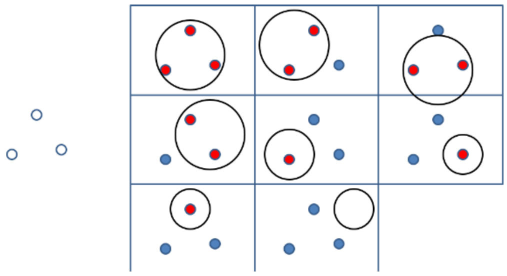
למידת מכונה הרצאה 4

ראינו שאם יש מסווג שעובד לפי חוק "פשוט" שיש לו טעות נמוכה, אז בהסתברות גבוהה הטעות שלו נמוכה על כל הדאטא.

יש עדיין בעיות:

- בעיה סטטיסטית: אם יש התפלגות שנותנת לכל נקודה תיוג בהסתברות חצי בדיוק, אי אפשר ללמוד את זה.
- בעיה חישובית: בדיקה אם גרף הוא 3-צביע. NP hard.

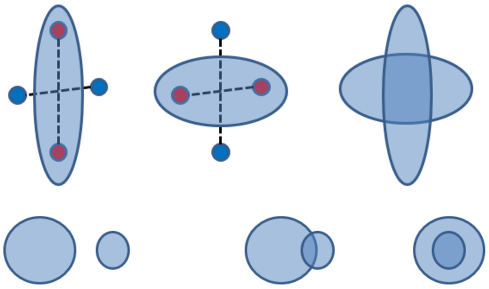
ראינו לדוגמה שלאוסף הקווים הדו-צדדיים יש מימד VC 3. כי הוא יכול לנפץ חלק מהקבוצות של 3 נקודות, ואף קבוצה של 4 נקודות.



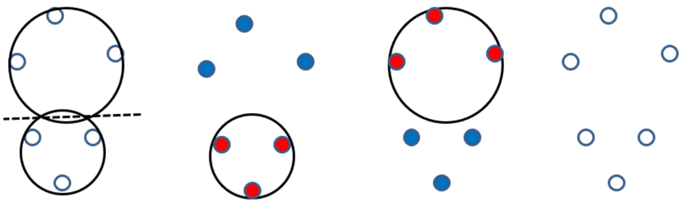
לעיגולים דו-מימדיים חד-צדדיים (בפנים – אדום) יש גם מימד VC 3:
קיים אוסף של 3 שאפשר לנפץ:

ולא קיים אוסף של 4. נחלק לשני מקרים:

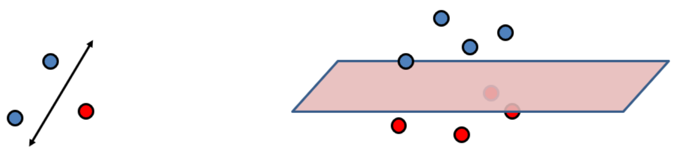
א. נקודה אחת היא בתוך הקמור של השלושה האחרים. אז במצב שבו הנקודות החיצוניות אדומות והפנימית כחולה, לא מכוסה ע"י החוק.



ב. אם הנקודות במצב כללי: נניח שיש מצב שבו הנקודות הנגדיות הן באותו צבע. נניח בשלילה שאפשר להפריד אותן על ידי עיגולים – שיש עגול שעוטף כל זוג אבל לא את הזוג השני. אז יהיו 4 איזורים זרים ואיזור חיתוך אחד. אבל זה לא אפשרי – אין מצב כזה של חיתוך מעגלים.

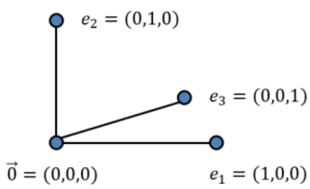


עוד הוכחה: ניקח אוסף שמנופץ ע"י מעגלים. לכל חלוקה של הקבוצה לתתי קבוצות A, B, יש שני עיגולים שמכסים את תתי הקבוצות. החיתוך בין העיגולים הוא קו ישר. כלומר עיגולים לא יכולים לנפץ יותר מאשר קו ישר, וכבר ראינו שזה 3.



טענה: לאוסף המישורים ה- $(d-1)$ מימדיים $(hyperplane)$ יש מימד VC $(d+1)$ במרחב R^d.

חסם תחתון: אוסף המישורים ה- $(d-1)$ מימדיים יכולים לנפץ $(d+1)$ נקודות במרחב R^d.

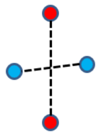


הוכחה: ניקח את הנקודות: $\vec{0}, e_1, \dots, e_d$. (נקודת האפס, ו-וקטורי הבסיס של המימד). בדוגמה אפשר לראות שאפשר להעביר מישור שיפריד כל תיוג. באופן כללי עבור n מימדים, בגלל שכל הווקטורים מאונכים, תמיד נוכל למצוא מישור שיפריד.

