: \mathcal{X} – מרחב הדוגמאות, \mathcal{Y} – מרחב התוצאות, D – התפלגות מעל $\mathcal{Y} \times \mathcal{X}$ שלא ידועה לנו. $\mathbb{R}_{\geq 0} \rightarrow \mathcal{Y} \times \mathcal{X} : \ell$:
 פונקציית loss שמקבלת שני תוצאות ומחזירה מספר אי-שלילי. המשימה – בהינתן סט אימון $S = \{(X_i, Y_i)\}_{i=1}^n$ שנדגמו מ- D מצלחת ה- D . נרצה להציג היפותזה, מסווג $\mathcal{Y} \rightarrow \mathcal{H}$ שambilא למינימום את: $L_D(h) = \mathbb{E}_{(X,Y) \sim D} [\ell(Y, h(X))]$. אנחנו לא יודעים את D ולכן לא ניתן לפתור את זה בצורה ישירה. גישת פתרון – קובעים מחלוקת היפותחות $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ ומוחזירים את היפותזה שמנצערת את ה-loss האמפירי, כאשר הממוצע הוא על סט האימון: $L_S := \frac{1}{n} \sum_{i=1}^n \ell(Y_i, h(X_i))$



aicot המסוג שאנו מוחזרים תלויה בפרמטרים מרכזים:

- f_D^* - המסוג האופטימלי.
- h_D^* - מי שיש לו את ה-loss population loss הבי נמור בחלוקת, זה תלוי בחלוקת ולא בסט האימון.
- h_S^* - מי שמביא את הפרדיקציה הבי טובה לסט האימון, ERM למשל.
- \bar{h} - היפותזה שנבחרה, שערוך.

ניתן לפרק את ה-population error ל-3 גורמים עיקריים:

1. approximation error: במה המכלה מסוגלת לבטא פונקציות, במה היא "עשירה".
2. estimation error: שגיאת הכללה, האם עשינו overfit או לא. אם יש לנו אינספור נקודות בסט האימון, כנראה שההיפותזה שעבודה טוב על סט האימון תעשה טובה גם על העולם האמיתי.
3. training error: נובעת מכך שלא תמיד מצליחים למצאו אלגוריתם ERM, ולכן יש שגיאה שנובעת מהאימון.

כל הפקטורים האלה עומדים בלב הלמידה הסטטיסטית, והכול זה trade-off ביןיהם. בעיית המזעור של ה-loss האמפירי נבחרת להיות קמורה, אךשה training error אפס.

למידה קלאסית (מה שראינו עד בהקורס):

נחשב על מסוגים לנאריים, כאשר $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \mathbb{R}^k$. יש לנו היפותחות של מפרדים לנאריים: $\{W \in \mathbb{R}^{k,d} \rightarrow x \mapsto Wx\}$. במודלים האלה בעית המזעור האמפירי היא קמורה. בוחרים loss ומודלים לנאריים כדי שזו תהיה בעיה קמורה.

כאן אנחנו מוצאים את bias-variance trade-off: אם מגדילים את \mathcal{H} , ה-loss approximation error יורד אבל יש יותר סיכוי לעשות overfit לדאטא ולכן ה-loss estimation error יורד ולהיפך.

למידה عمוקה (רשתות נירונים):

ברשותנו נירונים המצביעו שונה. נחשב על זה כטרנספורמציות שומרכבות אחת על השניה, הן לא לנאריות אלא פרמטריות. ככל טרנספורמציה זו קוראים שכבה. מבחינת אופטימיזציה: בלמידה عمוקה בעית מזעור השגיאה אינה קמורה. במקורה הגורע זה ממש לא טוב, כי זה בעיה קשה. בפועל, \mathcal{H} אינו יחיד כי יש מן היפותחות שעשוות fit לדאטא, ו-GD מצליח אייכשו למדוע אותה. לא בדיקות יודעים למה. מבחינת הכללה ואקספרסיביות המצביעו יותר מזרע.

- פחות או יותר ניתן להראות אמפירית שחלק מה-ERM מצללים מצין ואחרים לא. עם דאטא טיפוסי, הפתרונות ש-GD מספק מצללים טוב. לעומת זאת, **לאופטימיזציה יש חלק גדול בהכללה, לעומת למידה סטטיסטית.**
- אם ניקח \mathcal{H} גדול יותר עם יותר פרמטרים, שגיאת הכללה יורדת אבל גם ה-loss approximation error יורד. זה בכניגוד ללמידה סטטיסטית שם \mathcal{H} גדולה מדי מסוכנת לנו כי יכול לגרום fit-overfit.

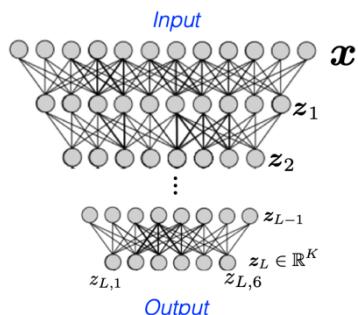
תרגול 8 (למידה عمוקה)

[מודול למידה عمוקה](#)

למידה عمוקה היא אחת ההצלחות הגדולות בתחום מדעי המחשב. מדובר במודלים שմבינה פרקטית הגינו להצלחה שלא הייתה כהווגמתה בשנים קודמות. יש סיבות רבות לכך שהוא שזה מפטיע שזה אכן עובד כך. אחת הסיבות, היא שהfonקציות שלhn אנחנו עושים אופטימיזציה כשאנו מאמנים רשותות, הן מאוד לא קמורות. נקבל objective שווייה מאוד לא קמור, אבל עדין מאפטמים אותו על ידי SGD, באופן מדרים, המשקלים מתכנסים כמעט בכל המקרים בשארת מספיק גדול, למינום גלובלי. עדין אין הבנה תאורטית מלאה של התופעות האלה. נגידו מחלוקת היפותזות באופן הבא:

$$\mathcal{H} = \left\{ \mathbf{x} \rightarrow \mathbf{h}_\theta \left(g_\xi \left(f_\phi(\mathbf{x}) \right) \right) : \theta \in \Theta, \xi \in \Xi, \phi \in \Phi \right\}$$

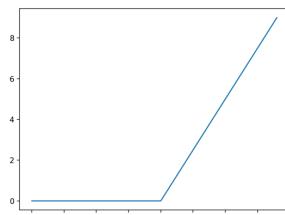
כאשר f_ϕ, g_ξ, h_θ הן פונקציות לא LINEARITIES, שאנו מרכיבים אחת על השנייה. אם הן הן LINEARITIES, הפונקציה הכללית מ- \mathbf{x} לפט היא יייתה גם LINEARITY. אנחנו לא רצאים מסווגים LINEARITIES, אלא מסווגים יותר משוכלים. זו למשל רשות עם 3 שכבות, וכל היפוטזה היא רשות נוירונים, שהוא מסווג. נבצע אופטימיזציה עם SGD מעל loss function שבחור.



הארQUITECTURA הבסיסית – feed-forward fully-connected: הרעיון הוא שהfonקציות שלhn הן רצף של פעולה של פונקציה LINEARITY, שאחריה מפעלים פונקציה פשוטה שהיא לא LINEARITY. נעשה זאת הרבה פעמים. נניח שהקלט לרשות הוא וקטור \mathbf{x} , נסמן $\mathbf{z} = \mathbf{z}_0$. באופן איטרטיבי נפעיל את התחילה הבא: ניקח את מה שמגע מהשכבה הקודמת \mathbf{z} ככפול אותו במטריצה \mathbf{W} (נפעיל פונקציה LINEARITY ונכבל וקטור חדש) ואז עליינו נפעיל פונקציה לא LINEARITY שפועלת ככשה-כניתה על הוקטור שקיבלנו, היא מכונה activation. activation שפועלת ככשה-כניתה על בהינתן מטריצות $\mathbf{W}_1, \dots, \mathbf{W}_L$ נגידו באופן רקורסיבי $1 - L = l$:

$$\mathbf{z}_{l+1} = \sigma(\mathbf{W}_{l+1}\mathbf{z}_l)$$

עבור סיוג BINARIES: השכבה האחורונה \mathbf{z} תהיה בMiami 1 וניקח את argmax כדי לדעת איך מסווג. כלומר (\mathbf{z}_L) argmax $f(\mathbf{x}; \mathbf{W}) = \text{sign}(\mathbf{z}_L)$ וניקח את ה-argmax בMiami 1 ל-0.1, ..., $L = 1$.



fonkציות activation: activation

$$\bullet \quad \sigma(z) = \text{sigmoid}(z) = \frac{1}{1+e^{-z}}$$

$$\bullet \quad \sigma(z) = \text{ReLU}(z) := \max\{0, z\}$$

אימון באמצעות SGD:

בהתאם $\{x_i, y_i\}_{i=1}^n$ ותוצאות $\{+1, -1\}$ לא נוכל לאמן בלי loss, את zero-one או loss hinge/logistic loss על הfonקציות surrogate שמקבל תיגזע ואת הפלט של הרשות מהשכבה האחורונה \mathbf{z} : נניח SGD על הfonקציה:

$$f(\mathbf{W}) = \frac{1}{n} \sum_i \ell(y_i, z_L(x_i; \mathbf{W}))$$

בשביל זה נctrיך לדעת לחשב ביעילות את הגראינטום: $\nabla_{\mathbf{W}} \ell(y, z_L(x; \mathbf{W}))$. לא ניכנס לפרטים של האלגוריתם הכללי של back-propagation, אך נסתכל על רשות מאוד פשוטה כדי לקבל תחושה.

