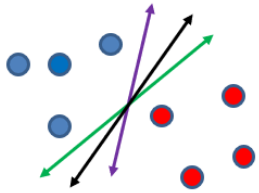
למידת מכונה הרצאה 3

ראינו איך ללמוד כשהחוקים שלנו הם:

- אינסוף אינטרוולים
- אינסוף מלבנים
- כל אוסף סופי של חוקים



מה לגבי אוסף אינסופי של חוקים? הרעיון שלנו:

יש מקרים שהרבה חוקים (אפילו אינסוף) הם בפועל חוק אחד:

כלומר, גם אם יש אינסוף קווים שאפשר לצייר שמפרידים בין נקודות, אפשר לצמצם את זה למספר סופי של **קבוצות** של חוקים, כאשר בכל קבוצה כל החוקים מתנהגים אותו דבר. אנחנו רוצים לדעת כמה **תיוגים** או צביעות אפשריות יש. ואז להכניס את זה לניתוח עבור קבוצה סופית של חוקים. לכל קבוצת נקודות יש מספר סופי של צביעות, ולכל צביעה יש מספר סופי של קווים בפועל.

יש בעיה שאפשר לצפות ראש – לא כל צביעה אפשר להפריד ע"י קו:

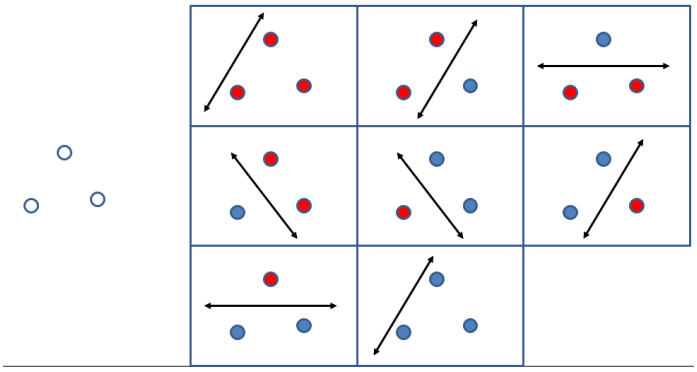
ניפוץ – shattering

תכונה של **אוסף** של חוקים.

יהיו H אוסף של חוקים, S קבוצת נקודות. נאמר ש-H **מנפץ** את S אם עבור כל צביעה אפשרית של S, יש חוק ב-H שמשיג אותו.

דומה (אבל לא בדיוק זהה) לשאלה של כמה צביעות יש. כי יש צביעות שאין קו שיכול להפריד בהן (כמו חלק מהצביעות של 7 נקודות).

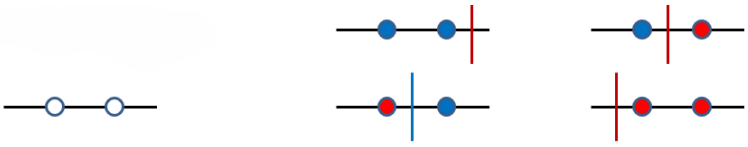
לדוגמה: אוסף הקווים במישור יכול לנפץ **חלק** (אבל לא את כל) הקבוצות של 3 נקודות במרחב:



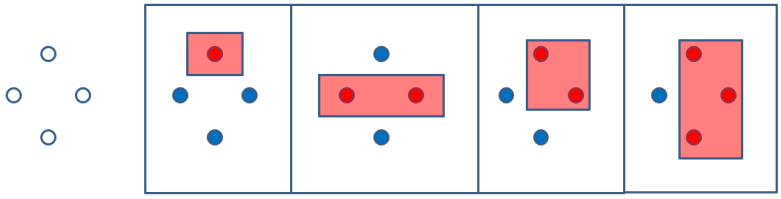
האוסף **לא** מנפץ את הקבוצה של 3 נקודות על אותו קו. כי אי אפשר להשיג את הצביעה כחול-אדום-כחול.

אוסף האינטרוולים החד-כיווני – שמגדיר תמיד את צד ימין אדום (בה"כ) – יכול לנפץ נקודה אחת: יש 2^1 תיוגים:

אוסף האינטרוולים הדו-כיווני יכול לנפץ שתי נקודות: יש 2^2 תיוגים:



אוסף המלבנים החד-כיווניים (בפנים אדום, בחוץ כחול, בה"כ) המקביל לצירים, יכול לנפץ ארבע נקודות (לא כל קבוצה של 4 נקודות – מספיק להראות דוגמה אחת). יש 2^4 תיוגים, נדגים 4 מהם:



מימד VC – Vapnik-Chervonenkis dimention:

מימד ה-VC של אוסף חוקים H, הוא המספר המקסימלי של נקודות שהוא יכול לנפץ. כדי להוכיח שמימד VC של קבוצת חוקים הוא n, צריך להוכיח:

1) שהקבוצה יכולה לנפץ קבוצה של n נקודות (אפשר ע"י דוגמה).
2) שהקבוצה **לא** יכולה לנפץ **אף** קבוצה של n+1 נקודות (צריך להוכיח פורמלית).

