למידת מכונה

9 באוגוסט 2020

מרצה: ד"ר ליעד גוטליב

דרישות: 50% פרויקט גמר , 50% מטלות

ספרים שימושים:

- https://www.cse.huji.ac.il/~shais/UnderstandingMachineLearning/index.html
- https://cs.nyu.edu/~mohri/mlbook/
- http://neuralnetworksanddeeplearning.com/

מאגרי מידע:

- https://www.kaggle.com/
- http://yann.lecun.com/exdb/mnist/ - מאגר של ספרות
- http://archive.ics.uci.edu/ml/index.php - מאגר של פרחים
- Hope College data sets:

https://web.archive.org/web/20180422082424/http://www.math.hope.edu/swanson/statlabs/data.html

הרצאה 1

בעיות סיווג הן הבעיות הקלאסיות של למידת מכונה , למשל הפרחים.

<u>דוגמאות ללמידת מכונה:</u>

- זיהוי גבר/אישה על פי משקלים וגובה
- זיהוי ספאם במערכת אימייל.
- זיהוי פנים על בסיס מאגר תמונות
- האלגוריתם של $Netflix$ להצעות סרטים (הסיפור עם בלוק־בסטר)
- זיהוי אקו־לב בעייתי (להתקף לב למשל)
- דוגמה נבדוק האם $random$ ב $java$ אכן אקראי

משפט הופדינג:

יהיו $X_1, X_2, ..., X_n$ משתנים מקריים בלתי תלויים, ברנולים. (לכל $p \in [0,1]$). אז בעבור $X = \frac{\sum_{i=1}^{n} x_i}{n}$ מתקיים ש:

$$Pr\left((X-p) > t\right) < 2e^{-2nt^2}$$

אצלנו: ניקח מטבע הוגן אז המשתנים המקריים מקיימים ש: $P\left(X=H\right) = P\left(X=T\right) = \frac{1}{2}$, וכעת נזרוק מטבע המון פעמים, נרצה ש: בהסתברות δ־1 יקיים ש: $p = X \pm t$

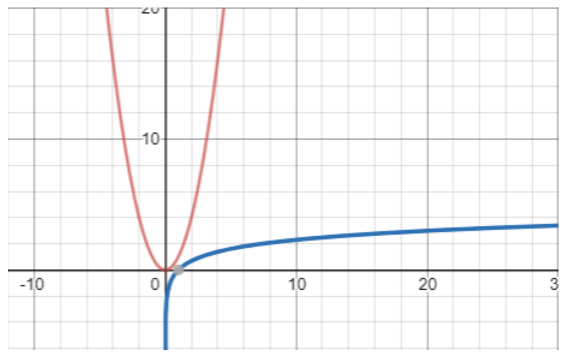
כלומר נסמן ב δ את הטעות שלנו לדוגמה $\delta = 0.01 \Leftarrow$ כלומר בהסתברות $0.99 = 1 - 0.01$ הX (המומצע) יהיה מאוד קרוב ל p עד כדי t , למשל $t = 0.1$

לסיכום: נרצה שבהסתברות 0.99 יתקיים ש: $X = 0.5 \pm 0.1$, כעת כל שנותר זה להציב:

$$\delta \geq 2e^{-2nt^2}$$
$$\Downarrow$$
$$\ln\left(\delta\right) \geq \ln\left(2e^{-2nt^2}\right) = \ln\left(2\right) - 2nt^2$$
$$\Downarrow$$
$$n \geq \frac{\ln 2 - \ln \delta}{t^2} = \frac{\ln 2 + \ln\left(\frac{1}{\delta}\right)}{t^2}$$

נשים לב שיש כאן חוסר סימטריה עבור התלות של n ב δ ולתלות ב t :

- התלות δ היא למעשה $-\ln\left(\delta\right)$ וזו פונקציה מונוטונית מאוד איטית $\Leftarrow$ כלומר עבור δ שקטן בפי מאה , צריך בסה"כ להגדיל את n ב $\ln(100)$
- לעומת זאת התלות ב t^2 ־ היא פונקציה מונוטונית מאוד מהירה $\Leftarrow$ עבור t מאוד קטן = כמה קרוב אני לאמת $\Leftarrow$ צריך הרבה מאוד דגימות.



נראה זאת בקוד, נבחר $\delta = 0.001$ ו$t = 0.01$

```
public class Coins {

    public static void main(String [] args) {
        double delta = .001; // probability of failure
        double t = .01;  // closeness to the true bias
        int n = (int)(Math.log(2/delta) / (t*t)); // sample size
        System.out.println(n);
        int sum = 0;
        for(int i=0; i<n; i++) {
            double flip = Math.random();
            if(flip < .5) sum++;
        }
        System.out.println("With probability " + (1-delta) + " coin bias is within "
        + t + " of " + sum/(double)n);
    }

}
```

וקיבלנו עבור הרצה אחת :

With probability 0.999 coin bias is within 0.01 of 0.500111

המשך דוגמאות־ מהמצגת:

https://www.cs.bgu.ac.il/~inabd171/wiki.files/lecture2_handouts.pdf

בעית המלצר החדש:

המלצר צריך בתוך שבוע ללמוד מה העדפות של כל לקוח, והתשלום הוא בהתאם לידע שהם מפגינים.

- זוהי בעיית תיוג קלאסית
- המלצר/הלומד צריך לצור חוק שדרכו יוכל לחזות את העדפות הלקוחות
- אצלנו:

- המלצר= הלומד
- הלקוחות = הדוגמאות
- המשקאות הנחברים = התיוגים
- התשלום = הצלחת הניבוי

- יש שני שלבים:
  - $training - phase$ שלב הלמידה
  - $test - phase$: שלב בו בוחנים את האלגוריתם שהצענו

- הצעה:
  - לרשום מה כל לקוח אוהב ולהציע לו את ההעדפה שלו בשבוע השני.
    - הבעיה: מה עושים עם לקוחות חדשים?
  - בשיעור הבא נראה: שחוק טוב הוא חוק פשוט שתופס הכל.
  - נדגים אצלנו חוק טוב: אם הצלחנו להבין שכל הגברים מזמינים קפה, וכל הנשים מזמינות תה. אם זה החוק שהצלחנו ללמוד, אז בקלות נוכל להכיל אותו לכל העולם.

- פרמול:
  - $X =$ the set of all possible examples
  - $Y =$ the set of all possible labels
  - A training sample: S = $((x_1, y_1), \ldots, (x_m, y_m))$
  - A learning algorithm is any algorithm that has:
    - Input: A training sample S
    - Output: A prediction rule $\hat{h}_s : X \to Y$

דוגמת $freinds$

- X = $\{"Monica", "Phoebe", "Ross", "Joey", "Chandler"\}$
- Y = $\{"Juice", "Tea", "Coffee"\}$
- D is a distribution over $X \times Y$.
- D defines a probability for every pair $(x, y) \in X \times Y$.
- This is denoted $P_{(X,Y)\sim D}\left[(X,Y) = (x,y)\right], or\ D((x,y))$.
- A possible D:

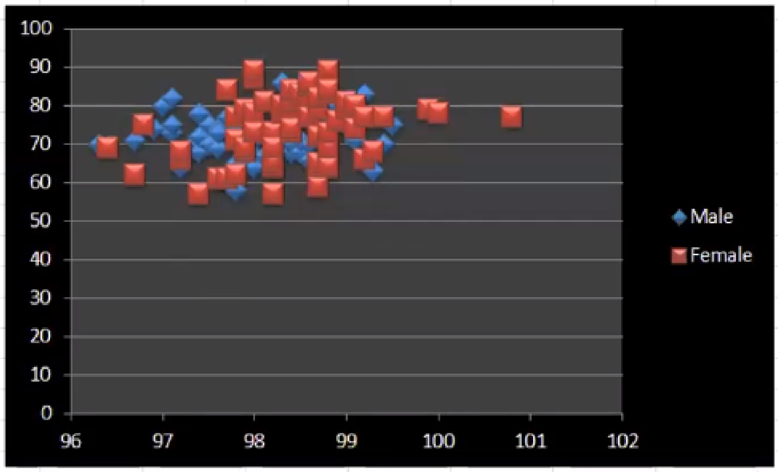
נניח ש:

	Juice	Tea	Coffee
Monica	0	20%	0
Phoebe	25%	10%	0
Ross	0	0	20%
Joey	0	0	0
Chandler	25%	0	0

- הערה: סוכמים ל1 את **כל** הטבלה, ולכן $Phoebe$ יכולה להזמין לפעמים מיץ ולפעמים תה.
  - ניתן להסתכל על זה כאילו יש שתי נקודות $P\left(Phoebe,Tea\right)=0.10$ וגם $P\left(Phoebe,Tea\right)=0.25$

דוגמת טמפטורת גוף לפי מגדר וקצב לב

https://web.archive.org/web/20180422082424/http://www.math.hope.edu/swanson/statlabs/data.html



ברור שיש כאן איזשהי התפלגות שונה בין גברים לנשים, גם אם לא ברור בדיוק מהי.

דוגמת הסופרבול

https://www.youtube.com/watch?v=owGykVbfgUE

הם יצרו את הפרוסמת על בסיס המון נתונים שהם למדו והם הבינו שהמון נשים צופות בגמר, ושהמון נשים קונות את השמפו לגבר שלהם.