איך נגזר פונקציה שהיא הרכבה של פונקציות (כמו ברשות)? כל השרשרת:

$$f(x) = y(z(x)) \Rightarrow \frac{df}{dx}(x) = \frac{d}{dz}y(z(x)) \cdot \frac{d}{dx}z(x)$$

$$f(x) = g(u_1(x), \dots, u_m(x)) \Rightarrow \frac{d}{dx}f(x) = \sum_i \frac{\partial}{\partial u_i}g(u_1(x), \dots, u_m(x)) \cdot \frac{d}{dx}u_i(x)$$

אז במקרה של רשות נוירונים עם 2 שכבות כאשר h פונקציית activation, נניח את הfonקציה $(h(w_2 \cdot h(w_1 \cdot x)) \cdot \ell)$. נקבל:

$$\frac{\partial}{\partial w_1} \ell(h(w_2 \cdot h(w_1 \cdot x))) = \ell'(h(w_2 \cdot h(w_1 \cdot x))) h'(w_2 \cdot h(w_1 \cdot x)) \cdot w_2 \cdot h'(w_1 \cdot x) \cdot x$$

$$\frac{\partial}{\partial w_2} \ell(h(w_2 \cdot h(w_1 \cdot x))) = \ell'(h(w_2 \cdot h(w_1 \cdot x))) h'(w_2 \cdot h(w_1 \cdot x)) \cdot h(w_1 \cdot x)$$

רקע:

בלמידה מפוקחת לכל מופע היה תיוג, והמטרה שלנו הייתה לחזות אותו. גם כאן זו המטרה, אבל ההבדל הוא שבלמידה מפוקחת ה-*train* וה-*test* היו תהליכי נפרדים. כאן ה-*setting* יותר מתאים בMOVED שמקבלים דוגמאות אחת אחרי השניה, ונותנים פרדיקציות לדוגמאות האלו ומודדים אותן על הפרדיקציות האלה. יש לנו תפקוד כפול – גם לבבא, וגם לשפר את הנבוי תוך כדי התהילה.

למשל, אם משתמש מגע לאתר בו יש פרסום, נרצה להציג לו פרסום הקשור לו לחוץ עליון. כל משתמש הוא דוגמה שיש לה תחומי עניין, הרגאי גליה וכו'. נרצה בזמן את פגוע ולהביא לכל אחד את הפרסומות הביקולוניות אליו, אבל גם לשפר את הנבוי עם הזמן. אחרי שככל משתמש רואה פרסום, יוכל לדעת אם צדקנו או לא – האם המשתמש לחוץ על הפרסומת – ולהשתמש בה כדי לשפר את המודל. ה-*framework* של הלמידה שלנו פועל כך:

1. האלגוריתם מקבל דוגמה לא מתוגנת x_t .
2. האלגוריתם עושה פרדיקציה של התיוג \bar{y}_t . נתרץ במרקחה שבו יש מרחב היפותחות \mathcal{H} ו- $h(x_t) = h^*(x_t)$ עבור $\mathcal{H} \in h$.
3. האלגוריתם מקבל את התיוג הנוכחי y_t . נניח במרקחה שלנו ריאלייזציות, כלומר קיימ $\mathcal{H} \in h^*$ כך ש- $(x_t) \in h^*$. התיוגים נקבעים על ידי היפותזה בשלשה בחלוקתם שלנו.
4. האלגוריתם סופג הפסד של 1 אם $y_t \neq \bar{y}_t$ ו-0 אחרת.

נרצה באופן כללי לעשות מספר קטן של טעויות. נחסום את מספר הטעויות (mistake bound).

אלגוריתם פשוט – CONsistent

נניח כי $\infty < |\mathcal{H}|$. ככל צעד t , נגדיר את \mathcal{H}_t להיות קבוצת היפותחות שקובסיסטנטיות עם הדוגמאות עד זמן t (כולל). לעומת זאת מתקיים בזמן t כי $y_{t-1} = h(x_{t-1}) = y_1, \dots, h(x_1) = y_t$. $\mathcal{H}_t \Leftrightarrow h(x_1) = y_1, \dots, h(x_{t-1}) = y_{t-1}, h_t \in \mathcal{H}_t$.

משפט: אלגוריתם CON מבצע בכל היוטר $1 - |\mathcal{H}|$ טעויות.

בالمור, מספר הטעויות חסום על ידי $1 - |\mathcal{H}|$ ולכן נמשיך "לנצח" לבצע בכל היוטר בדמות כזו של טעות, וכך לא נטהה יותר. הוכחה:

- אבחנה 1 – $\mathcal{H}_{t+1} \subseteq \mathcal{H}_t$. אם יש לנו היפותזה שנמצאת ב- \mathcal{H}_{t+1} היא קובסיסטנטית עם כל הדוגמאות בזמן t וגם ב- \mathcal{H}_{t-1} ומכך זה נובע.
- אבחנה 2 – אם יש טעות בזמן t , אז: $1 - |\mathcal{H}_t| \leq |\mathcal{H}_{t+1}|$. אם טעהנו בזמן t , ההיפותזה שעלה בסיסה הייתה הפרדיקציה נלקחה מזמן t . היא טעהה, ולכן מפרקת היא מפרקת החוצה במהלך האיטרציה. באיטרציה הבא אם זרקנו לפחות אחת מכל נקבול $1 - |\mathcal{H}_t|$.
- אבחנה 3 – לכל t מתקיים $1 \geq |\mathcal{H}_t|$, מכיוון ש- $h^* \in h$ מהנחה הריאלייזציות.

נגיד ש- m זה מספר הטעויות שנעשות עד זמן t :

$$1 \leq |\mathcal{H}_{t+1}| \leq |\mathcal{H}_1| - m = |\mathcal{H}| - m \Leftrightarrow m \leq |\mathcal{H}| - 1$$

אלגוריתם HALving

נניח כי $\infty < |\mathcal{H}|$ ו- $\{0,1\} = \mathcal{Y}$. האלגוריתם פועל אותו דבר כמו CON, אבל הפרדיקציה מבוססת על majority:

$$\begin{aligned} n_0 &= |\{h \in \mathcal{H}_t : h(x_t) = 0\}| \\ n_1 &= |\{h \in \mathcal{H}_t : h(x_t) = 1\}| \end{aligned}$$

אם $\bar{y}_t = 0$ אז $n_0 > n_1$, ואחרת $1 - \bar{y}_t = 0$.

משפט: אלגוריתם HAL עושה בכל היוטר $|\mathcal{H}| \log_2 |\mathcal{H}|$ טעויות. הוכחה:

- אבחנה 1 – $\mathcal{H}_{t+1} \subseteq \mathcal{H}_t$ כמו ב-CON.
- אבחנה 2 – אם יש טעות בזמן t , אז: $|\mathcal{H}_{t+1}| \leq \frac{1}{2} |\mathcal{H}_t|$, כי זרקנו לכל היוטר $\frac{1}{2}$ מהדוגמאות שהוא.
- אבחנה 3 – לכל t מתקיים $1 \geq |\mathcal{H}_t| \geq |\mathcal{H}|$ כמו ב-CON.

נגיד ש- m זה מספר הטעויות שנעשות עד זמן t :

$$1 \leq |\mathcal{H}_{t+1}| \leq \left(\frac{1}{2}\right)^m |\mathcal{H}_1| = \left(\frac{1}{2}\right)^m |\mathcal{H}| \Leftrightarrow m \leq \log_2 |\mathcal{H}|$$



המחר שנסלט באנו הוא שבדי לששות פרדיקציה צריך להרים את ה-*instance* working set כדי שנוכל לעשות את ה-majority. זאת לעומת CON, שם השתמשו כל פעם רק בהיפותזה אחת. מצד שני, כדי לשמור את ה-working set גם בכמה צריך לבדוק את הקונסיסטנטיות כל פעם, אז זה בורא לא מחר גודל. באופן כללי, שימוש נאי של HAL/CON דורש $(|\mathcal{H}|)O$ פעולות בלבד. לעומת זאת, לפעמים ל- \mathcal{H} יש מבנה מסוים, ואז ניתן ממש אוטם במימוש הרבה יותר מהה.

מפרידים לנארים עם אלגוריתם ה-perceptron:

במקרה זה $\{+1, -1\} = \mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}^d$ ו- $w \in \mathbb{R}^d$: $x \cdot sign(w) \rightarrow y$. נניח realizability: כלומר קיימים w^* כך שמתקיים $y_t = sign(w^* \cdot x_t)$ לכל t . באופן מחלקת ההיפותזה תהיה אינסופית, וכן אי אפשר להשתמש בהבטחה של HAL/CON.

תיאור האלגוריתם:

1. מתחילה $w_1 = 0$.
2. עבור $t = 1, 2, 3, \dots$:
 - a. בהינתן דוגמה x_t , עושים פרדיקציה \bar{y}_t לפי $x_t \cdot w_t$.
 - b. אם יש טעות, כלומר $y_t \neq \bar{y}_t$ נכניס את העדכון הבא: $w_{t+1} = w_t + y_t x_t$.
 - c. אם אין טעות, לא מעדכנים.