לאוסף האינטרוולים החד-כיווני יש מימד-VC 1: הוא יכול לנפץ רק נקודה אחת:

הוא לא יכול לנפץ **אף** קבוצה של 2 נקודות, כי הוא לא יכול לגרום להם להיות בצבעים שונים (כי הוא חד-כיווני).

לאוסף האינטרוולים הדו-כיווני יש מימד-VC 2, כי הוא יכול לנפץ שתי נקודות ולא יכול לנפץ 3, כי הוא לא יכול להפריד אף קבוצה של 3, כי הוא לא יוכל לתת צבע נפרד לנקודה האמצעית:

ראינו שאוסף הקווים הדו-צדדיים במישור יכול לנפץ **חלק** (אבל לא את כל) הקבוצות של 3 נקודות במרחב, ולכן יהיה לו מימד-VC לפחות 3. הטענה שלנו היא שהמימד הוא לא 4, כי הוא לא יכול לנפץ **אף** קבוצה של 4 נקודות.

טענה: לא קיימת אף קבוצה של 4 נקודות שאוסף הקווים יכול לנפץ. נפריד לשני מקרים:

א. נקודה אחת נמצאת בתוך (או על) **הקמור** של השלושה האחרים.
ב. הנקודות במצב כללי.

עבור מקרה א: קו לא יכול לתת לנקודה האמצעית צבע שונה מהנקודות החיצונית. כי הוא יצטרך לעבור בין הנקודה הפנימית לאחד הקודקודים, כלומר לעבור דרך המשולש. אז הוא מפריד 2 קודקודים מהשלישי, אבל הם באותו צבע.

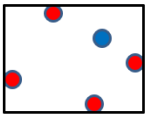
עבור מקרה ב: הקו לינארי, אז אם שתי נקודות הן מצד אחד של הקו, גם כל נקודה ביניהן תהיה מאותו צד. יש נקודה אחת שהיא החיתוך של שני הקווים, אז היא צריכה להיות משני הצבעים:

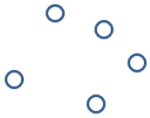
אז לסיכום: אוסף הקווים הדו-צדדיים במישור יכול לנפץ חלק מהקבוצות של 3 קווים (לא את כולם), ולכן יהיה לו מימד-VC לפחות 3. הוא לא יכול לנפץ **אף** קבוצה של 4 נקודות ולכן המימד לא יהיה 4. בסה"כ – המימד הוא 3.

לאוסף המלבנים החד-כיווניים (בפנים אדום, בחוץ כחול) המקביל לצירים, יש מימד-VC 4: יכול לנפץ חלק מהקבוצות של 4 נקודות, כמו שהראינו. הוא לא יכול לנפץ אף קבוצה של 5. נפריד לשני מקרים:

א. נקודה אחת נמצאת בתוך (או על) **הקמור** של האחרים.
ב. הנקודות במצב כללי.

עבור מקרה א: אם הנקודות החיצוניות אדומות, הן חייבות להיות בתוך המלבן. מלבן הוא צורה קמורה, ולכן הנקודה האמצעית חייבת להיות בפנים גם. (ההוכחה עובדת עבור כל צורה קמורה).





עבור מקרה ב: ניקח את הנקודות עם ערכי X ו-Y מינימליים ומקסימליים. כל מלבן שמכיל אותן מכיל גם את החמישי. אי אפשר לצבוע את ה-4 הראשונים באדום והאחרון בכחול.

ראינו שעבור קבוצת נקודות עם 2 תיוגים אפשריים, יש 2^n תיוגים אפשריים, אבל החוקים לא יכולים לכסות את כולם. אנחנו רוצים לשאול באופן כללי, כמה תיוגים שונים הקווים יכולים לתת. וזה בעצם אומר לנו כמה חוקים שונים יש בפועל.

למת סאור – Sauer's lemma: (נקרא גם Sauer-Shelah-Vapnik-Chervonenskis).

נתון: אוסף S של n נקודות, ואוסף חוקים עם מימד-VC d. יהי $\Pi(H,S)$ מספר הצביעות האפשריות ש-H מגדיר ל-S. אזי:

$$\Pi(H,S) \le \sum_{i=0}^{d} \binom{n}{i} \le \left(\frac{en}{d}\right)^d = O(n^d)$$

שזה יכול להיות הרבה פחות מ- 2^n. ככל שאוסף החוקים פשוט יותר (מימד VC נמוך), הו איכול לתת פחות צביעות. מספר הצביעות קובע בפועל את מספר החוקים האמיתי.