חזרה על הסתברות:

- **התפלגות:** דוגמאות:
  - ההתפלגות של מטבע הוגן היא $f\left\{H,T\right\}=\left\{\frac{1}{2},\frac{1}{2}\right\}$
  - ההתפלגות של קוביה הוגנת היא: $f\left\{[6]\right\}=\left\{\frac{1}{6},....,\frac{1}{6}\right\}$
- **גבולות:**
  - $Pr\left(X\vee Y\right)\leq Pr\left(X\right)+Pr\left(Y\right)$
  - $Pr\left(X\wedge Y\right)\leq\frac{Pr(x)+Pr(Y)}{2}$
  - ניתן לראות בעזרת דיאגרמות ון
- **תוחלת:**
  - $E\left[X\right]=\sum\limits_{x}x\cdot Pr\left(X=x\right)$
  - לדוגמה בקוביה $\sum\limits_{i\in[6]}i\cdot\frac{1}{6}=3.5$

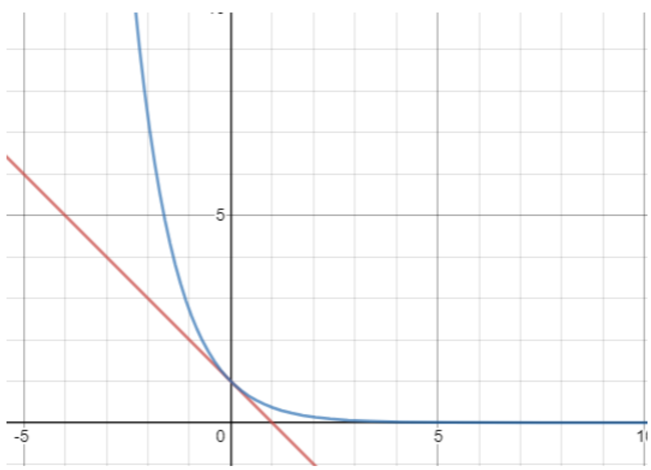
הרצאה 2

תזכורת:

- בשיעור שעבר הראינו את בעיות ה $callisfication$ / סיווג
- פורמלית על סמך מדגם S -(X כל העולם , והתפלגות D)
- צרך למצוא חוק ($hypothesis$)) שיתאים (=שיהיה טוב) לכל דגימה בהתפלגות D מהעולם X
  - דוגמה: נניח שרוצים לחזות מי יתקבל למדעי המחשב - ונניח שהתנאי קבלה הוא רק על פי ציון פסיכומטרי (מעל רף מסוים)
  - אפשרות לחוק היא לקחת את המינימום על מי שהתקבל ולהחליט מעט פחות $\Leftarrow$ יש לנו טעות רק בכיוון אחד
  - אפשרות נוספת היא לקחת את האמצע בין המינימום על המתקבלים והמקסימום על אלו שלא, ובכך יהיה נמנע מטעות בשתי הצדדים (המספר שנקבל הוא הערכה)

טענה : $1-x\leq e^{-x}$

בגרף ($desmos$) :



הוכחה:

- נסמן $\left\{ \begin{array}{ll} f(x)=1-x & g(x)=e^{-x} \\ \Rightarrow f'(x)=-1 & g'(x)=-e^{-x} \end{array} \right\}$
- עבור $x=0$ $\Leftarrow$ $f(x)=1=g(x)$

עבור $x<0$:

$$g'(x)<-1=f'(x)$$

$$\frac{f'}{g'}=\frac{\cancel{-}1}{\cancel{-}e^{-x}}\underset{x<0}{=}\frac{1}{n}\rightarrow 0$$ (לופיטל)

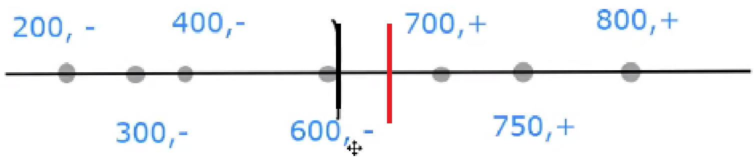
- עבור $x>0$:

$$g'(x) > -1 = f'(x)$$

$$\frac{f'}{g'} = \frac{\cancel{-}1}{\cancel{-}e^{-x}} \underset{x>0}{=} \frac{1}{\frac{1}{n}} \to n$$ (לופיטל)

- סה"כ תמיד $f \leq g$

נחזור לדוגמת הפסיכומטרי שלנו:



נסמן כל אדם כזוג : ציון, התקבל/לא התקבל

ונניח שהקו השחור הוא הסוף על פי הלמידה שלנו - n, ושהקו האדום זוהי האמת - $truth$, את הטווח בינהם נסמן בε, יתקיים ש:

וכעת נוכל לשאול את השאלה, מה ההסתברות שהטעות היא ε , כלומר שמבין n טעינו בε (למשל 0.01) אנשים. אז עבור $\varepsilon = 0.01$ אם בחרנו אדם אחד, ההסתברות שנפספס את מרחב הטעות היא $(1-\varepsilon)$ ועבור n אנשים (בכל n הפעמים נפספס את ε), ומהלמה נקבל :

$$(1-\varepsilon)^n \leq \left(e^{-\varepsilon}\right)^n = e^{-\varepsilon n}$$

נשים לב שהפונקציה דועכת מאוד מהר, ושואפת לאפס:

כעת נסמן:

$$e^{\varepsilon n} \leq \delta$$
$$\ln\left(e^{-\varepsilon n}\right) = -\varepsilon n \leq \ln(\delta)$$
$$n \geq \frac{-\ln(\delta)}{\varepsilon} = \frac{\ln\left(\frac{1}{\delta}\right)}{\varepsilon}$$

וכעת נשים לב: δ קטן = יש לנו חוק טוב, כלומר הסתברות לטעות מאוד נמוכה, והקטנת הδ משפיעה באופן מינורי על הגודל של n (בגלל הln)

ε = על איזה אחוז מהאנשים אני טועה

ללמידה הזו קוראים

$$PAC = \underbrace{\text{Probably}}_{\delta}\ \underbrace{\text{approximetaly}}_{\varepsilon}\ \text{correct}$$

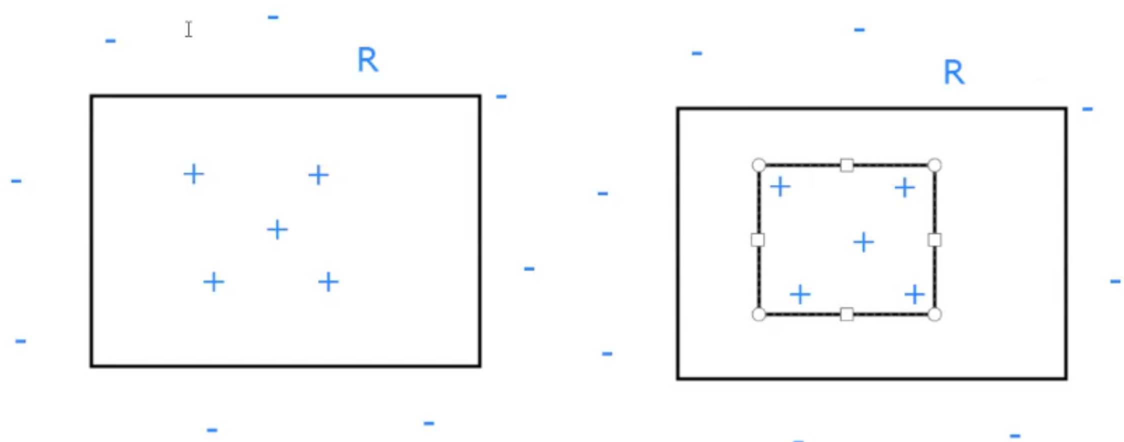
דרך נוספת להסביר:

δ היא: הטענה שלי צודקת ב $1-\varepsilon$ מהאנשים, ε הוא הסיכוי לטעות בדגימה.

<u>דוגמה נוספת:</u>

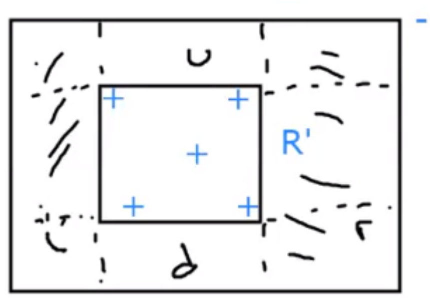
יש לנו גרף שמתאר גבהים ומשקלים של תינוקות , כאשר $+$ הוא תינוק "תקין" ו $-$ תינוק בטווח בעייתי.

נניח גם שמשרד הבריאות, פרסם מלבן (R) שאומר מי הם התינוקות שנמצאים בטווח התקין, ואנחנו רוצים לנחש/ללמוד אותו



אז על בסיס ה+ים שאנחנו מכירים ציירנו מלבן משלנו (R'), וכעת אנחנו רוצים לשאת מה ההסתברות שטעינו, מהו אחוז מהתינוקות שיהיה בטווח הטעות?

נסמן את אזור הטעות ε , וכעת נצייר 4 מלבנים חופפים:



אז ההסתברות לכל מלבן קטן היא $\frac{\varepsilon}{4}$

וכעת נוכל לשאול, מה ההסתברות שהמלבן שלי טועה ביותר מε מהתינוקות

$$Pr\left(error(R')>\varepsilon\right)\leq Pr\left(\text{S miss } \textbf{up} \text{ or } \textbf{down} \text{ or } \textbf{left} \text{ or } \textbf{rigth}\right)$$

$$\leq Pr\left(\sum_{s\in\text{sides}}\text{S miss s}\right)\leq 4\left(1-\frac{\varepsilon}{4}\right)^{n}\leq 4e^{\frac{-\varepsilon n}{4}}$$

וכעת אם נסמן

$$4e^{\frac{-\varepsilon n}{4}}<\delta \iff n>\frac{4\left(\ln\left(\frac{1}{\delta}\right)+\ln 4\right)}{\varepsilon}$$

נקבל שוב ש:

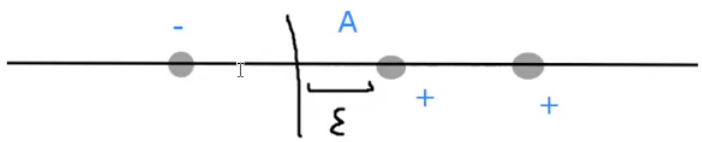
- טענה: $hypothesis$ נכונה ב $1-\varepsilon$ מהתינוקות
- וטענה זו נכונה בהסתברות $1-\delta$

הראנו עד כה טכניקה של קו וריבוע, אך בעיגול זה כבר לא עובד.