אינטואיציה: נניח שבזמן t קיבל את x_t ואת השגיאה עליו (טעינו). אז לאחר העדכון מתקיים:

$$w_{t+1} = (w_t + y_t x_t) \cdot x_t = w_t \cdot x_t + y_t \|x_t\|^2$$

אם התוווג היה חיובי, התגובה של x_t (מה שהפרדיקטור נותן לפני שימושים *sign*) עלתה, ולהיפך. ככל יותר ניכנס תיקון לבזון הנכון.

חסם של השגיאה של perceptron: יהי S רצף כל דוגמאות מתויגות קונסיסטנטיות עם פונקציה $sign(w^* \cdot x) \rightarrow x$ עבור $\gamma = \min_{(x,y) \in S} y w^* \cdot x$ עם אורך ייחידה. נסמן בתוור $\|x\|$.

$$\max_{(x,y) \in S} y w^* \cdot x = R. \text{ נסמן ב-}\gamma\text{ את המargin של } w: \gamma = \frac{R}{\|w\|}.$$

או מספר הטעויות שהפרסptron עשה על S הוא **קטן או שווה** $\frac{R}{\gamma}$.

- נשים לב שכאן אנחנו לא מניחים כלום על איך הדוגמאות מוצירות. למידה online יותר מחמירה מבוסנת זהה מלמידה מפרקחת, כי הדוגמאות יכולות להילך מהתפלגות "לא טוביה".
- נשים לב שהמשמעות של $x \cdot w^*$ היא ש- $sign(w^* \cdot x) = 1$ תמיד נכון. נכפיל את זה ב- y וזה תמיד יהיה חיובי.

טענה 1: אם יש טעות בזמן t , אז $\gamma \geq w_t \cdot w^* + w_{t+1} \cdot w^*$.
 הוכחה: נניח ש- x_t היה חיובי ($y_t = +1$). אז $w_{t+1} = w_t + x_t \cdot w^* \geq w_t \cdot w^* + w_t \cdot w^* = w_t \cdot w^* + x_t \cdot w^* \geq w_t \cdot w^* + \gamma$. בולם $\gamma \geq 0$. באותו אופן נטפל במקרה בו x_t היה שלילי.

טענה 2: אם יש טעות בזמן t , אז $\|w_{t+1}\|^2 \leq \|w_t\|^2 + R^2$. בולם המשקלים לא גדלים מהר מדי.
 הוכחה: אם x_t היה חיובי אז:

$$\|w_{t+1}\|^2 = \|w_t + x_t\|^2 \leq \|w_t\|^2 + \|x_t\|^2 + [2w_t \cdot x_t] \leq \|w_t\|^2 + R^2$$

נשים לב כי $0 \leq x_t \cdot w_t$ כי יש טעות בזמן t . אם x_t חיובי, אז $-1 = sign(w_t \cdot x_t) \leq 0$. באותו אופן נטפל במקרה בו x_t היה שלילי.

הוכחת המשפט: טענה 1 אומרת שאחריו מ- m טיעויות מתקיים $\gamma \cdot m \geq w_t \cdot w^*$, כי כל פעם שיש טעות התגובה קטנה בפחות γ . יש מ- m טיעויות ולכן m טיעויות $\gamma \cdot m \geq R \cdot m$ (התחלנו מ-0).

טענה 2 אומרת שאחריו מ- m טיעויות $\|w_t\|^2 \leq \|w_{t+1}\|^2 \leq \|w_t\|^2 + R^2$ כי כל פעם שעשינו טעות זה גודל בכל היתר R^2 ולכן לאחר m טיעויות זה גודל ב- $R^2 \cdot m$. בולם: $R \cdot m \leq \sqrt{m} \|w_t\|$. כיוון ש- $\|w_t\| \leq \|w^*\|$ הוא וקטור יחידה, מתקיים $\|w_t\| \leq \|w^*\| \cdot \sqrt{m}$ (קושי-שורץ?). נקבל:

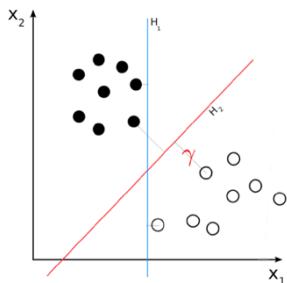
$$m \cdot \gamma \leq w_t \cdot w^* \leq \|w_t\| \cdot \|w^*\| \leq \sqrt{m} \cdot R \Leftrightarrow \sqrt{m} \leq \frac{R}{\gamma} \Leftrightarrow m \leq \left(\frac{R}{\gamma}\right)^2$$

המקרה הלא ריאלייזבלי – Regret Minimation

במקרה זה, אי אפשר להבטיח שום דבר אבסולוטי על מספר הטעויות שנעשה. מה שמסתכלים עליו נקרא regret אחריו זמן T : $Regret(T) = \sum_{t=1}^T \Delta_{zo}(\bar{y}_t, y_t) = \min_{h \in \mathcal{H}} \sum_{t=1}^T \Delta_{zo}(h(x_t), y_t)$. אנחנו משווים את מה שקרה, למאה שהיא קורה אם היינו לוקחים את ההיפותזה הבזetta מתוך המחלקה שהיא הכי טובה. להשיג regret לנאר ב- T זה טריוני, מה שמעניין זה להגיע לפחות regret sub לנאר.

תרגול 9 (online)

הדאטה מגיע בצורה סדרתית. בזמן \hat{t} אנחנו מוחיבים לחזות תיוג כלשהו, ואז מקבלים את התיוג האמתי וונענש אם טעינו. נספוג 1 על טעות ו-0 אם צדקנו. המטרה היא לטעות כמה שפחות. נניח את המקרה ה-realizable, יש היפוטזה במחלקה שמסוגת את כל הדאטה בצורה נבונה.



נזכיר במסוגים לינאריים עם $0 = b$ בה"ב. נסתכל על דאטה שהוא פריד לינארית – קיימים w^* שמסוג את הדאטה בצורה נבונה, עם $margin$ שהוא z . ההיפוטזה היא $(b) h(x) = sign(w \cdot x + b)$, התוצאות הם $\min_{x \in S} \frac{|w^* \cdot x|}{\|w^*\|} = \gamma$.

ראינו את אלגוריתם perceptron והוכחנו שבמקרה זהה כמות הטוויות $\frac{R^2}{\gamma^2} \leq M$ כאשר לכל γ מתקיים $R \leq \|x\|$. האם זה החסם הכי טוב? נראה שלא. התלות ב- γ היא הדוקה. נבנה דוגמה ונוכיח שהאלגוריתם עוזה עליה $\frac{1}{\gamma^2}$ טוויות, או אפילו שיפור את התלות ב- γ (נבחר את הנורמה להיות 1).

טענה (מטלה 6, תשפ"ב-א): לכל $1 < \gamma < 0$ קיים מספר $0 < d$, וקטור $w \in \mathbb{R}^d$ וסדרת של m נקודות: $(x_1, y_1), \dots, (x_m, y_m)$ אשר:

1. הנורמה שלן היא $1 : \|x_i\| = 1$.
2. הנקודות פרידות לינארית עם γ : $\frac{y_i w^* \cdot x_i}{\|w^*\|} \geq \gamma$.
3. אלגוריתם perceptron מבצע $\left\lfloor \frac{1}{\gamma^2} \right\rfloor$ טוויות.

הוכחה: הבחירה תהיה עם הבסיס הסטנדרטי. נבחר את המימד $d = \left\lfloor \frac{1}{\gamma^2} \right\rfloor$ כאשר $x_i = e_i$ ו- $y_i = -1$ לכל $i \leq m \leq i-1$. נראה שה- w -שה-perceptron מחזיק בזמן γ והוא צירוף של וקטורי הבסיס עד 1 – $w_i = -\sum_{j=1}^{i-1} e_j$.

• בסיס: $0 = w_1$ וזה נכון באופן ריק.

• צעד: בזמן i מקבלים נקודה $e_i = x_i$ ומסוג לפיה $1 = y_i w^* = sign(w_i \cdot e_i) = sign(0) = sign(w_i \cdot x_i) = sign(w_i \cdot x_i)$ זאת כי מהנתן האינדוקציה w הוא צירוף של וקטורי בסיס שלא מכיל את e_i , ולכן יהיה תיוג חיובי. האלגוריתם יטעה (כי התיוג שליל) והוא יעדכן את המשקלים:

$$w_{i+1} = w_i + y_i x_i = -\sum_{j=1}^i e_j$$

נשאר להראות את 2: נבחר את $(1, 1, \dots, 1)$, וקטור מנורמה 1, שלילי, ושווה בכל ה-components שלו. נחשב:

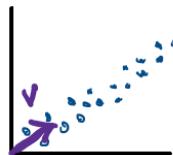
$$\frac{y_i w^* \cdot x_i}{\|w^*\|} = \frac{y_i w^* \cdot e_i}{\|w^*\|=1} = y_i w^* \cdot e_i = (-1) \cdot \left(-\frac{1}{\sqrt{d}}\right) = \frac{1}{\sqrt{d}} = \frac{1}{\sqrt{\left\lfloor \frac{1}{\gamma^2} \right\rfloor}} \geq \gamma$$

הערה: אפשר לבנות דוגמה $x_i = Re_i$ ולהראות חסם אחר.