חסם עליון: אוסף המישורים ה- $(d-1)$ מימדיים **לא** יכולים לנפץ $(d+2)$ נקודות במרחב R^d.

הוכחה – טענת עזר:

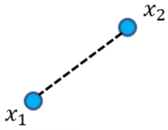
משפט ראדון: כל אוסף של $(d+2)$ נקודות במרחב R^d ניתן לחלוקה לשתי קבוצות זרות שיש חיתוך בין הקמורים שלהם.



מכאן הטענה תנבע ישירות, כי החיתוך שייך לשני הקמורים ולכן אמור להיות בשני הצבעים. וזו סתירה.

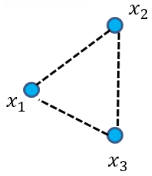
← ראינו כבר את משפט ראדון עבור שני מימדים.

נוכיח את משפט ראדון. הקדמה: קמור. נתייחס לנקודות בתור וקטורים במרחב.

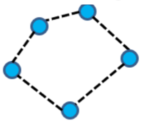


בהינתן שתי נקודות x_1, x_2, כל נקודה על הקו המחבר אותן מקיימת: $\alpha_1 x_1 + \alpha_2 x_2$, כאשר $\alpha_1 + \alpha_2 = 1, \alpha_i \geq 0$.

לדוגמה נקודת האמצע ביניהן: $\frac{x_1+x_2}{2} = \frac{1}{2}x_1 + \frac{1}{2}x_2$.



עבור 3 נקודות, כל נקודה בתוך המשולש מקיימת: $\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3$, כאשר $\alpha_1 + \alpha_2 + \alpha_3 = 1, \alpha_i \geq 0$.



עבור n נקודות, כל נקודה בתוך הקמור מקיימת: $\sum_i \alpha_i x_i$, כאשר $\sum_i \alpha_i = 1, \alpha_i \geq 0$.

נחזור להוכחה: יהיו נקודות $x_1, \dots x_{d+2}$.

נמצא קבועים שונים מאפס: $a_1, \dots a_{d+2}$ המקיימים: $\sum_{i=1}^{d+2} a_i = 0,\ \sum_{i=1}^{d+2} a_i x_i = \vec{0}$.
(סה"כ $d+1$ משוואות, עם $d+2$ נעלמים. תמיד יש פתרון לא טריוויאלי).

אם $a_i > 0$, נשים את x_i בתוך קבוצה I. אחרת, בתוך קבוצה J. הקמורים של I, J בוודאות נחתכים כי:

נגדיר: $A = \sum_{x_i \in I} a_i =^{\aleph} \sum_{x_j \in J} -a_j$ (מתקיים כי יש מספר זוגי של משתנים והסכום הוא 0).

כלומר (נחלק ב-A): $1 = \sum_{x_i \in I} \frac{a_i}{A} = \sum_{x_j \in J} -\frac{a_j}{A}$.

אז: $\sum_{x_i \in I} \frac{a_i x_i}{A} = \sum_{x_j \in J} -\frac{a_j x_j}{A}$, (כי $\sum_{i=1}^{d+2} a_i x_i = \vec{0}$).

כלומר הצירופים הלינאריים האלה הם נקודה שנמצאת בתוך שני הקמורים.

זה מוכיח את משפט ראדון, ומכאן נובעת הטענה שלנו:
אוסף המישורים ה- $(d-1)$ מימדיים **לא** יכולים לנפץ $(d+2)$ נקודות במרחב R^d.

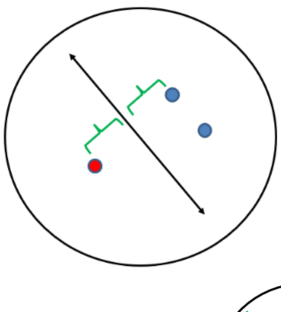
בסה"כ הוכחנו: **לאוסף המישורים ה- $(d-1)$ מימדיים ($\boldsymbol{hyperplane}$) יש מימד VC $(d+1)$ במרחב R^d.**

מכאן נובעת בעיה: אם הדאטא שלנו בהרבה מימדים, זה אומר שהמימד-VC שלנו גדול. אז החסם הכללה שאנחנו מכירים פחות יעיל. כזכור, עבור מימד-VC d קיבלנו:

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

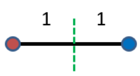
כאשר ה-d גדל, n צריך לגדול בהרבה כדי שהטעות תהיה קטנה מ-1 (ובשאיפה, קרובה לאפס). זאת בעיה במיוחד כאשר יש יותר פיצ'רים מאשר נקודות. אז באופן כללי, נרצה לחסום את המימד-VC שלנו. דרך אחת לעשות את זה היא לחסום את מספר הנקודות שחוק יכול לנפץ:

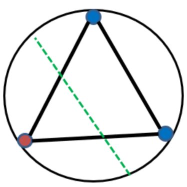
מישורים עם מרווח:



מישור עם מרווח γ יכול לנפץ רק $O(R/\gamma)^2$ נקודות בתוך רדיוס R. בלי תלות במימד.