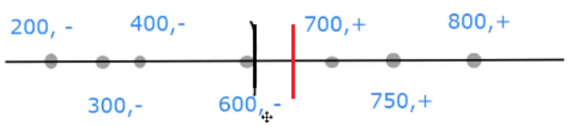
נציין גם שבאנליזה של שיטת המלבנים - הגענו לאזור של טעות על פי ה R (האמת), ולא על פי R' , אנחנו לא יכולים להגדיר את האזור לפי הניחוש ואז לשאול על הניחוש, היות ויש אינסוף ניחושים. במילים אחרות, לא ניתן להגיד:

$$Pr\left[\text{hypothesis error} > \varepsilon\right] \leq Pr\left[\text{S missed region A}\right]$$

בציור זה כמו לצייר קו ואז להגדיר ε= טווח טעות:



ומה שעשינו בדוגמה עם הפיסכומטרי זה היה לשאול מה ההפרש בין הניחוש לאמת, וזה מה שמגדיר את ε :



כעת נראה מדוע חוק טוב, הוא טוב לכולם העולם.

נניח ויש ל $h \in H$ חוקים בעולם , אז א:

- נדגום S כך ש $S=|n|$, ונבחר את החוק הכי טוב $h \in H$, ואכן זהו חוק עיקבי, מה כעת ניתן לומר על כל העולם?
- ניתן לומר שהחוק הזה יהיה עיקבי לכל העולם, הסבר:

$$Pr\left(\text{exist h in H : h consistent on } S \text{ and error(h)}{>}\varepsilon\right)$$

$$\begin{aligned}
&= Pr\left[(h_1 \in H \text{ consistent on } S \text{ and error}(h_1){>}\varepsilon) \vee \right.\\
&\quad \left. h_2 \in H \text{ consistent on } S \text{ and error}(h_2){>}\varepsilon \vee \ldots.\right]\\
&\leq \sum_{i=1}^{|H|} Pr\left(h_i \in H \text{ consistent on } S \text{ and error}(h_i){>}\varepsilon\right)\\
&\leq \sum_{i=1}^{|H|} Pr\left(h_i \in H \text{ consistent on } S \mid \text{error}(h_i){>}\varepsilon\right)
\end{aligned}$$

$$|H|\left(1-\varepsilon\right)^n \leq |H|\, e^{-\varepsilon n} \leq \delta$$

$$\Downarrow$$

$$|H|\, e^{-\varepsilon n} \leq \delta \iff n \geq \frac{\ln\left(\frac{1}{\delta}\right)+\ln|H|}{\varepsilon}$$

- ובכך הוכחנו שאם יש לנו חוק טוב (שבחרנו מתוך מספר סופי של חוקים אפשריים) הוא יהיה טוב לכל העולם - מה שאומר שיש לנו אפשרות ללמוד
  - הבעיה: מי אמר שיש מספר סופי של חוקים, הרי אפילו אם נקח עגולים יש להם אינסוף אפשרויות למרכז מעגל ולרדיוס?
  - תשובה: נוכיח משפט בשבוע הבא שהרעיון שלו שלמרות שיש אינסוף עיגולים/קווים אפשריים, החוקים הרלוונטים לנו הם מספר סופי.

משפט מרקוב

$$Pr\left(X>c\right)<\frac{E(x)}{c}$$

משפט $\left(1+x\right)ln\left(1+x\right)>x-\frac{x^{2}}{3}$

<u>הוכחה:</u>

מתקיים ש (ניתן להראות על ידי טיילור):

$$\ln\left(1+x\right)=x-\frac{x^{2}}{2}+\frac{x^{3}}{3}-...>x-\frac{x^{2}}{2}+\frac{x^{3}}{3}$$

ולכן:

$$\left(1+x\right)\ln\left(1+x\right)>x-\frac{x^{2}}{2}+\frac{x^{3}}{3}+x^{2}-\frac{x^{3}}{2}+\frac{x^{4}}{3}$$

$$>x+\frac{x^{2}}{2}-\frac{x^{3}}{6}...>x+\frac{x^{2}}{3}$$

משפט צ'רנוף

יהיו $x_{1},...,x_{n}$ משתנים מקרים בלתי תלויים, ברנוליים. ויהיה $X=\sum_{i=1}^{n}x_{i}$ ותהיה $E\left(x_{i}\right)=p$, ויהיה $m=E\left(X\right)=np$

אז:

$$Pr\left(X>\left(1+\delta\right)m\right)\leq\frac{e^{-m\delta^{2}}}{3}$$

הוכחה:

$$Pr\left(X>\left(1+\delta\right)m\right)\overset{1}{=}Pr\left(e^{tx}>e^{t(1+\delta)m}\right)\overset{2}{\leq}\frac{E\left(e^{tx}\right)}{e^{t(1+\delta)m}}$$

1. העלנו את הביטוי לe^{t} .2. מרקוב.

$$E\left(e^{tx}\right)\overset{1}{=}E\left(e^{t\sum_{i}x_{i}}\right)\overset{2}{=}E\left(\prod_{i=1}^{n}e^{tx_{i}}\right)\overset{3}{=}\prod_{i=1}^{n}E\left(e^{tx_{i}}\right)=$$

$$\overset{4}{=}\prod_{i=1}^{n}pe^{t}+\left(1-p\right)\cdot1\overset{5}{=}\prod_{i=1}^{n}1+p\left(e^{t}-1\right)\overset{6}{\leq}\prod_{i=1}^{n}e^{\left(e^{t}-1\right)}$$

$$=e^{m\left(e^{t}-1\right)}$$

1. הגדרת X . 2 חוקי חזקות. 3.נסמן $Y_i = e^{tx_i}$ אז Y_i בת"ל. 4. הגדרת תוחלת. 5. סדרנו. 6. $(1+x) \le e^{x} \Leftrightarrow (1-x) \le e^{-x}$

$$\overset{7}{\le} \frac{E\left(e^{tx}\right)}{e^{t(1+\delta)m}} \overset{8}{\le} \frac{e^{m\left(e^{t}-1\right)}}{e^{(1+\delta)tm}} \overset{9}{=} e^{m\left[\left(e^{t}-1\right)-(1+\delta)t\right]}$$

$$\overset{10}{=} e^{m\left[\left(e^{\ln(1+\delta)}-1\right)-(1+\delta)\ln(1+\delta)\right]} \overset{9,11}{\le} e^{m\left[\delta-\delta+\frac{\delta^2}{3}\right]} = e^{\frac{m\delta^2}{3}}$$

7. מה שהגענו ממרקוב. 8. נציב את את שהוכחנו ב1־6 . 9. חוקי חזקות . 10 נציב $t = \ln(1+\delta)$ .11 מהלמה $(1+\delta)\, ln\,(1+\delta) > \delta - \frac{\delta^2}{3}$

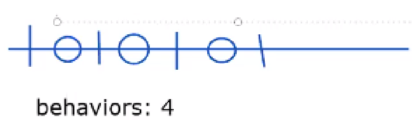
הרצאה 3

תזכורת:

בשיעור שעבר הזכרנו שמספר הקווים שאנחנו יכולים למתוח הוא מוגבל, כי יש המון קווים שמחזירים את אותו תיוג. והיום נוכיח שאכן מספר הקווים (= תוצאות שונות) הוא סופי.

נתחיל להסביר את הרעיון דרך דוגמת האינטרוול:

נניח שיש 3 נקודות, עד הקו זה ־ ומהקו זה +, אז יש בסה"כ 4 מקומות שאפשר לשים את הקו:



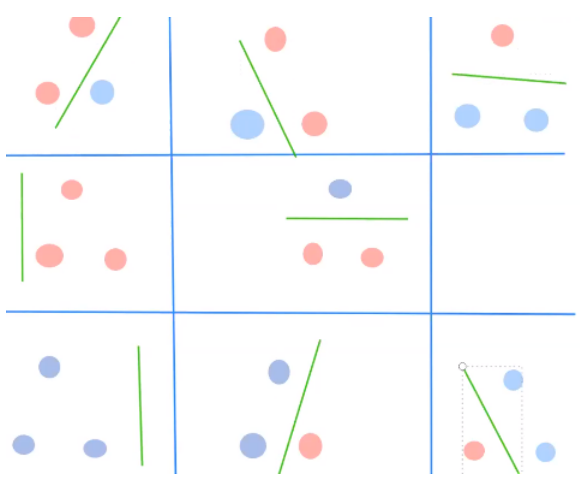
ואם היינו רוצים כל האפשרויות על 3 נקודות אז זה $2^3 = 8$,

חסרון שקיים באינטרוול הוא שהמצב הבא בלתי אפשרי (בהנחה והאינטרוול משמאל לימין)

כי ניתן רק להגדיר קו שממנו והלאה הכל + , ואי אפשר "לחזור ל -"

לתופעה זו קוראים $shattering$ (לנפץ/לפצל): בהנתן מקבץ נקודות מגודל k יש סה"כ 2^k אפשרויות תיוג $(behavoirs)$

נראה דוגמה של $k = 3 \Leftarrow$ יש $2^k = 2^3 = 8$ אפשרויות:



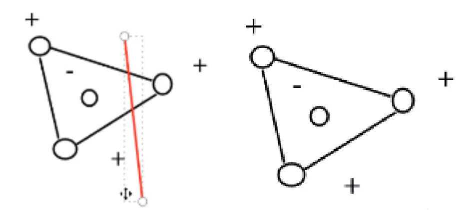
כעת נוכיח של $k = 4$, זה לא אפשרי ומכאן נגיע למסקנה ש$VC - dimension$ של קווים הוא לכל היותר 3 .

טענה: כל אוסף אינסופי של קווים לא יכול לתייג (לפצל/לנפץ/ $shatter$) קבוצה של 4 קווים

הוכחה:

נפצל למקרים:

מקרה ראשון: נקודה אחת בתוך 3 השאר, למשל:

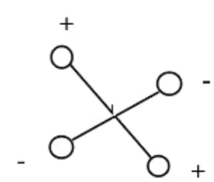


תיוג זה אינו אפשרי כי אם ננסה להעביר קו ישר כלשהו נגרום לאחד מה+'ים להיות "בחוץ", כמו בציור עם הקו האדום.