למידה לא מפוקחת

בלמידה לא מפוקחת אנחנו לא עובדים עם תוצאות בכלל, הדאטה אינם מותג. כל מה שיש לנו זה $\mathcal{X} \in \mathbb{R}^n, \dots, \mathbf{x}_1$, והמטרה שלנו היא למצוא מבנה בדוחהו תוך הדאטה שיהיה שימושי במשימה בהמשך. הרבה פעמים בעולם האמתי יותר קל להציג דוגמאות לא מתיוגות. הקושי כאן הוא שהמטרה לא מוגדרת היטב, גישות שונות של למידה לא מפוקחת מגדרות objective שונת:



$$\mathbf{x} \approx \mathbf{z}; \mathbf{z} \approx \mathbf{y}$$

1. **הזרת מימדים (Dimensionality Reduction)**: הדאטה מוגע מימדי d , אבל הוא יושב בתת-מרחב d -ממדי $d < r$. למשל $\mathbb{R}^2 = \mathcal{X}$ אבל אפקטיבית ניתן לתאר את הנקודות על ידי סקלר בדוחהו כפול וקטור d -ממדי.
2. **סיווג לקבוצות (Clustering)**: מחלקים את הדאטה לקבוצות, קלאסרים. נוכל ליצג כל קלאסער על ידי אב טיפוס אחד שנשמר בזיכרון ועוד נציג את הדאטה בצורה יותר מצומצמת.
3. **יצור דאטא (Generative modeling)**: לומדים התפלגות שמאפשרת לנו ליצור דאטא. מניחים שהדאטה נוצר \mathcal{D} מתוך התפלגות, ולאחר מכן נלמד אותה נוכל ליצור דאטא מתוך התפלגות של העולם.

Dimensionality Reduction

מעבר בין \mathcal{X} לתת-מרחב V :

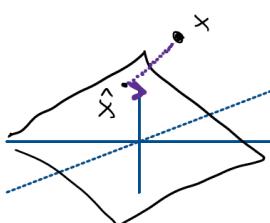
נניח שמרחב הקלט שלנו הוא $\mathbb{R}^d = \mathcal{X}$, ונסתכל על מקרה שבו הדאטה $\mathbf{x}_n, \dots, \mathbf{x}_1$ נמצא בתת-מרחב d -מימי $d < r$. תת-מרחב זה של \mathbb{R}^d ניתן ליצג על ידי בסיס אורותונורמלי v_r, \dots, v_1 , שעבורו מתקיים $\sum_{i=1}^r v_i = v_j$.

נסמן את הוקטוריים במטריצה $V = [v_1 \dots v_r]$. נשים לב שמאורתונורמליות מתקיימים $V^T V = I$. השתמש ב- V כדי להתייחס לבסיס שעמודות המטריצה פרושות $\{v_r, \dots, v_1\}$.

אם $\mathbf{x} \in \mathcal{X}$: במקרה זה הוא נמצא בתת-המרחב V , אז קיימים וקטור מקדמים $r \in \mathbb{R}^r$ כך ש: $\mathbf{x} = Vr = \sum_{i=1}^r a_i V_i = Va$. נכפיל משמאלה ב-

V^T ונקבל: $\mathbf{x} = V^T Vr = V^T Va = a \Rightarrow a = V^T \mathbf{x}$. זה היצוג של הנקודה במרחב המקורי \mathbb{R}^d , וזה היצוג של הנקודה בתת-המרחב. נחשוב על הפעולות הבאות:

- encoding: המעבר מ- \mathbf{x} לתת-המרחב באמצעות $Va = \mathbf{x}$.
- decoding: המעבר לבסיס המקורי מ- \mathbf{x} באמצעות $a = V^T \mathbf{x}$.



אם $\mathbf{x} \notin \mathcal{X}$: נחפש $\hat{\mathbf{x}} \in \mathcal{X}$ שהוא הכי קרוב ל- \mathbf{x} : $\hat{\mathbf{x}} = \arg \min_{x' \in \mathcal{X}} \|x' - \mathbf{x}\|^2$. ככל מרחק הקירוב הטוב ביותר ל- \mathbf{x} בתת-המרחב V . ניתן לפתור את זה באמצעות המינימיזציה הבאה: $\|Va - \mathbf{x}\|^2 = \min_{a \in \mathbb{R}^r} \|Va - \mathbf{x}\|^2$. נשים לב ש- $f(a) = (Va - \mathbf{x})^T(Va - \mathbf{x})$ היא פונקציית כפולה. לכן, נוכל להשוו את הגרדיינט של הפונקציה ל-0 ונקבל: $\nabla f(a) = V^T(Va - \mathbf{x}) = 0 \Leftrightarrow V^T V a - V^T \mathbf{x} = 0 \Leftrightarrow a = V^T \mathbf{x}$.

קיבלו את ה- a שմגדיר את הפונקציה ולכן $\hat{\mathbf{x}} = Va = VV^T \mathbf{x}$. גם כאן יש לנו encoding ואז decoding חזרה ל- \mathbb{R}^d . אבל בעצם הנקודה לאחר decoding שונה מהמקורית – הטלנו לתוך תת-המרחב V (ביצוג אחר), ואז חזרנו ל- d מימדי המקוריים ואז הגענו לנקודה אחרת – **הטלה האורתוגונלית של \mathbf{x} על V** , זו הנקודה הקרובה ביותר ל- \mathbf{x} בתת-המרחב V .

בעיית PCA (Principal Component Analysis):

נניח שיש לנו $\mathcal{X} = \mathbb{R}^d \in \mathcal{X} = \mathbf{x}_n, \dots, \mathbf{x}_1$ 我们知道 מאמנים שהם יושבים (בערך) בתת-מרחב d -מימי. נרצה למצוא תת-מרחב שימזר את השגיאה:

$$\min_{V \in \mathbb{R}^{d,r}} \sum_{i=1}^n \|x_i - VV^T x_i\|^2 \text{ s.t. } V^T V = I$$

זו בעיית PCA: אנחנו מנסים למצוא תת-מרחב d שאם נטיל את כל הנקודות עליו, סכום השגיאה הריבועית הוא הקטן ביותר. אנחנו מיצגים את תת-המרחב ע"י מטריצה עם d עמודות שהן אורותונורמליות. וקטוריים v_r, \dots, v_1 (העמודות של V) שפוטרים את הבעה נקראים **principal components**.

היצוג על ידי V הוא אינו יחיד, אבל מספיק לנו אחד זהה. מדובר על בעיה שאינה קמורה, האילוצים הם פולינומיים ולא לינאריים. למעשה זה בון המקרים הבודדים בהם הבעיה לא קמורה, אבל יש לה פתרון סגור – יש נוסחה לחישוב PCA.



פתרון הבעה: מתקיים לכל :

$$\|x_i - VV^T x_i\|^2 = (x_i - VV^T x_i)^T (x_i - VV^T x_i) = \|x_i\|^2 - 2x_i^T VV^T x_i + x_i^T VV^T VV^T x_i \\ = \|x_i\|^2 - x_i^T VV^T x_i \underset{tr(scalar)}{=} \|x_i\|^2 - tr(x_i^T VV^T x_i) \underset{tr(ABC)=tr(CBA)}{=} \|x_i\|^2 - tr(V^T x_i x_i^T V)$$

נסכום על כל i ונקבל:

$$\sum_i \|x_i\|^2 - tr(V^T x_i x_i^T V) = \sum_i \|x_i\|^2 - \sum_i tr(V^T x_i x_i^T V) \underset{tr \text{ linearity}}{=} \sum_i \|x_i\|^2 - tr\left(\sum_i V^T x_i x_i^T V\right) \\ = \sum_i \|x_i\|^2 - tr\left(V^T \left(\sum_i x_i x_i^T\right) V\right)$$

זה שקול לבעה הבאה:

$$\max_{V: V^T V = I} tr\left(V^T \left(\sum_i x_i x_i^T\right) V\right) \Leftrightarrow \max_{V: V^T V = I} tr\left(V^T \left(\frac{1}{n} \sum_i x_i x_i^T\right) V\right)$$

מקסימיזציה על הביטוי שקופה למקסימיזציה על הביטוי מוכפל בקבוע $\frac{1}{n}$. נסמן $\Sigma = \frac{1}{n} \sum_i x_i x_i^T$ בתור ה covariance של הדאטא שאינו ממורץ (כי לא חסכנו את הממוצע). סה"כ בעית ה-PCA מירוגמת:

$$\max_{V: V^T V = I} tr(V^T \Sigma V)$$

אפשר לבתוב את Σ כ X : $\Sigma = \frac{1}{n} X^T X$ כאשר $X \in \mathbb{R}^{d,d}$ היא PSD, ולכן ניתן לעשות לה פירוק עצמי $X = UDU^T$

באשר $U \in \mathbb{R}^{d,d}$ אורטוגונלית, $D \in \mathbb{R}^{d,d}$ אלכסונית שעל האלכסון שלה ע"ע ממוניים לפי גודל $d \geq \dots \geq \lambda_1$. נסמן ב- u_i את

$$\text{העמודה של } U \text{ שמתאימה ל-} \lambda_i, \text{ כלומר } U = \begin{bmatrix} | & & & \\ u_1 & u_2 & \dots & u_d \\ | & & & \dots \\ & & & | \end{bmatrix}_{n \times d}$$

משפט: פתרון ל-PCA הוא $u_r = v_r, \dots, u_1 = v_1$. לעומת זאת, בבחירה לגיטימית של Σ וקטורים עצמיים של Σ שמתאים ל- Σ הערכיהם העצמיים הגדולים ביותר.