ההוכחה:

נזכיר: $\binom{n}{i} = \binom{n-1}{i} + \binom{n-1}{i-1}$

הסבר קומבינטורי לשוויון: בצד ימין, הchoose הראשון מתאר מצב שלא בחרנו את האיבר הראשון (ואז צריך לבחור i מתוך כל השאר). הchoose השני זה המצב שכן בחרנו את האיבר הראשון (ואז צריך לבחור $i-1$ מתוך $n-1$).

נראה דוגמה: עבור קבוצת נקודות x_1, x_2, x_3, נקרא לצבע אחד 0 ולשני 1, ונגדיר h_i את כל אחת הצביעות:

	x_1	x_2	x_3
h_1	0	0	0
h_2	0	0	1
h_3	0	1	0
h_4	0	1	1
h_5	1	0	0
h_7	1	1	0

נמספר את כל הנקודות, ונעבור על כל החוקים ב-H: לכל חוק, אם יש חוק אחר שזהה לו בכל התיוגים חוץ מהנקודה האחרונה, נשים אותם בקבוצות נפרדות:

	x_1	x_2	x_3
h_1	0	0	0
h_3	0	1	0
h_5	1	0	0
h_7	1	1	0

	x_1	x_2	x_3
h_2	0	0	1
h_4	0	1	1

נשים לב: אם נמחק את הנקודה האחרונה, אפשר למחוק את כל H_2.

ל-H יש מימד VC d כלשהו. מה המימד של H_1, H_2?

כמו שרואים, מספר התיוגים של H_1, H_2 לא משתנה אם מוחקים את הנקודה האחרונה. נגדיר: $S' = \{x_1, x_2\}$. ונשים לב שמתקיים:

$$\Pi(H_1,S') = \Pi(H_1,S), \qquad \Pi(H_2,S') = \Pi(H_2,S)$$

מספר התיוגים זהה. ההוספה של הנקודה האחרונה לא דרשה עוד חוק ולכן, מספר החוקים זהה.

ובאופן טריוויאלי: $\text{VC-dim}(H_1) \overset{\text{א}}{\le} \text{VC-dim}(H) \overset{\text{ב}}{=} d$

א – כי פחות חוקים לא יכולים לנפץ יותר נקודות. ב – לפי הגדרה.

נשים לב: ניקח כל קבוצה ש-H_2 יכולה לנפץ. H יכולה לנפץ את הקבוצה הזאת בתוספת x_3.

ולכן מתקיים: $\text{VC-dim}(H_2) \overset{\text{א}}{\leq} \text{VC-dim}(H) - 1 = d - 1$.

א – כי היא לא יכולה לנפץ את הנקודה האחרונה, כי לפי איך שבחרנו את החוקים ל-H_2, כל החוקים שם מסכימים לגבי הנקודה האחרונה.

אז ההוכחה למשפט, באינדוקציה על n ביחס ל-d:

בסיס: עבור $n = d$, (המימד VC שווה למספר הנקודות), מתקיים: $\sum_{i=0}^{d}\binom{d}{i} = 2^d$. מספר הדרכים לתייג את כל הנקודות.

נניח עבור $n - 1$ נקודות או פחות, ו- $d - 1$ נקודות או פחות.

נוכיח עבור n נקודות ומימד VC d. (לא בהכרח שווים). מתקיים:

$$\Pi(H,S) \leq \Pi(H_1,S) + \Pi(H_2,S) = \Pi(H_1,S') + \Pi(H_2,S') \leq \sum_{i=0}^{d}\binom{n-1}{i} + \sum_{i=0}^{d-1}\binom{n-1}{i}$$
$$= \sum_{i=0}^{d}\binom{n-1}{i} + \sum_{i=1}^{d}\binom{n-1}{i-1} = \sum_{i=0}^{d}\binom{n-1}{i} + \sum_{i=0}^{d}\binom{n-1}{i-1} = \sum_{i=0}^{d}\binom{n}{i}$$

אז ראינו בהרצאה הקודמת חסם עליון לטעות, שהיה תלוי בגודל קבוצת החוקים. אם במקום $|H|$, נכניס את $\left(\frac{en}{d}\right)^d$, נקבל (עד כדי קבועים):

$$e(h) \leq \frac{2}{n}\left(d \cdot \log_2 \frac{2en}{d} + \log_2 \frac{2}{\delta}\right)$$

$$e(h) \leq \bar{e}(h) + \sqrt{\frac{8d \cdot \ln \frac{2en}{d} + 8 \cdot \ln \frac{4}{\delta}}{n}}$$

מה המשמעות? יש חשיבות לפשטות של החוקים שלנו. ככל שהמימד VC שלו נמוך יותר, הוא חוסם את הטעות. כלומר – חוק פשוט הוא יותר אמין. ובכלל, הוא נותן לנו לחסום טעות גם במצב שיש אינסוף חוקים.