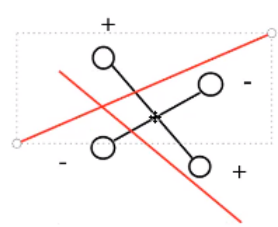
קצת יותר פורמלי:

- מתיחת קו פרושה שצד אחד + וצד שני -
- כאשר מנסים למתוח קו אדום אם:
  - לא חתכנו אף קו שחור - אז התיוג של הפנימי זהה לחיצוניים
  - אף חתכנו בין פנימי לחיצוני בפרט חתכנו שני קווים שחורים, ולכן או שקודקוד חיצוני קיבל - או שהפנימי קיבל +
- בכל מקרה , קיבלנו סתירה.

מקרה שני: הנקודות בסדר כללי כשלהו, אז חייב להתקיים שיש חלוקה:

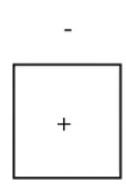


וכל ניסיון למתוח "קו אדום", לא יצליח:

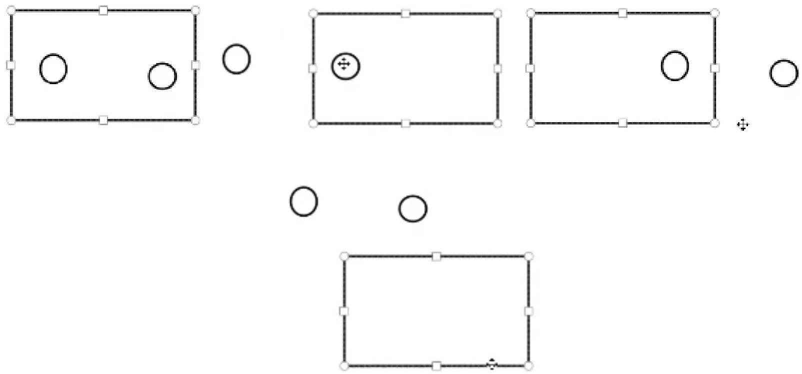


הבהרה : $shatter$ (לנפץ) - זה לתת את כל האפשרויות, דוגמה :

נניח שיש לנו מלבן ושתי נקודות, המלבן מגדיר חלוקה (בחוץ - , בפנים +):

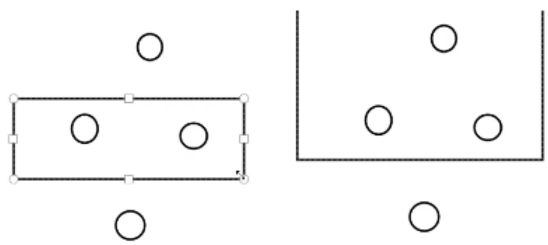


אז עבור 2 נקודות, המלבן באמת מנפץ את כל האפשרויות:



עבור 3 נקודות זה גם עובד

כעת עולה השאלה, האם קיים אוסף של 4 נקודות שמלבן **מאונך לצירים** יכול לנפץ? תשובה - כן.

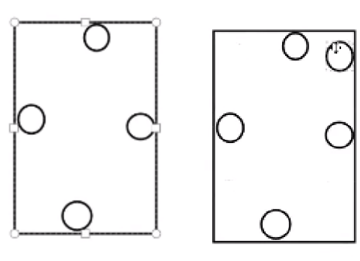


(סה"כ יש 16 אפשרויות)

נמשיך - האם קיים אוסף של 5 נקודות שמלבן יכול לנפץ? תשובה - לא

נראה זאת:

- יהיו 4 נקודות מתוך החמישה כך שלקחנו את הנקודה הכי קיצונית מכל כיוון - הכי למעלה, הכי למטה, הכי ימינה, הכי שמאלה $(min, max\,(x, y))$
- נבנה מלבן הכי צמוד שאנחנו יכולים סביבם:



- ואז ננסה להבין, איפה הנקודה החמישית?
- מהבניה היא חייבת להיות איפשהו בתוך המלבן $\Leftarrow$ תיוג בו הנקודה החמישית היא $(-)$ והרביעיה $(+)$ לא אפשרי

מסקנה: $VC - dimension$ של אינסוף מלבנים הוא 4 .

Sauer–Shelah lemma

https://moodlearn.ariel.ac.il/mod/url/view.php?id=1116793&redirect=1

רקע: את המשפט המציאו סאוור ושלח בנפרד, ולכן קוראים כך למשפט. היו שני רוסים $vapnic$ ו $cherovenenski$ שהמציאו אותו שנה קודם, אבל לא פרסמו באנגלית (ונשכחו).

הרעיון: עד כה הראנו שעל עולם H בגודל m יש 2^m תיוגים, המשפט הבא יראה שבפועל יש רק m^d תיוגים כאשר $d = vc-dim\left(H\right)$

(3) לפני, תזכורת מדיסקרטית (הבינום של ניוטון):

$$(1+y)^n = \sum_{k=0}^{n} \binom{n}{k} y^k$$

lemma: let P(s) all possible label set assigned to sample S using H than: $P\left(S\right) \le \sum\limits_{i=0}^{d} \binom{m}{i} \backsim O\left(m^d\right)$

while $d = vc - dim\left(H\right)$

proof:

ראשית, נראה ש$\sum\limits_{i=0}^{d} \binom{m}{i} \backsim O\left(m^d\right)$ **:**

$$\sum_{i=0}^{d} \binom{m}{i} \overset{1}{\le} \sum_{i=0}^{m} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i} = \left(\frac{m}{d}\right)^d \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^i \overset{3}{=} \left(\frac{m}{d}\right)^d \left(1+\frac{d}{m}\right)^m$$

$$\overset{4}{\le} \left(\frac{m}{d}\right)^d e^d = \left(\frac{em}{d}\right)^d \overset{5}{\sim} O\left(m^d\right)$$

1. $m \ge d \ge i$ ⇐ הביטוי גדול מ1. 2. d קבוע ⇐ ניתן להוציאו. (3) תזכורת. 5 . ללא הקבועים

נותר להראות ש $P\left(S\right) \le \sum\limits_{i=0}^{d} \binom{m}{i}$

ראשית עוד תזכורות מדיסקרטית (זהות פסקל):

$$\begin{matrix} \binom{m}{d} = \binom{m-1}{d} + \binom{m-1}{d-1} \\ \binom{m}{d} = 0, if\ d<0 \vee m<d \end{matrix} \qquad (\#)$$

נראה באינדוקציה:

בסיס:

מהתזכורת כל מקרה ש $m<d$ שווה לאפס, ולכן מקרה הבסיס הראשון שמעניין $m=d$ ⇐ $\sum\limits_{i=0}^{d} \binom{m}{i} = \sum\limits_{i=0}^{m} \binom{m}{i}$

אם נספר את הסיפור של הביטוי, משמעו מה הן כל תתי הקבוצות האפשריות מתוך קבוצה בגודל m $\iff$ ס"ך האפשרויות הוא 2^m (כ"א יכול להיות או לא להיות בקבוצה)

צעד: נניח ל $m<d$ **ונוכיח ל** m **:**

צ"ל:

$$P\left(s\right) = \left|H\right| \le \sum_{i=0}^{d} \binom{m}{i}$$

שלב ראשון: <u>$|H|=|H_1|+|H_2|$</u>

נסמן את סט החוקים ב H וכל חוק ב H יהיה חוקים ב h_i ונקודות ב x_i , ונניח שהסיווג של החוק h_i הוא ל0 ו 1 , לדוגמה:

$$\begin{array}{c|ccccc}
\multicolumn{6}{c}{\boldsymbol{H}}\\
 & x_1 & x_2 & x_3 & x_4 & x_5\\
\hline
h_1 & 0 & 1 & 1 & 0 & 0\\
h_2 & 0 & 1 & 1 & 0 & 1\\
h_3 & 0 & 1 & 1 & 1 & 0\\
h_4 & 1 & 0 & 0 & 1 & 0\\
h_5 & 1 & 0 & 0 & 1 & 1\\
h_6 & 1 & 1 & 0 & 0 & 0
\end{array}
\Rightarrow
\begin{array}{c|ccccc}
\multicolumn{6}{c}{\boldsymbol{H_1}}\\
 & x_1 & x_2 & x_3 & x_4 & x_5\\
\hline
h_1 & 0 & 1 & 1 & 0 & 0\\
 & & & & & \\
h_3 & 0 & 1 & 1 & 1 & 0\\
 & & & & & \\
h_5 & 1 & 0 & 0 & 1 & 1\\
h_6 & 1 & 1 & 0 & 0 & 0
\end{array}
,
\begin{array}{c|ccccc}
\multicolumn{6}{c}{\boldsymbol{H_2}}\\
 & x_1 & x_2 & x_3 & x_4 & x_5\\
\hline
 & & & & & \\
h_2 & 0 & 1 & 1 & 0 & 1\\
 & & & & & \\
 & & & & & \\
h_5 & 1 & 0 & 0 & 1 & 1\\
 & & & & &
\end{array}$$

נרצה לפצל את הסט H לשתי קבוצות, לפי הרעיון הבא: נתבונן ב x_5 , כל שני חוקים שמסכימים על קוד ה x_i חוץ מx_5 נפצל לשתי קבוצות, ושאר החוקים (בדוגמה h_3,h_6) נכניס באופן דיפולטיבי לH_1

- $\Leftarrow$ מתקיים ש $|H|=|H_1|+|H_2|$,

שלב 2: <u>נראה ש: עבור $|H_1|,|H_2|$ יש $m-1$ נקודות</u>

- נספור את מספר החוקים ב H_1 . נשים לב שבH_1 כל החוקים שונים:
- $(*)$ נתבונן ב H ־ אם היינו זורקים את x_5 , אז מתקיים ש החוקים h_1 וh_2 זהים, כנ"ל ל h_4 ו h_5
- מהבניה של H_1,H_2 ו $(*)$,נוכל להסיק שכל החוקים ב H_1 שונים זה מזה, כנ"ל ל H_2 .
- לכן ניתן להוריד מH_1,H_2 את x_5
- כעת אם נכליל את מספר הנקדות הנקודות ל m , נקבל שבחלוקה לH_1,H_2 כל סט חוקים עובד עם $m-1$

שלב 3: <u>טענה: $VC-dim\,(H_2)=VC-dim\,(H)-1$</u>

- נניח שיש לנו אוסף נקודות שאנחנו יכולים ל"נפץ" ב H_2 (לתת כל תיוג אפשרויות)
- כעת נוסיף את x_5 , אז אוסף החוקים האלה לא יצליח להכליל את x_5 ־ כי הם בונים על זה שx_5 לא קיים $\Leftrightarrow$ x_5 תמיד 1
- כעת אם נוסיף את כל החוקים מH שיודעים להתמודד עם $x_5=0$ $\Leftarrow$ נוכל לנפץ את x_5
- לכן ה$VC-dim\,(H_2)+1=VC-dim\,(H)$, כנדרש.