הערה: נשים לב כי Σ היא מטריצה $p \times p$, ואנחנו צריכים למצוא את Σ ה- r הערכים הגדולים (יש $p-r$ בעסה"ב).

הוכחה: נוכיח בשני שלבים:

1. $u_r = v_r, \dots, u_1 = v_1$ מקבל ערך objective, כלומר נתון ערך ל- $(V^T \Sigma V)$.
2. נראה כי הערך מ-(1) הוא הערך גדול שbowton להשיג.

נניח כי $u_r = v_r, \dots, u_1 = v_1$, מתקיים:

$$tr(V^T \Sigma V) = \sum_{i=1}^r v_i^T \Sigma v_i = \sum_{i=1}^r u_i^T \Sigma u_i = \sum_{i=1}^r u_i^T UDU^T u_i = \sum_{i=1}^r \lambda_i$$

נניח ש- V הוא כזה $I = V^T V \in \mathbb{R}^{r,r}$. נסמן $Z = V^T \Sigma V = tr(V^T UDU^T V)$.

$$ZZ^T = V^T UU^T V = V^T V = I \Rightarrow tr(ZZ^T) = tr(I) = r \Rightarrow tr(Z^T Z) = r$$

בולם $r = \sum_{i=1}^d \|z_i\|^2$ כאשר z_i הוא העמודה ה- i של Z . נסמן $c_i := \|z_i\|^2$. מתקיים $\sum_{i=1}^d c_i \leq r$ לכל i . זה נכון כי ראים כי $I = ZZ^T$. בולם השורות של Z הן אורותנו-NORMALיות ולכן אנחנו יכולים להוציא שורות ל- Z ולהפוך אותה למטריצה אורטוגונלית. הנורמה של העמודות של המטריצה זו היא 1 מאורתוגונליות, ולכן העמודות של המטריצה Z המקורית חייבת להיות עם נורמה לכל היותר 1. מתקיים:

$$tr(V^T \Sigma V) = tr(ZDZ^T) = tr(Z^T ZD) = tr(DZ^T Z) = \sum_{i=1}^d \lambda_i \|z_i\|^2 \leq \sum_{i=1}^d \lambda_i c_i \leq \sum_{i=1}^r \lambda_i$$

מכיוון ש- λ_i לא ניתנים לשינוי, מתקבל למסhum את c_i , נחשוב עליהם מקדים שיכולים להיות בכל היותר 1.

bianing

בහינתן DATA encoding $A = \begin{bmatrix} - & a_1 & - \\ \cdots & \cdots & \cdots \\ - & a_n & - \end{bmatrix}_{n \times r}$ המתאים תחת PCA. כלומר $A = X\Sigma V^T$ כאשר $\Sigma = \frac{1}{n}X^T X$ והוא הוא ה- i -ה- n של Σ . מתקיים:

$$\Sigma = \begin{bmatrix} | & & & | \\ u_1 & \cdots & u_r \\ | & & \cdots & | \\ u_1 & \cdots & u_r \\ | & & \cdots & | \end{bmatrix}_{n \times d}$$

$$\frac{1}{n}A^T A = \frac{1}{n}(XV^T)XV = \frac{1}{n}V^T X^T X V = V^T \Sigma V = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ 0 & \cdots & 0 \\ 0 & \cdots & \lambda_r \end{bmatrix}$$

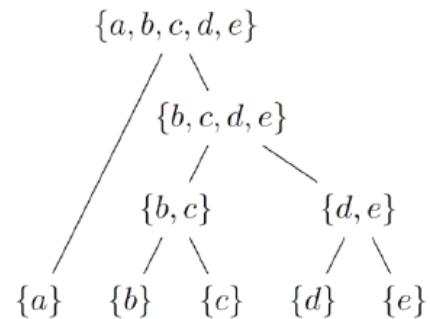
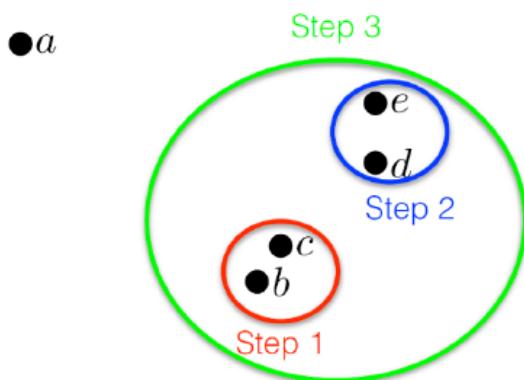
ערכנו לבסיס ייצוג שבו ה- x covariance matrix היא אלכסונית, הפיצ'רים של הדאטה הם חסרי קולרציה ומונורמלים. אם אנחנו

עשימים scaling של כל a_i על ידי $\lambda_1^{-\frac{1}{2}}, \dots, \lambda_r^{-\frac{1}{2}}$, כלומר מיצגים את a_i על ידי XV מטריצה המתקבלת היא מטריצת היחידה. תהליך זה נקרא whitening.

Clustering

בבעיית clustering אנחנו מקבלים נקודות $S = \{x_1, \dots, x_n\} \subseteq \mathbb{R}_{\geq 0}^d$, ויש לנו ממד מרחק d . המטרה שלנו היא לחלק את הדאטה הנתון לקבוצות שנקראות קלאסטרים, כך שהמרחב בין נקודות באותו קלאסטר הוא קטן יותר מהמרחב בין נקודות בקלאסרים שונים.

Linkage-Based Clustering: בשלב ממצב שכל נקודה היא קלאסטר בפני עצמה, ואז באופן איטרטיבי נחבר את הקלאסטרים הכנים. בכל שלב יש לנו valid clustering האלגוריתם בחרתו הבסיסית לא יגיד מתי לעצו, נקבל קלאסטר היררכי ונוכל להחליט לפיה כל מינו שיקולים – למשל אם המרחק בכל הקלאסטרים הוא סף מסוים, או להגיד מספר קלאסטרים שנרצה לקבל. קיבלנו סוג של עץ שנראה如下. כדי שהת hollow יהיה מוגדר היטב, נדרש להגיד מהו ממד המרחק בין הקלאסטרים.

בחירה למרחק:

$$d(A, A') = \min\{d(x, x'): x \in A, x' \in A'\} : \text{Single} \quad .1$$

$$d(A, A') = \max\{d(x, x'): x \in A, x' \in A'\} : \text{Max} \quad .2$$

$$d(A, A') = \frac{1}{|A| \cdot |A'|} \sum d(x, x') : \text{Average} \quad .3$$

מגבילות של השיטה זו:

1. מתחאים את האלגוריתם על ידי תהליך, ולא מדובר ברווח מהו objective שמתמצער בינו לבין אלגוריתמים אחרים.
2. רק לחשב את המרחק בין כל הנקודות לכל הנקודות זה $O(n^2)$, ובמזהה merge-merge שצריך לעשות כדי להגיע לסוף היא $O(n^3)$ ולכן סה"כ נקבל $O(n^3)$.



k-Means Clustering: שיטה זו מושכנת לגישה שנקראה prototype-based clustering, בה הkowskiים מוצגים על ידי אב טיפוס $\mathcal{X} \in \mathbb{R}^d$, והנקודות בkowskiים קרובות אליו טיפוס שהוא μ_j . למשל כאשר $\mathcal{X} = \mathbb{R}^d$ והמרקח הוא אוקלידי, אז שיטתו של k-means היא:

$$I(x; \mu) = \arg \min_{j \in \{1, \dots, k\}} \|x - \mu_j\|^2$$

ה-objective של הבעיה: עבור x נגדיר את המרחק שלו מקבוצת אבי הטיפוס μ :

$$d(x; \mu) = \min_j \|x - \mu_j\|^2 = \|x - \mu_{I(x; \mu)}\|^2$$

אנחנו נרצה למצאו $\{\mu_1, \dots, \mu_k\} \subseteq \mathcal{X} = \mathbb{R}^d$ כך שהדבר הבא ממזער:

$$f(\mu) = \sum_{i=1}^n d(x_i, \mu) = \sum_{i=1}^n \min_j \|x_i - \mu_j\|^2$$

למזור את f במקרה בו יותר זה דבר קשה חישובי, זו בעיה שאינה קמורה. נובל לנסות להריץ GD, אבל יש אלגוריתם ייעודי שモבettaה שהוא באופן מונוטוני מוריד את ה-objective, וזה אלגוריתם k-means.

האלגוריתם: מתחילה קבוצה של אבי טיפוס $\{\mu^{(0)}\}$. לאחר מכן באופן איטרטיבי קובעים $\{\mu^{(t+1)}\}$ בהתאם:

1. Assignment step: לכל $S \in \mathcal{X}$ מוצאים את האינדקס של האב טיפוס הקרוב ביותר $I(x; \mu^{(t)})$. נסמן על ידי S_j^t את הנקודות שהוקטו לאב טיפוס j באיטרציה t .

2. Re-estimation step: מעדכנים את האב טיפוס בצורה הבאה: $\mu^{(t+1)} = \frac{1}{|S_j^t|} \sum_{x \in S_j^t} \mu^{(t)}$. מומצע הנקודות הקיימות.

נראה שה-objective שלנו הוא מונוטוני לא עולה, ולכן ניתן לעצור את האלגוריתם לפי מספר איטרציות ידוע מראש, או מתי שהירידה ב-objective היא מאוד קטנה. בפרט, נראה כי $f(\mu^{(t+1)}) \leq f(\mu^{(t)})$.