אינטואיציה: המרווח הכי גדול מתקבל כשהנקודות במרחק שווה מהמשטח.
עבור n נקודות, נשתמש ב- $(n-1)$ מימדים. ככל שיש יותר נקודות, המרווח קטן.





עבור $R = 1$: בשתי נקודות יש לנו מרווח 1. 4 נקודות – מרווח $1/\sqrt{2}$.
באופן כללי עבור k נקודות, המרווח הוא $1/\sqrt{k/2}$.

$\gamma \leq 1/\sqrt{k/2} \Rightarrow k \leq 2/\gamma^2$.
כלומר אם קבענו את המרווח, זה חוסם את מספר הנקודות.

אלגוריתם *Winnow* (1988): נחזור לדוגמה של הבית קפה, לשאלה אם לקוח רוצה קפה או תה. ההנחה היא: רוב הפיצ'רים לא רלוונטיים, ואם אפילו אחד מהפיצ'רים החשובים מופיע, זה אומר + (קפה). בדוגמה כאן, F_3, F_4 הם הקובעים.

	F_1	F_2	F_3	F_4	label
Person	old	male	tall	resident	Coffee
X_1	1	0	1	0	+
X_2	0	1	0	1	+
X_3	1	1	0	0	-
X_4	0	1	0	0	-

אלגוריתם *winnow* לומד איזה פיצ'רים הם החשובים.

נגדיר: n הוא מספר הפיצ'רים הכולל, r הוא מספר הפיצ'רים החשובים. $F_j(X_i)$ הוא הערך של הפיצ'ר ה-j של הנקודה X_i. האלגוריתם ייתן משקל לכל פיצ'ר.

1. נגדיר $w_1 = \cdots = w_n = 1$.
2. לכל וקטור X_i:
   a. אם $\sum_{j=1}^{n} w_j F_j(X_i) \geq n$, ננחש +.
      אחרת, ננחש −.
   b. אם טעינו לגבי X_i:
      i. אם ניחשנו +, לכל j כך ש $F_j(X_i) = 1$, נגדיר $w_j = 0$.
      ii. אם ניחשנו −, לכל j כך ש $F_j(X_i) = 1$, נגדיר $w_j = 2w_j$.
3. אם הייתה טעות, נחזור ל-2. נמשיך לרוץ על כל הנקודות עד שיש מעבר שלם בלי טעויות.

טענה: יהיו לכל היותר $2r \log n$ טעויות.

בצורה מסוימת, מספר הטעויות שיכולות לקרות בדרך לקביעת החוק, מגדיר עד כמה החוק הזה פשוט (קשור ללמידה ע"י דחיסה, *out of scope* לסמסטר הזה).

נוכיח את הטענה:

יש 2 סוגים של שגיאות: סוג א – *false positive*, וסוג ב – *false negative*. כזכור, יש r פיצ'רים חשובים.

נשים לב שברגע שפיצ'ר מגיע למשקל n, הוא לא יגדל יותר – כי תמיד ננחש + עבור כל נקודה שבה הוא מופיע. כלומר, טעות ב' יכולה לקרות לכל היותר $r \log n$ פעמים. כי אחרי $r \log n$ טעויות כאלה, לכל פיצ'ר חשוב יש משקל לפחות n.

כמו כן, אחרי כל טעות מסוג א', הורדנו את סכום המשקלים בלפחות n. כי המשקל של הפיצ'רים בנקודה הזו היה לפחות n (לכן ניחשנו +), ואיפסנו את כל המשקלים האלה. כמה פעמים זה יכול לקרות? רק טעות מסוג ב' מוסיפה משקל, וכבר ראינו שהיא יכולה לקרות לכל היותר $r \log n$ פעמים. ובכל פעם, מוסיפים לכל היותר n משקל. כי סכום המשקל היה קטן מ-n, והכפלנו אותו. אז אם כל פעם מורידים לפחות n, ובסך הכל אפשר להוסיף לכל היותר $nr \log n$ משקל למערכת, אז אפשר לעשות את זה לכל היותר $r \log n$ פעמים.

בסה"כ – לכל היותר $2r \log n$ טעויות.

האלגוריתם נותן לנו מסווג לינארי (מישור, נרחיב בהרצאה הבאה). מקבלים בעצם וקטור (המשקלים), ומכפילים אותו בווקטור (אדם אחד). זה בעצם מוצא מישור שמפריד בין הנקודות. ובמהירות די יעילה. ואנחנו כבר יודעים שהממד-VC של אותו מישור יהיה קבוע לפי ה-n.