שלב 4: <u>הנחת האינדוקציה</u>

- ראינו ש:

$$P\,(S)=|H|=|H_1|+|H_2|$$

- מהנחת האינדוקציה, והטענה ל $VC-dim\left(H_{2}\right)$:

$$\leq\sum_{i=0}^{d}\binom{m-1}{i}+\sum_{i=0}^{\boldsymbol{d-1}}\binom{m-1}{i}\overset{1}{=}\sum_{i=0}^{d}\binom{m-1}{i}+\sum_{i=0}^{\boldsymbol{d}}\binom{m-1}{i-1}=\sum_{i=0}^{d}\binom{m-1}{i}+\binom{m-1}{i-1}\overset{2}{=}\sum_{i=0}^{d}\binom{m}{i}$$

1. הוספנו 1 ל d וחיסרנו 1 מ i $\Leftarrow$ הוספנו מקרה של $i=-1$ ששוה ל0 מהתזכורת $(\#)$. 2. גם מהתזכורת $(\#)$ (זהות פסקל)

$\square$

למסקנה קיבלנו מספר חוקים הרבה יותר קטן מ"אינסוף הקווים" שציירנו. אם נזכר במשפט משיעור שעבר:

For a set H :

$$Pr\left(\text{exsits h}\in H\text{ has emprical error 0 but true error }>\varepsilon\ \right)<\left|H\right|e^{-m\varepsilon}$$

נוכל לשפר אותו ולקבל ש:

$$<\left|H\right|e^{-m\varepsilon}\leq\left|m^{d}\right|e^{-m\varepsilon}=e^{d\ln(m)-m\varepsilon}$$

וכעת ננסה להעריך את גודל המדגם שלנו:

$$\begin{gathered}e^{d\ln(m)-m\varepsilon}<\delta\\ d\ln\left(m\right)-m\varepsilon<\ln\left(d\right)\\ m\varepsilon>\ln\left(\tfrac{1}{\delta}\right)+d\ln\left(m\right)\\ m>\tfrac{\ln\left(\frac{1}{\delta}\right)+d\ln(m)}{\varepsilon}\end{gathered}$$

נקבל שהm צריך להיות קצת יותר גדול מ $VC-dim$ ו $\frac{\ln\left(\frac{1}{\delta}\right)}{\varepsilon}$, ושאנחנו מקבלים מספר מוגבל של חוקים.

לסיכום, בחלק זה של הקורס הראנו שהלמידה אפשרית, ואפשר ללמוד עם סט חוק יחסית פשוט , ועם מספר מוגבל.

הרצאה 4

בשיעור שעבר:

- הזכרנו את המונח של $Shattring$ = לנפץ, ואת המשפט:
  - סט של פונקציות $m\ Shattering$ נק' $\iff$ ישנה פונקציה לכל 2^{m}
  - הראנו שלקו יש $VC-dim=3$
    - הראנו שאכן 3 (חסם תחתון) בציור
    - והראנו שאוסף של 4 אפשרויות , הוא בלתי אפשרי
  - נעיר שההוכחה מאוד דומה למישור

הכללת הקו למקרה כללי:

הגדרה: $hyperplane$ (על־מישור) הוא מרחב ממימד $n-1$ בתוך מרחב ממימד n

משפט $VC-dim$ של $hyperplane$ במימד d הוא $d+1$

הוכחה:

$LowerBound$: נסמן את הנקודות $0,e_1,e_2....e_d$

דוגמאות: ב$2d$ זה יראה כך: $(0,0),(0,1)\ (1,0)$ ובשלושה: $(0,0,0),(0,0,1)\ (1,0,0)\ (0,1,0)$

הוכחה - נפצל למקרים:

- אם כל הנקודות $(+)$ או $(-)$ נציב את ה$hyperplane$ ב $x=-1$ במקביל לציר ה y
- אם נקודה אחת $(+)$, והאחרות $(-)$ (או הפוך) , ה $hyperplane$ יעבור דרך הנקודה היוצאת דופן בניצב ל ל $(0,e_i)$
- באופן כללי: נבחר את ה$hyperplane$שיעבור דרך $(-\backslash+)$ים ולא דרך ה $(-\backslash+)$

$UpperBound$: אצלנו צ"ל להראות שלא קיים סט עם $d+1$ שניתן לנפץ

נשתמש במשפט של רדון:

any set of d+2 points in d-dimensional space can be partitioned into two sets whose convex hull intersects

הסבר:

לכל סט של $d+2$ נקדות במרחב d קיימת חלוקה לשתי קבוצות כך שהקמור (הקו המחבר את הנקודות בקבוצה) של האחת יחתוך את הקמור של השני

דוגמאות מהקו:

הנקודות נחשבות כמי שהקמור של אחת בתוך השני

הוכחה

- נחלק את הנקודות לשני סטים A,B לפי המשפט של רדון (הקמורים נוגעים אחד בשני)
- לכן חוק לינארי שינסה לתייג את A כ $(+)$, יהיה חייב לתת $(+)$ גם לחיתוך
- כנ"ל ל B נסיון לתת ל B $(-)$ יהיה חייב לתת $(-)$ גם לחיתוך
- וזו סתירה

משפט:

The convex k-gons in 2d (inside is +) has VC-dimension 2k+1

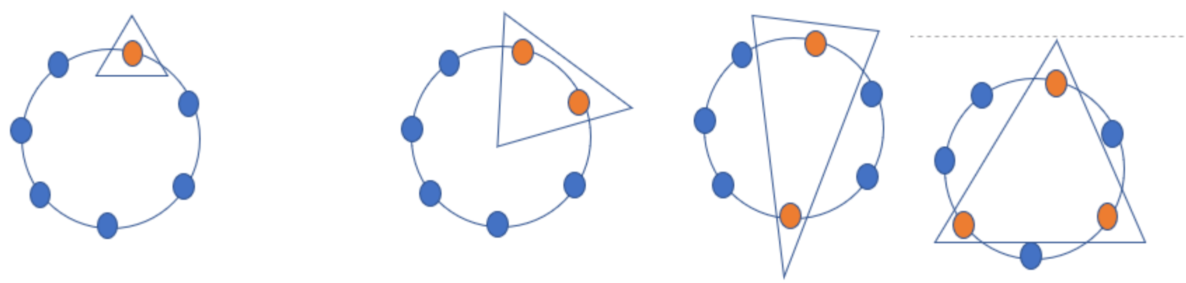
$3-gon, 4-gon$:

הוכחה:

ניקח את דוגמת המשולש $\Leftarrow 3-gon \Leftarrow Vc-dim = 2\cdot 3+1 = 7$

$LowerBound$

נצייר:



וכן הלאה..

$UpperBound$:

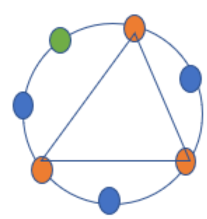
הגדרה:

אם נקודה אחת בקמור של האחרות היא לא יכולה $(-)$ ושהאחרות יהיו $(+)$

לכן נניח , שהם בצורה כללית

צריך להראות שכל הנקודות הן על העיגול

ואז עבור המקרה של $+,-,+$ צריך להראות שיש מקסימום למספר האזורים שאתה יכול לחלק ולמשל במשולש שמחלק לשלושה כמו בדוגמה:



יכול להסתדר עם 6 נקודות (3 לאדום ו3 לכחול), ויודע גם להתמודד עם נקודה (הירוקה) נוספת שיכולה להצטרף לכל אחד מהקבוצות אבל לנקודה השמינית (שבגלל שהן לסירוגין) שצבעה יהיה שונה לא יהיה כיצד לתייג אותה (כל ה"קווים" תפוסים)

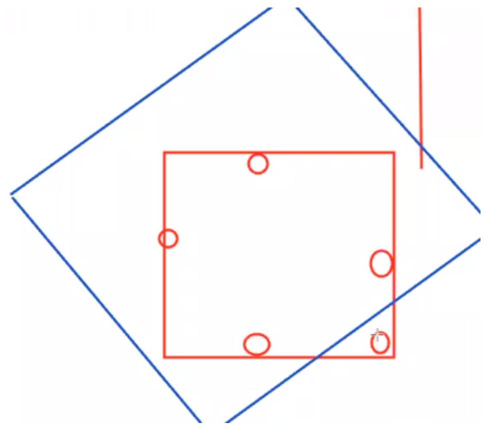
כעת נרצה לעלות במימדים (הוכחה פורסמה לפני חודש) :

The VC-dimension of k-gon in d-dimensional spase is $O(dk \log d)$

תיקון משיעור שעבר: בנוגע לVc של מלבנים וצריך לדייק שלא סתם מלבן, אלא:

ה $VC-dimension$ של מלבן **המאונך לצירים** ב$2d$ הוא 4

דוגמה שזה לא עובד אם לוקחים כל מלבן:



תזכורת - הלמה של סאוור

Sauer's lemma: Pi(S) is all possible labals assigned to S by rule set R is then

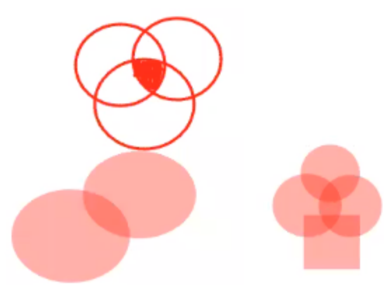
$$Pi\left(S\right)=\sum_{i=0}^{d}\binom{n}{i}<\left(\frac{em}{d}\right)^{d}$$