נגדיר $\Delta_k = \{(r_1, \dots, r_k) : r_j \geq 0, \sum_j r_j = 1\}$ – probability simplex.

טענה: $d(x; \mu) = \min_{r \in \Delta_k} \sum_{j=1}^k r_j \|x - \mu_j\|^2$ נבחר.

הוכחה: לפי הגדרה $d(x; \mu) = \min_j \|x - \mu_j\|^2 \geq d(x; \mu)$. לכל $r \in \Delta_k$ כיוון $\sum_j r_j = 1$ אז $\sum_j r_j \|x - \mu_j\|^2 \geq d(x; \mu)$. כלומר מומצע משקל של מרחקים מאבי טיפוס, כאשר את מקדמי המומצע נקבעו סקלרים, ולכן הוא גדול שווה מהסקלר הקיים בקבוצה זו. מצד שני, נגדיר $r_j = I(x; \mu)$ ו-0 אחרת. בפרט $r_j \geq 0$ ו-0 אחרת. המשגה על אב טיפוס הקרוב ביותר. זה מוביל לכך ש- $d(x; \mu) = \sum_{j=1}^k r_j \|x - \mu_j\|^2 = d(x; \mu)$.

משפט: $f(\mu^{(t+1)}) \leq f(\mu^{(t)})$

הוכחה: מהטענה הקודמת מתקיים:

$$f(\mu) = \sum_{i=1}^n \min_{r \in \Delta_k} r_{ij} \|x_i - \mu_j\|^2 = \min_{r_1, \dots, r_n \in \Delta_k} \sum_{i=1}^n \sum_{j=1}^k r_{ij} \|x_i - \mu_j\|^2$$

נגדיר פונקציה שתלויה גם ב- Δ_k : $g(\mu, r) = \sum_{i=1}^n \sum_{j=1}^k r_{ij} \|x_i - \mu_j\|^2$ ו- μ ב- \mathbb{R}^d . גם ב- r מוגדר $r_j = I(x_i; \mu)$. מכך נובע ש: $\min_{\mu, r} g(\mu, r) = \min_r g(\mu, r)$.

נראה שבעצם הצעדים של k-means מיען מינימום של $g(\mu, r)$: משאירים את μ קבוע וקובעים את r כדי למזור את הפונקציה g , ואז קובעים את μ ומוחרים לפי r , וכך הלאה בצורה איטרטיבית. סה"כ מובettaה ש- g ירד. בכל פעם שאנו מזערים את μ לפי r אנחנו מוגעים לערך של r .

אחרי t איטרציות k-means מיציר את μ^t . יהי $\Delta_k = \{r_1^t, \dots, r_n^t\} \subseteq \mathbb{R}^d$. ההשומות שנותן את כל המשקל על האב טיפוס הכי קרוב ב- μ^t . בפרט, $r_{ij}^t = 1$ אם $r_i^t \in S_j^t$ ו-0 אחרת. אז:

$$g(\mu^t, r^t) = \min_r g(\mu^t, r) = f(\mu^t)$$



חשיבותו נחושב על מה קורה אם רוצים למצער לפני μ כאשר נקבע את ההשמה:

$$\min_{\mu} g(\mu^t, r^t) = \min_{\mu} \sum_{i=1}^n \sum_{j=1}^k r_{ij} \|x_i - \mu_j\|^2$$

זה ביטוי קמור ולבן לגזר ולהשווות ל-0 ניתן לנו:

$$\mu_j = \frac{1}{\sum_i r_{ij}} \cdot \sum_i r_{ij} x_i = \frac{1}{|S_j^t|} \cdot \sum_{x \in S_j^t} x$$

למעשה עוברים על כל הדוגמאות מ-1 עד n עם ה- j של האב טיפוס הרלוונטי. יהיה בכך רק אם x הוקצה לאב טיפוס ה- j . זה בדיקת ממציע הנקודות שהוקצתו לאב טיפוס j , וזה בדיקת מה שהאלגוריתם עושה בשלב ה-re-estimation.

כלומר נקבל: $g(\mu^{t+1}, r^t) = \min_{\mu} g(\mu, r^t)$. סה"כ:

$$\begin{aligned} f(\mu^t) &= g(\mu^t, r^t) \geq \min_{\mu} g(\mu, r^t) = g(\mu^{t+1}, r^t) \geq \min_r g(\mu^{t+1}, r) = g(\mu^{t+1}, r^{t+1}) = f(\mu^{t+1}) \\ &\quad \text{ולכן } f(\mu^t) \geq f(\mu^{t+1}). \end{aligned}$$

Generative Models

אנו מניחים שקיים התפלגות כלשהי בעולם $p \sim x_i$ באופן IID, ואנו רוצים ללמוד p היפוך ל- x_i בכל הניתן. נתונות לנו נקודות $X \in (x_n, x_1, \dots, x_1)$ ללא תיוגים. מטרתנו היא להשיג תובנה מהדatta שימושה לשימוש אוטומטי ללמידה בהמשך. במודלים גנרטיביים המטריה היא ללמידה התפלגות $(x)p$. אם הצלחנו, יוכל ליצור דאטא ברצוננו, ואם יהיו תיוגים, גם ידע על התפלגות יאפשר לנו לסוגם בצורה אופטימלית.

נדיר משפחה של התפלגות \mathcal{F} (סוג של מחלקה היפותזות) ונשתמש בדआטה כדי לבחור התפלגות ספציפית מתוך המשפחה. נרצה משפחה עשרה מספיק כדי להכיל התפלגות מסוימת מדויקות. מצד שני, לא נרצה משפחה עשרה מיידי כי אחרת נעשו לדआטה ללא משמעות.

KL Divergence: בהינתן שתי התפלגות q, p , נגדיר את המדריך באופן הבא:

$$D_{KL}[p|q] = \sum_x p(x) \cdot \log \frac{p(x)}{q(x)} = \mathbb{E}_{x \sim p} \left[\log \frac{p(x)}{q(x)} \right]$$

- המדריך הזה תמיד אי-שלילי: $D_{KL}[p|q] \geq 0$
- מתקיים $D_{KL}[p|q] = 0 \Leftrightarrow p = q$
- המדריך אינו סימטרי: $D_{KL}[p|q] \neq D_{KL}[q|p]$

Maximum Likelihood: שיטה יסודית ונפוצה לשערון. בוחרים את התפלגות במשפחה שמנדרת את ה- p^* . כלומר:

$$\arg \min_{p \in \mathcal{F}} D_{KL}[p^*|p]$$

נשים לב כי מתקיים:

$$\begin{aligned} \arg \min_{p \in \mathcal{F}} D_{KL}[p^*|p] &= \arg \min_{p \in \mathcal{F}} \mathbb{E}_{x \sim p^*} \left[\log \frac{p^*(x)}{p(x)} \right] \stackrel{\text{log + E linearity}}{=} \arg \min_{p \in \mathcal{F}} \mathbb{E}_{x \sim p^*} [\log p^*(x)] - \mathbb{E}_{x \sim p^*} [\log p(x)] \\ &= \arg \max_{p \in \mathcal{F}} \mathbb{E}_{x \sim p^*} [\log p(x)] \end{aligned}$$

רוצים למקסם תוחלת של משתנה מקרי $(x)p$ והתפלגות שלו היא פונקציה של x , באשר התפלגות של x היא p^* . לכן, יוכל לקחת ממוצע אמפירי ולקבל קירוב אמפירי לבעה של מציאות p . נסמן ב- \hat{p} את המשער של p .

$$\hat{p} = \arg \max_{p \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \log p(x_i)$$

דרך אחרת לפרש את זה, הוא לחפש p שתחתיו התפלגות הדआטה היא מקסימלית:

$$\hat{p}_{ML} = \arg \max_{p \in \mathcal{F}} \sum_{i=1}^n \log p(x_i) = \arg \max_{p \in \mathcal{F}} \log \left(\prod_{i=1}^n p(x_i) \right) = \arg \max_{p \in \mathcal{F}} \prod_{i=1}^n p(x_i)$$



משפחות פרמטריות: בפועל נבוד עם משפחה פרמטרית כלשהו $\{p(x; \theta) | \theta \in \Theta\}$, ואז ניקח \log ונקבל סכום של לוגים:

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \log p(x_i; \theta) = \arg \max_{\theta \in \Theta} \sum_{i=1}^n \log p(x_i; \theta)$$

וקיבלנו בעית אופטימיזיה – הביטוי האחרון נקרא **log likelihood**.

דוגמאות:

מקרה	קלט	פרמטר	התפלגות
1) התפלגות ברנולי	$\mathcal{X} = \{0,1\}$	$\Theta = [0,1]$	$p(x=1; \theta) = \theta \Rightarrow p(x=0; \theta) = 1 - \theta$
2) גאוסיאן עם תוחלת לא ידועה ושונות יחידה	$\mathcal{X} = \mathbb{R}$	$\Theta = \mathbb{R}$	$p(x; \theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$
3) גאוסיאן עם תוחלת ושונות לא ידועות	$\mathcal{X} = \mathbb{R}$	$\Theta = \mathbb{R} \times \mathbb{R}_{>0}$	$p(x; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{1}{2\theta_2}(x-\theta_1)^2}$
4) התפלגות קטגורית	$\mathcal{X} = [k]$	$\Theta = \Delta_k$	$p(x; \theta_x) = \theta_x$ $\Delta_k = \{v \in \mathbb{R}^k : \sum_{i=1}^k v_i = 1, v_i \geq 0\}$

למ' עברו דוגמה 1: כאשר מתקיים $x_i; \theta) = \theta^{x_i} (1 - \theta)^{1-x_i}$. לכן:

$$\ell(\theta) = \sum_{i=1}^n \log p(x_i; \theta) = \sum_{i=1}^n x_i \log \theta + (1 - x_i) \log(1 - \theta) = n_1 \log \theta + n_0 \log(1 - \theta)$$

באשר $n_1 = \#x_i = 0$ -ים $n_0 = \#x_i = 0$ -ים. נרצה למצסם את $\ell(\theta)$, נתעלם מהאלוץ ונראה שהוא מתקיים אוטומטית. הפונקציה היא קעורה שכן מספיק להראות $\ell'(\theta) = 0$ כדי לקבל את נקודת המקסימום.

$$\ell'(\theta) = \frac{n_1}{\theta} + \frac{n_0}{1 - \theta} = 0 \Leftrightarrow \theta = \frac{n_1}{n_0 + n_1} = \frac{n_1}{n}$$

למ' עברו דוגמה 3: בהינתן x_1, \dots, x_n ה- ℓ -ה-log likelihood יהיה:

$$\ell(\theta) = -\frac{n}{2} \log(2\pi\theta_2) - \frac{1}{2\theta_2} \sum_i (x_i - \theta_1)^2$$

עבור θ_2 נתון, נמנסם ביחס ל- θ_1 . זו פונקציה קעורה (פרבולה בוכה). לכן מספיק לגזר ולהשווות ל-0:

$$\frac{\partial}{\partial \theta_1} \ell(\theta) = \frac{1}{\theta_2} \sum_i (x_i - \theta_1) = 0 \Leftrightarrow \theta_1 = \frac{1}{n} \sum_i x_i$$

נשים לב שאין תלות ב- θ . לכן, באשר $\ell(\theta)$ ממוקסם, חייב להיות x_i $\theta_1 = \frac{1}{n} \sum_i x_i$. המשערך של התוחלת הוא בדיקת הממוצע האמפירי וכן זה מאד אינטואיטיבי. נמנסם ביחס ל- θ_2 את הדבר הבא:

$$f(\theta_2) = \ell\left(\theta_1 = \frac{1}{n} \sum_i x_i, \theta_2\right) = -\frac{n}{2} \log(2\pi\theta_2) - \frac{1}{2\theta_2} \sum_i \left(x_i - \frac{1}{n} \sum_i x_i\right)^2$$

מתקיים:

$$\frac{\partial f(\theta_2)}{\partial \theta_2} = -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2} \sum_i \left(x_i - \frac{1}{n} \sum_i x_i\right)^2 = 0 \Leftrightarrow \theta_2 = \frac{1}{n} \sum_i \left(x_i - \frac{1}{n} \sum_i x_i\right)^2$$

קיבלו את השונות האמפירית תחת התוחלת האמפירית.



TLV עבר דוגמה 4: בהינתן $x_1, \dots, x_n \in [k]$ נסמן ב- j את מספר הפעמים ש- j הופיע.

$$\ell(\theta) = \sum_{i=1}^n \log p(x_i; \theta) = \sum_{i=1}^n \log \theta_{x_i} = \sum_{j=1}^k n_j \log \theta_j$$

נרצה למקסם את $\ell(\theta)$ בך ש- $0 \geq \theta_j \geq 1$. נתעלם מהailoz θ_j ונראה אחר בך שהוא מתקיים גם כבה. נחשב:

$$\mathcal{L}(\theta, \lambda) = \ell(\theta) - \lambda \left(\sum_j \theta_j - 1 \right)$$

מתנאי KKT, עבור θ אופטימלי קיים ג'ך ש:

$$\frac{\partial}{\partial \theta_j} \mathcal{L}(\theta, \lambda) = 0 \Leftrightarrow \frac{n_j}{\theta_j} - \lambda = 0 \Leftrightarrow \theta_j = \frac{n_j}{\lambda}$$

מכיוון ש- θ_j פתרונות פיזיולוגיים לבעה שלנו, מתקיים כי $1 = \sum_j \frac{n_j}{\lambda}$ ולכן $n_j = \lambda$ ו- $\lambda = \frac{n}{\lambda}$. נשים לב שעדיין מתקיים האילוז $0 \geq \theta_j$.

שער ביסיאני:

בגישה הביסיאנית, כל דבר הוא משתנה מקרי, והפרמטר שאנו חושם הוא דגימה של המשתנה המקרי θ . ההתפלגות המשותפת של θ ו- x_1, \dots, x_n :

$$P[X_1 = x_1, \dots, X_n = x_n, \Theta = \theta] = P[\Theta = \theta] P[X_1 = x_1, \dots, X_n = x_n | \Theta = \theta]$$

התפלגות המותנית של θ של המדגם x_1, \dots, x_n

$$P[\Theta = \theta | X_1 = x_1, \dots, X_n = x_n] = \frac{P[X_1 = x_1, \dots, X_n = x_n, \Theta = \theta]}{P[X_1 = x_1, \dots, X_n = x_n]} = \frac{P[\Theta = \theta] P[X_1 = x_1, \dots, X_n = x_n | \Theta = \theta]}{P[X_1 = x_1, \dots, X_n = x_n]}$$

$$= \alpha \cdot P[\Theta = \theta] P[X_1 = x_1, \dots, X_n = x_n | \Theta = \theta]$$

כאשר α קבוע מכפלה שאינו תלוי ב- θ .

בהינתן ה-*prior* אפשר לשער את θ :

- באמצעות MAP נחשיר:
- $\theta^* = \arg \max_{\theta} P[\Theta = \theta | X_1 = x_1, \dots, X_n = x_n] = \arg \max_{\theta} P[\Theta = \theta] P[X_1 = x_1, \dots, X_n = x_n | \Theta = \theta]$
- אם ה-*prior* הוא יוניפורמי (שווה לקבוע שלא תלוי ב- θ) אז $\text{MAP} = \text{MLE}$
- $P[\Theta = \theta | X_1 = x_1, \dots, X_n = x_n]$ פשוט מחדירים את התוחלת של ההתפלגות [expected value]

הערה: נשים לב שמתקיים:

$$\begin{aligned} \log P[\Theta = \theta | X_1 = x_1, \dots, X_n = x_n] &= \log \alpha \cdot P[\Theta = \theta] P[X_1 = x_1, \dots, X_n = x_n | \Theta = \theta] = \\ &= \log P[\Theta = \theta] + \log P[X_1 = x_1, \dots, X_n = x_n | \Theta = \theta] = \log P[\Theta = \theta] + \sum_{i=1}^n P[X_i = x_i | \Theta = \theta] \end{aligned}$$

בכל ש- α גדול (יש יותר דוגמאות) אז האימפקט של ה-*prior* קטן.

דוגמה: נתונים מגע מගאוסיאן עם תוחלת לא ידועה ושונות יחידה: ($X; \theta; q$)
נניח prior שהוא $P[\Theta = \theta] = N(\theta; 0; 1)$. ה-*prior* $P[\Theta = \theta] = N(\theta; 0; 1)$.

$$\begin{aligned} P[\Theta = \theta | X_1 = x_1, \dots, X_n = x_n] &= P[\Theta = \theta] P[X_1 = x_1, \dots, X_n = x_n | \Theta = \theta] \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\theta^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\sum_{i=1}^n (x_i - \theta)^2} = e^{-\frac{1}{2}\theta^2} \cdot e^{-\frac{1}{2}\sum_{i=1}^n (x_i - \theta)^2} = e^{-\frac{1}{2}((n+1)\theta^2 - 2\theta \sum_{i=1}^n x_i)} \\ &= e^{-\frac{(n+1)}{2}(\theta - \frac{1}{n+1} \sum_{i=1}^n x_i)^2} = N\left(\theta; \frac{1}{n+1} \sum_{i=1}^n x_i; \frac{1}{n+1}\right) \end{aligned}$$

از שער ה-*MAP* הוא התוחלת של θ שהוא $\theta_{MAP} = \frac{1}{n+1} \sum_{i=1}^n x_i$.



תרגול 10 (לא מפוקחת)

:PCA

בعالום של למידה לא מפוקחת אין לנו תיוגים, ונרצה לבצע פעולה על הדאטא כפי שהוא ולהסיק ממנו מסקנות. אחת המטרות האפשריות היא להוריד את המידע שלו – לפעמים נוח לנו להסתבל על הפיצרים הבולטים בדאטא ולהוריד את המידע כך שנוכנש לשמר את האינפורמציה שיש בדאטא. השיטה הבסיסי הוא הורדת מידע ליניארית – PCA.

נראה שאפשר להציג את בעיית-PCA כבעיה שקופה אבל אחרת – מיקסום של השונות. נראה שהכיוונים במרחב של אותם ה- u 'ים שפותרון בעיית-PCA נוטן, הם הכיוונים שעיליהם השונות בדאטא היא מקסימלית. ומה נרצה שונות מקסימלית? כיון שיש לנו עליות שונות גבורה, הוא כיון שמספר הרבה על היחס בין נקודות הדאטא. אם יש כיון שבו כל הדאטא יושב בנקודה אחת, זה כיון חסר משמעות, ולא מספר כלום על הדוגמאות. לכן נרצה כיונים שיש עליהם שונות גבורה.