מה שטוב בסאוור שהוא מקטין לנו את האפשרות כאשר הוא מסתכל על התוצאות האפשרויות ולא על הנקודות. כעת נלמד משפט המשך (יועיל לשאלה 3 במטלה)

let $h_1, h_2... \in H$ be class of rule with VC-dimension d. Let H' be the set of s rules (h_i and / or h_2) then

$$VC=dim\left(H'\right)<2ds\log\left(2s\right)$$

הרעיון: לקחנו את החוקים מ H ויצרנו מהם קבוצות חוקים חדשות למשל:



והמשפט אומר שה$Vc-dim$ של אוסף החוקים החדש קטן מ $2ds\log\left(2s\right)$

רעיון ההוכחה:

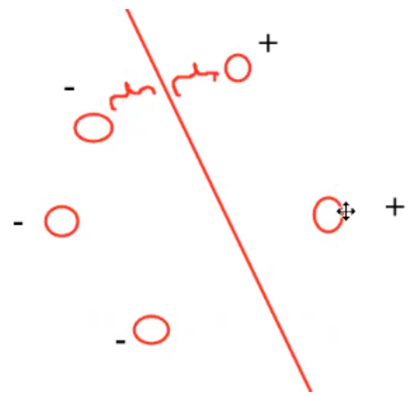
- Given any sample of m points, Sauer's lemma gives $Pi\left(S,H\right)<\left(\frac{em}{d}\right)^{d}$
- and Pi$(S,\boldsymbol{H'})<\left(\frac{em}{d}\right)^{ds}$
  - כאשר הגידול הוא בs , כי אם לדוגמה יש לי s חוקים עם 4 תוצאות אפשריות לכל אחד, אז החיבור בינהם לכל היותר מכפיל בs
  - כלומר אנחנו יודעים מהו מספר התוצאות המקסימלי, וכעת נרצה לשאול את השאלה ההפוכה, שבהנתן מספר התוצאות מהי הסדרה הכי גדולה שהוא יכול לנפץ $\iff$ מה הm הכי גדול שהוא יכול לנפץ $\iff$ מה הm הכי גדול שבעבורו $\left(\frac{em}{d}\right)^{ds}\geq2^{m}$

$$\left(\frac{em}{d}\right)^{ds}\geq2^{m}$$
$$if\ \ m=2ds\log3s$$
$$\log\left(3s\right)<\frac{9s}{2e}\ \forall s>2$$

סיימנו את החלק הקומבינטורי, ונעבור לחלק האלגוריתמי

הראנו ש $Vc-dim$ של $planes$ במימד d הוא $d+1$, הבעיה שהמימד של ה$data$ בקלות יכול להגיע למספור אסטרונומיים.

לפני מושג חדש: מישור עם מרווח $gamma$, בדוגמה יש את הקו שמפריד וה$gamme$ אלו הסוגריים ־ נרצה לקחת את המרווח הכי גדול



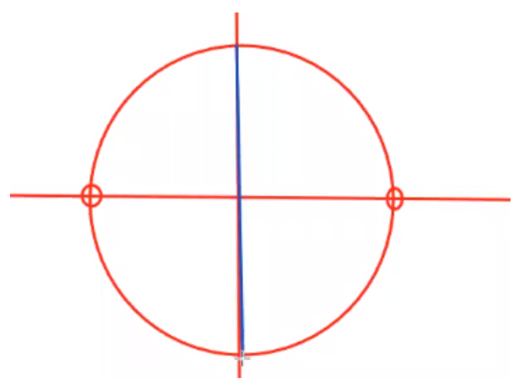
משפט:

VC-dimension of planes with margin γ is $\frac{1}{\gamma^2}$ independent of dimension

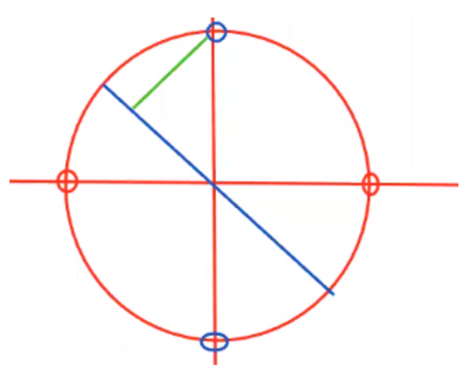
הערה: צריך לשים לב שמרחק מוגדר היטב.

אינטואיציה :

אם יש לנו 2 נקודות (מימד 1) המרחק נוכל לייצג את המרחק בינהם כך (כאשר הרדיוס = 1)



ובעבור 4 (2 מימדים) נקודות

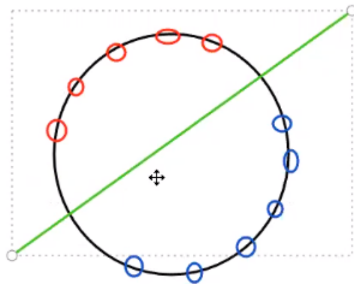


והמרחק הוא האורך של הירוק = $\frac{1}{\sqrt{2}}$ (עבור רדיוס = 1)

ובעבור מימד d המרחק הקסימלי הוא $\frac{1}{\sqrt{d}}$ $\Leftarrow$ שאוסף הנקודות שניתן לנפץ הוא $2d^2$

מסקנה: יש יחס בין גודל המרווח למספר הנקודות שניתן לנפץ

לכן נניח שיש לנו אוסף עם מרווח די גדול, כמו בדוגמה



(זה אומר שה $Vc-dim$ יחסית קטן) ונשאל כיצד מוצאים את הקו הכי טוב (עם המרווח המקסימלי)?

אם נעבור על כל האפשרויות, אז נמצא ונקבל זמן ריצה:

- יש סה"כ n^2 קווים אפשריים - כיון שצריך לבדוק את המרחקים בין כל 2 נקודות
- עבור כל קו צריך לבדוק מרחקים לכל n הנקודות
- סה"כ $O\left(n^3\right)$

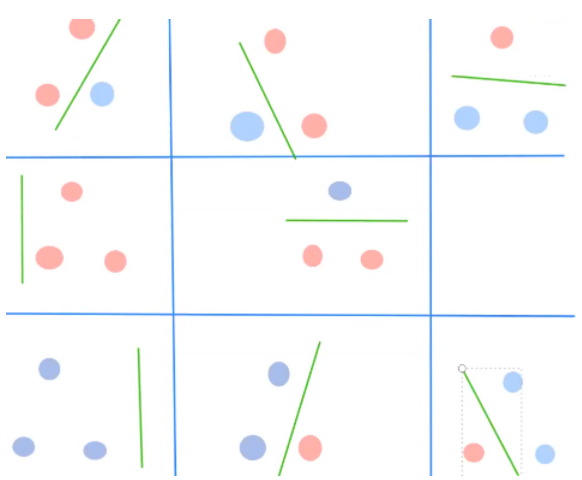
שיפור: אלגוריתם $preceptron$ זמן ריצה $\frac{n}{\gamma^2}$, וכנראה שזה אלגוריתם הלמידה הראשון שהמציאו, ואחד הפשוטים.

על כל אלה ועוד בשיעור הבא.

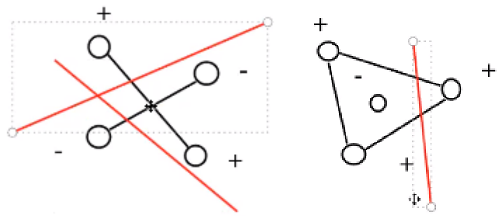
שיעור 5

תזכורות:

- vc-dimension of a line 2d $= 3$
  - proof : lower bound



- proof: upbbern bouund, two cases:



- vc-dimesnion of a hyperplane in d dimension = d+1
- proof: lowerbound:

$LowerBound$: נסמן את הנקודות $0, e_1, e_2....e_d$

ונעביר את המישור לפי התיוג

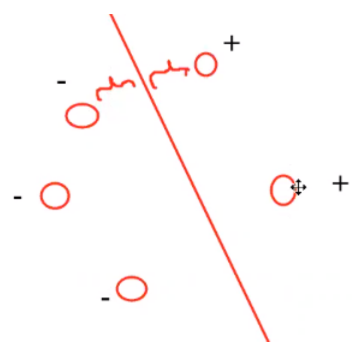
$UpperBound$:

אצלנו צ"ל להראות שלא קיים סט עם $d+1$ שניתן לנפץ. אם הן בקמור אחת של השניה - לא ניתן להפריד בינהם.

אחרת $\Leftarrow$ הן לא בקמור אחת של השניה $\Leftarrow$ נשתמש במשפט של רדון:

Any set of d+2 points in d-dimensional space can be partitioned into two disjoint sets A,B whose convex hull intersect

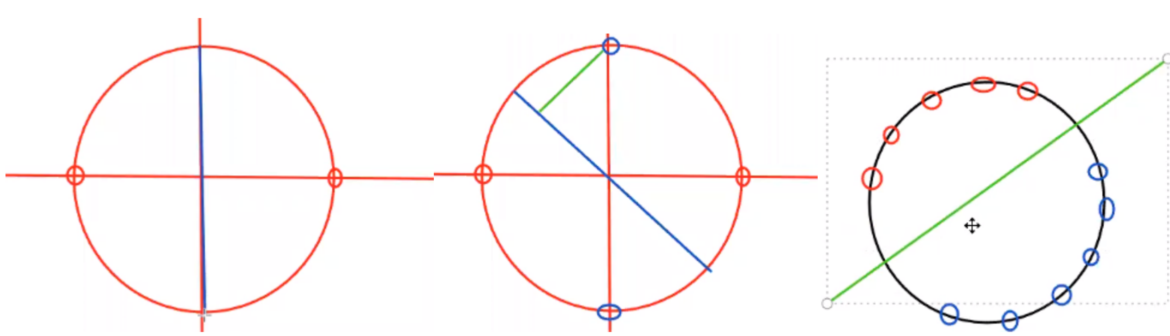
הזכרנו גם שבקלות ניתן להגיע מימדים מאוד גבוהים , והראנו שחיפשו איך ללמוד גם אם מימדים גבוהים מאוד, והראנו שאם ניתן למצוא מרווח כמו בציור:



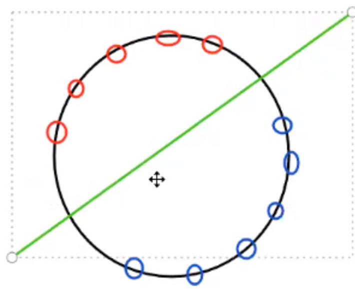
אז ניתן לבצע למידה גם במימדים מאוד גבוהים. וזה המשפט:

VC-dimension of planes with margin γ is $\frac{1}{\gamma^2}$ independent of dimension

אינואציה למשפט: המרווחים עבור 2 נקודות ו4 נקודות, ועבור n נקודות



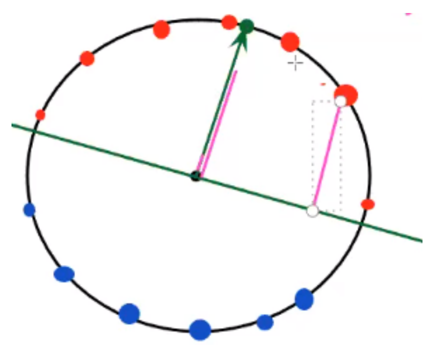
ניקח מעגל שהנורמה של כל הנקודות היא 1, ונחפש את המרווח האופטימלי:



אם נעבור על כל האפשרויות, אז נמצא ונקבל זמן ריצה: $O\left(n^3\right)$ - וזה אלגוריתם $Brute-force$

ועכשיו נלמד אלגוריתם חדש $Preceptron$ שזמן הריצה שלו $O\left(\frac{n}{\gamma^2}\right)$

לפני נזכר במושגים נורמה והטלה:



למשל מה המרחק בין מרכז המעגל לנקודה הירוק? תשובה: הנורמה (של הוקטור)

מה המרחק של כל נקודה אדומה מהמישור ? זו ההטלה על הוקטור לנקודה הירוקה

המכפלה הפנימית בין שתי נקודות במעגל (רדיוס 1) גדלה ככל שהן יותר קרובות

- אם יש 90° בינהן המכפלה נותנת 0
- אם זו נקודה כפול עצמו המכפלה נותנת 1

הערה: עבור הנקודות הכחולות אותו חישוב, רק התוצאה שלילית

אלגוריתם:

1. let $w_1 = 0$
2. iteration t = 1 ,2
   (a) At iteration t, go through all points for each point x
      i. if $\langle w_t,x\rangle < 0$ guess x is +
      ii. else ($\langle w_t,x\rangle \geq 0$) guess x is -
   (b) if guess is wrong
      i. x is +: $w_{t+1} = w_t + x$
      ii. x is - : $w_t - x$

(c) go to iteration t+1

(d) if no mistakes in this iteration -> stop

טענה:

preceptron algo' make only $\frac{1}{\gamma^2}$mistakes

מסקנה ־זמן ריצה:

- $\Leftarrow$ יש סה"כ $\frac{1}{\gamma^2}$ איטרציות $\Leftarrow$ בכל איטרציה עוברים על כל n הנקודות $\Leftarrow$ סה"כ $O\left(\frac{n}{\gamma^2}\right)$

נכונות:

נראה דרך 2 למות:

w^* is true hyperplane - המישור אותו אנחנו מחפשים

claim 1: $\langle w_{t+1}, w^* \rangle \geq \langle w_t, w \rangle + \gamma$

proof:

if w_t made mistake on x that is $(+)$ then:

$$\langle w_{t+1}, w^* \rangle = \langle (w_{t+1} + x), w^* \rangle = \langle w_{t+1}, w^* \rangle + \langle x, w^* \rangle = \langle w_{t+1}, w^* \rangle + \gamma$$

if w_t made mistake on x that is $(-)$ then the product of $\boldsymbol{\langle x, w^* \rangle = -\gamma} < 0$ and:

$$\langle w_{t+1}, w^* \rangle = \langle (w_{t+1} - x), w^* \rangle = \langle w_{t+1}, w^* \rangle - \boldsymbol{\langle x, w^* \rangle} = \langle w_{t+1}, w^* \rangle - (-\gamma') = \langle w_{t+1}, w^* \rangle + \gamma'$$

claim :2: $\|w_{t+1}\|^2 \leq \|w_t\|^2 + 1$

Proof:

if w_t made mistake on x that is $(\pm)$ then:

$$\|w_{t+1}\|^2 = \|w_{t+1} \pm x\|^2 = \|w_t\|^2 \pm \underbrace{\langle w_t, x \rangle}_{<0} + \underbrace{\|x\|^2}_{r} \leq \|w_t\|^2 + 1$$

claim 3: $\langle w_t, w^* \rangle \leq \|w_t\|$

proof:

$$\underbrace{\frac{w_t}{\|w_t\|}}_{\leq 1} \cdot \underbrace{w^*}_{=1} \leq 1 \iff \langle w_t, w^* \rangle \leq \|w_t\|$$

מכפלה של וקטור מנורמל (w^*) אם וקטור חלקי הנורמה שלו (ולכן קטן מ1) . המכפלה תהיה שווה 1 רק $w_t = w^*$

כעת, אם נסמן את מספר האיטרציות הסופי ב M נקבל ש:

$$M\gamma \overset{1}{\leq} \langle w_M, w^* \rangle \overset{3}{\leq} \|w_M\| \overset{2}{\leq} \sqrt{M}$$
$$\Downarrow$$
$$M \leq \frac{1}{\gamma^2}$$

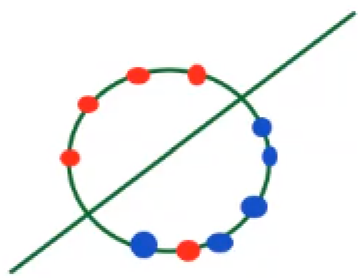
$\square$

Better perceptron

1. let $w_1 = 0$
2. iteration t = 1 ,2
   (a) At iteration t, go through all points for each point x
      i. if $\langle w_t, x \rangle \leq -\frac{\gamma}{2}$ guess x is +
      ii. if else $\left(\langle w_t, x \rangle \geq \frac{\gamma}{2}\right)$ guess x is -
      iii. consider $\langle w_t, x \rangle \in \left(-\frac{\gamma}{2}, \frac{\gamma}{2}\right)$ as mistake
   (b) if guess is wrong
      i. x is +: $w_{t+1} = w_t + x$
      ii. x is - : $w_t - x$
   (c) go to iteration t+1
   (d) if no mistakes in this iteration -> stop

האלגוריתם הזה נותן אפשרות להעריך את המרחק, וזמן ריצה די זהה (מוכפל בקבוע)

מה נעשה במקרה הבא?



הרי ה $precpetion$ לא יעצור לעולם (כי תמיד תהיה טעות) - כלומר מה נעשה כשאין הפרדה?

SVM - Supprot vector machines - alogorithem

רקע: וקניק (הרוסי שטוען שסאוור שלו) וקרוינה קורטז כתבו מאמר בנושא בשנות ה80 שהיא אישתו של מוהרי שהיה המרצה של גלעד.

https://moodlearn.ariel.ac.il/pluginfile.php/2013900/mod_resource/content/0/lect0125.pdf

מטרה: הפרדה בין הנקודות, עם מפריד גדול וללא שגיאות.

הבעיה: אולי זה לא אפשרי.

הרעיון: נסביר על מקרה שיש הפרדה, ונשליך ממנו על מקרה שאין הפרדה (כמו בציור)

assum data is separable

- where $\|w\|$

- $\max \gamma$ subject to: $\text{label}(x_i) * \langle wx_i \rangle > \gamma$

בשביל לעבור דרך ראשית הצירים - wx_i

- where $\gamma = 1$
  - if x_i, w not normalized, choose margin size to be $\gamma = 1$
  - $\max \gamma = \max\left(\min_x \left\{\frac{|wx+b|}{\|w\|}\right\}\right) = \max\left(\frac{1}{\|w\|}\right) = (x)$ המרחק המקסימלי בין המישור (b) לנקודה הכי קרובה (בציורים לעיל למשל הנקודה הכחולה בקצה)

~ המקרים הללו שקולים כי הערכים תלוים אחד בשני ~

כלומר:

SVM: $\min \|w\|$ subject to $\text{label}(x_i)*\langle w, x\rangle + b \geq 1$

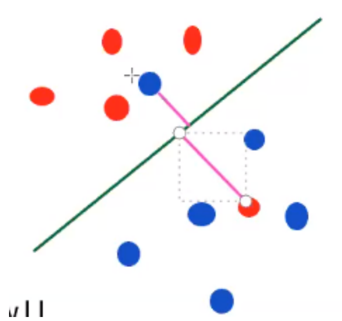
if data is non-separable.

Goal: fine hyperpalne with as few mistakes as possible?

problem: NP-hard, even hard to approximate

But has a SMV version: $\min\|x_i\|^2 + c\sum_{i\in[n]}\eta_i$ subject to $\text{label}(x_i) * \langle w, x_i\rangle \geq 1 - \eta_i$

בגרסה השניה - נותנים "מקום" לטעות ומשלמים בדיוק



לתשלום הזה קוראים $slack - variables$

שיעור 6

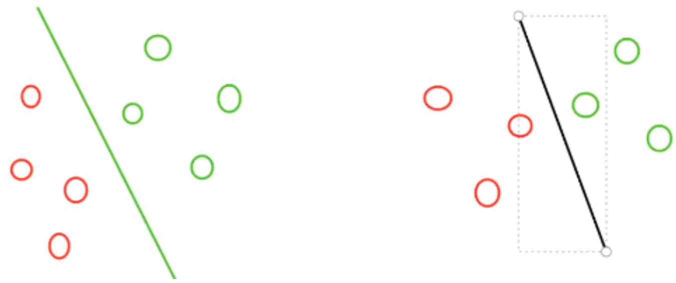
משפט:

Suppose a classifier form s family of VC-dim' d is consistent on smaple S ,the with high prob' ,the classifier has true error (m points)

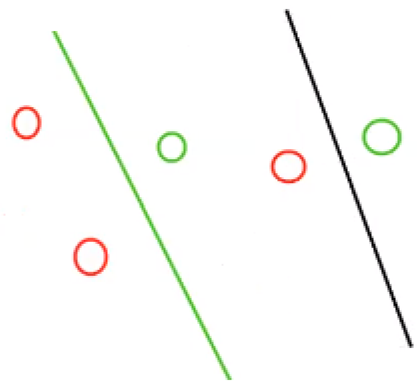
$$\frac{1}{m}\left(\log m + d\log\frac{m}{d} + O(1)\right) = O\left(\frac{d\log n}{n}\right) \cong O\left(\frac{d}{m}\right)$$

תכונה חשובה לאלגוריתמים שנלמד עכשיו - support vectors, הסבר:

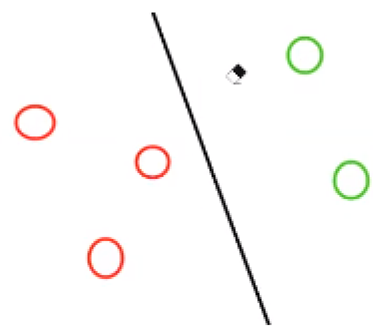
אוסף מתויג כך:



יהיה מתויג בדיוק אותו דבר = הקו היה נשאר באותו מקום , גם אם ימחקו נקודות:



לעומת מקרה שבו נמחקה נקודה מה $support\ vector$, ואז מיקום הקו משתנה.