שונות אמפירית: עבור נקודות a_1, \dots, a_n , השונות האמפירית מוגדרת להיות הממוצע של המרחק של הנקודה ה- i -הממוקעת של הנקודות, בטיבו. זה ממוצע הסטיות הריבועיות מה-mean. במו השונות שאנו מכירים מהסתברות רק שאנו חצינו רצים על מדגם ולא על משתנה מקרי:

$$\frac{1}{n} \sum_{i=1}^n \left(a_i - \frac{1}{n} \sum_{j=1}^n a_j \right)^2$$

בעיית מיקסום השונות: ראשית, נמרכז את הנקודות סביב 0 – נניח בה"כ כי $\bar{x}_i = x_i - \frac{1}{n} \sum_{j=1}^n x_j$, אחרת נגדיר x_j כ- $x_j - \frac{1}{n} \sum_{j=1}^n x_j$. בפונקציית השונות האמפירית, נסתכל על השונות של הדאטא אחרי שנטילו אותו על עמדות של מטריצה V . השונות של ההטלה:

$$\frac{1}{n} \sum_{i=1}^n \left(VV^T x_i - \frac{1}{n} \sum_{j=1}^n VV^T x_j \right)^2$$

כאשר x_i זה ההיטל של x_i על V – $\frac{1}{n} \sum_{j=1}^n VV^T x_j$ ממוצע ההיטלים שמתאפס (אם יצא את VV^T מה שנשאר מתאפס מההנחה). נרצה להראות של מיקסום את $VV^T x_i$ שקול לביעית-PCA.

קשר בין ההטלה של הוקטור x_i לוקטור עצמו. כאשר נסתכל על ההטלה ונסתכל על הריבב המשלים, נקבל שני וקטורים ניצבים (הטלה אורותוגונלית), נבע מכך שהעמדות של V מאונכות זו לזו. משפט פיתגורס נוכל לכתוב:

$$\|x_i\|^2 = \|VV^T x_i\|^2 + \|x_i - VV^T x_i\|^2$$

נסתכל על המטריצה V שמקסמת את השונות האמפירית של הדאטא:

$$\begin{aligned} \arg \max_{V \in \mathbb{R}^{d \times r}, V^T V = I} \sum_{i=1}^n \|VV^T x_i\|^2 &= \arg \max \sum_{i=1}^n \|x_i\|^2 - \|x_i - VV^T x_i\|^2 = \arg \max \sum_{i=1}^n -\|x_i - VV^T x_i\|^2 \\ &= \arg \min \sum_{i=1}^n \|x_i - VV^T x_i\|^2 \end{aligned}$$

ביחס ל- V הגורם הראשון הוא קבוע, ונקבל מקסימום של מינוס, שהוא כמו לבצע מינימום. הוכחנו שקיים של שתי הבעיות.

הערכים העצמיים: אם هو"ע זה הכיוונים שעל פניהם השונות היא מקסימלית, מה אנחנו מצפים שיהיה הע"ע? השונות עצמה. יהיו $\lambda_1, \dots, \lambda_d$ הע"ע של V . נסתכל על ה"ע" u_k , ועל השונות של הדאטא שמוטל עליו:

$$\begin{aligned} (u_k u_k^T x_i)^T (u_k u_k^T x_i) &= u_k^T (u_k^T x_i) u_k \\ &= u_k^T \underbrace{\left(\frac{1}{n} \sum_{i=1}^n x_i x_i^T \right)}_{\Sigma} u_k \\ &= \frac{1}{n} \sum_{i=1}^n \|u_k^T x_i\|^2 = \frac{1}{n} \sum_{i=1}^n (u_k^T x_i)^2 = u_k^T \left(\frac{1}{n} \sum_{i=1}^n (x_i x_i^T) \right) u_k \\ &= u_k^T (\Sigma u_k) = u_k^T u_k \lambda_k = \lambda_k \end{aligned}$$

:Maximum Likelihood

נתונה לנו מחלוקת התפלגיות \mathcal{F} , למשל משפחת התפלגיות נורמליות עם שונות σ^2 ותוחלת μ :

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

נרצה למצאו את ההתפלגות הספציפית \mathcal{F} שמסבירה את הדadata הבי. ההסתברות שנראה את הדadata שראינו, תחת ההתפלגות הזאת p , היא מקסימלית. זה נקרא **likelihood**.

$$\hat{\theta} = \arg \max_{\theta \in \mathcal{F}} \sum_{i=1}^n \log p(x_i)$$

התפלגות קטגורית (דרך LK): נסכל על קובייה, ההתפלגות של קובאה סופית של ערכים $\{K, \dots, 1, 2\}$ ובכל ערך ההסתברות היא θ_k . בקוביה חוקית כל ההסתברויות $\frac{1}{6}$. בפועל יש לנו $1 - k$ פרמטרים כי האחרון הוא 1 פחוות השאר. נשים לב כי יש לנו את האילוץ שסכום ההסתברויות הזה הוא 1. נתנות לנו n דוגמאות x_1, \dots, x_n ונרצה לשער את $\theta_K, \dots, \theta_1$. יש כאן אינטואיציה: אם יש התפלגות נורמלית עם תוחלת μ , המשער יהיה הממוצע של הדadata. במקרה שלנו: θ_1 זה ההסתברות שנראה 1, המשער של זה יהיה כמות הפעם שיצא לו 1.

חישוב: לפי הגדרה log likelihood הוא לעבור על הדוגמאות ולראות על \log ההסתברות לקבל את x_i :

$$\mathcal{L}(\theta_1, \dots, \theta_K) = \sum_{i=1}^n \log \theta_{x_i} = \sum_{k=1}^K n_k \log \theta_k = \sum_{k=1}^K \frac{n_k}{n} \log \theta_k$$

נסכם **במקום על הדוגמאות, על הקטגוריות** (יש לנו K כאליה), לכל קטgorיה ניקח את $\log \theta$, וממות הפעמים שזה מופיע בסכום זהה, זה כמות הפעמים שהקטgorיה הופיעה בדadata.vrker נקרא θ_k . אם ננחש שההשובה היא $\frac{n_k}{n} = \theta_k$.vrker ננחש שהכפלנו בסוף ב- $\frac{1}{n}$.

נשים לב שלכל $\Delta_k \in (\theta_1, \dots, \theta_K)$ מתקיים, כאשר נציב את הניחוש שלו ב-objective ונחשב את ההפרש:

$$\sum_{k=1}^K \frac{n_k}{n} \log \frac{n_k}{n} - \sum_{k=1}^K \frac{n_k}{n} \log \theta_k = \sum_{k=1}^K \frac{n_k}{n} \log \left(\frac{\frac{n_k}{n}}{\theta_k} \right) = D_{KL}(\hat{\theta} | \theta)$$

כאשר $\frac{n_k}{n} = \hat{\theta}_k$. כיוון ש-LK תמיד או-שלילי, נקבל:

$$\sum_{k=1}^K \frac{n_k}{n} \log \theta_k \leq \sum_{k=1}^K \frac{n_k}{n} \log \frac{n_k}{n}$$

הטלות מטיב תלויות: הדadata מכיל זוגות של הטלות מטיב, יש לנו שלושה פרמטרים $p_0, p_{0|0}, p_{1|0}$:

$$p_0 = 1 - p_{1|0}, p_{0|0} = 1 - p_{1|1}$$

ההסתברות לראות תוצאה i היא $p_{x_1} p_{x_2|x_1}(x_1, x_2) \in \{0, 1\}^2$. יש לנו n זוגות ונרצה לשער את הפרמטרים $p_1, p_{1|0}, p_{1|1}$.

- אפשר בדבר ראשון לנחש שהשער p_1 יהיה מספר הדוגמאות שבהן המטיב הראשון הוא 1 חלקו n .
- p_1 – כמות הדוגמאות שבהן היה (1, 0) חלקו כמות הדוגמאות שהראשון היה 0, נתנה על כך שבטבע הראשוני היה 0. באותו אופן גם על $p_{1|1}$.

$$\begin{aligned} \mathcal{L}(p_1, p_{1|0}, p_{1|1}) &= \sum_{i=1}^n \left(\log p_{x_1^i} + \log p_{x_2^i|x_1^i} \right) \\ &= (n_{11} + n_{10}) \log p_1 + (n_{01} + n_{00}) \log(1 - p_1) \\ &\quad + n_{11} \log p_{1|1} + n_{10} \log(1 - p_{1|1}) \\ &\quad + n_{01} \log p_{1|0} + n_{00} \log(1 - p_{1|0}), \end{aligned}$$

► Taking the derivative with respect to p_1 ,

$$\frac{n_{11} + n_{10}}{p_1} - \frac{n_{01} + n_{00}}{1 - p_1} = 0.$$

► We get that

$$\hat{p}_1 = \frac{n_{11} + n_{10}}{n_{01} + n_{00} + n_{11} + n_{10}} = \frac{n_{11} + n_{10}}{n}.$$

► Similarly,

$$\hat{p}_{1|1} = \frac{n_{11}}{n_{11} + n_{10}} \quad ; \quad \hat{p}_{1|0} = \frac{n_{01}}{n_{01} + n_{00}}.$$