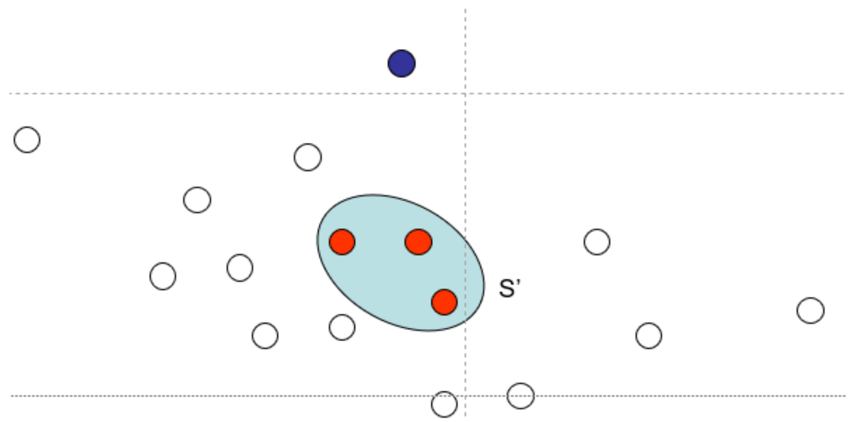
כמה זמן לוקח למצוא את ה $supprot\ vector$, אם ננסה כל אפשרות, עבור n נקודות ל d מימדים יש: $O\left(n^{d+1}\right)$, כלומר זמן ריצה מאוד בעייתי.

אלגוריתמי $Boosting$

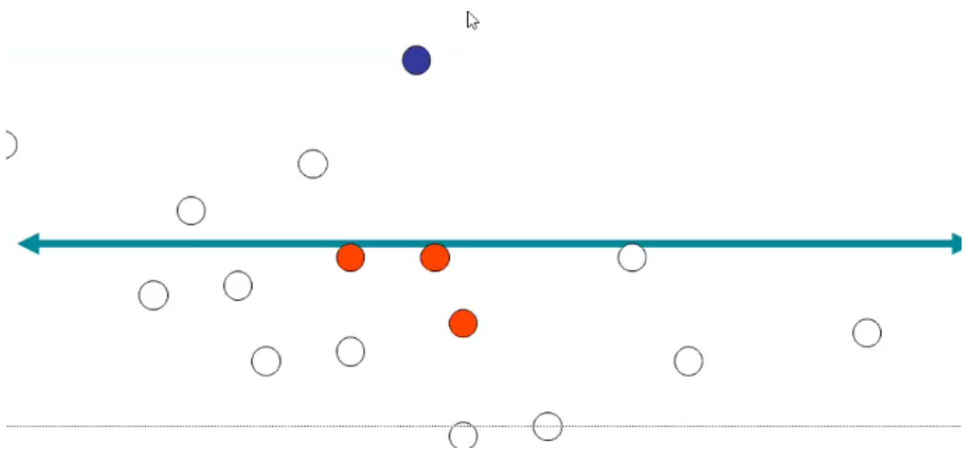
האלגוריתם של קלרקסון

לשם הדוגמה נניח ויש נקודה אחת כחולה ואנחנו רוצים למצוא את ה $margin$ הגדול ביותר בינה לבין האדמות

- אז נבחר, 3 נקודות רנדומלית



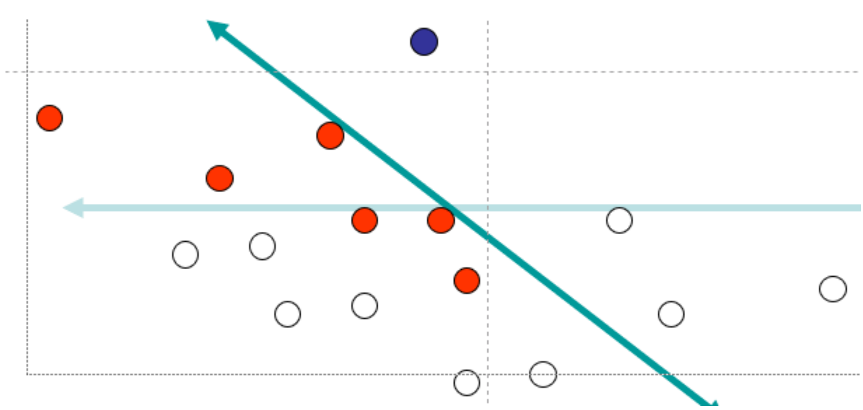
- ונעביר קו על פיהם = פעולה יקרה למציאת המרווח הכי גדול $\Leftarrow$ על פי המשפט מתחילת השיעור טעינו בלכל היותר ב $O\left(\frac{d}{m}\right)$ נקודות



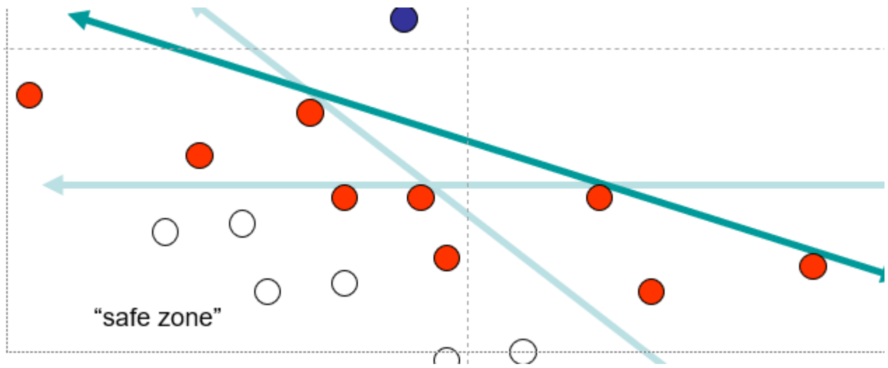
שתי תובנות:

- אם יש טעות $\Leftarrow$ אחת מהנקודות שטעיתי בהן שייכת ל $support\, vector$
- אם אין טעות $\Leftarrow$ העברתי קו אופטימלי

- לכן נדגום (נוסיף אותן ל S') מתוך ה"טעות" ונעביר קו חדש



- הקו החדש - עיקבי על המדגם המקורי $\Leftarrow$ לא טועה על הרבה נקודות



נכליל את המשפט מתחילת השיעור עם בחרתי רק n נקודות אז הטעות שלי היא לכל היותר $O\left(\frac{d_{vc}n}{m}\right)$

כעת, עולה השאלה כמה נקודות נרצה לדגום בכל איטרציה (ולמזער את הזמן הריצה)?

נבחר בכל שלב $|S'| = \sqrt[n]{d}$ נקודות נוספות, ואז נקבל:

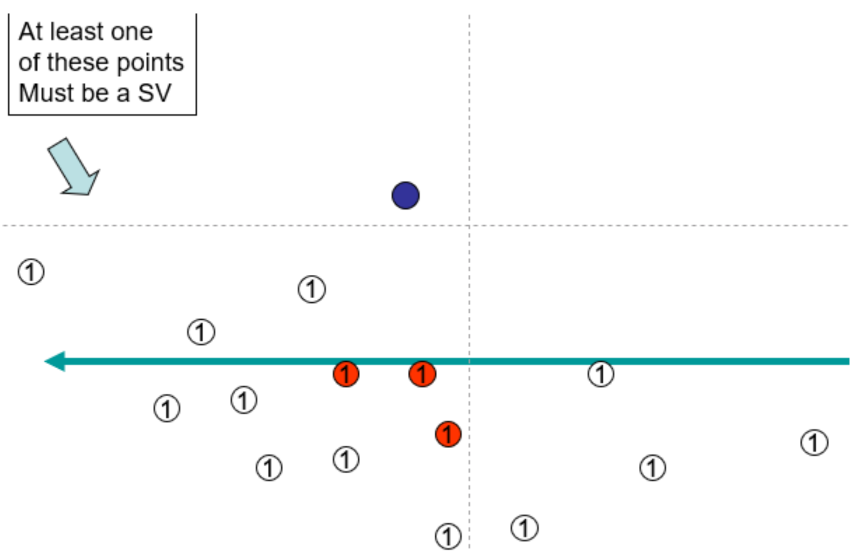
$$O\left(\frac{d_{vc}n}{m}\right) = O\left(\frac{d_{vc}n}{\sqrt[n]{d}}\right) = \left(\frac{\cancel{d_{vc}}n}{\cancel{d_{vc}}n^{0.5}}\right) = n^{0.5}$$

כמה אטרציות יש?

- אמרנו שעבור מימד d יש לכל היותר $d+1$ וקטורים , ולכן יש לכל היותר $O\left(d\right)$ איטרציות
- בכל איטרציה נבדוק $\sqrt{n}$ נקודות, ונוסיף אותם לוקטור
- לכן זמן הריצה $2d\sqrt{n} = O\left(d\sqrt{n}\right)$

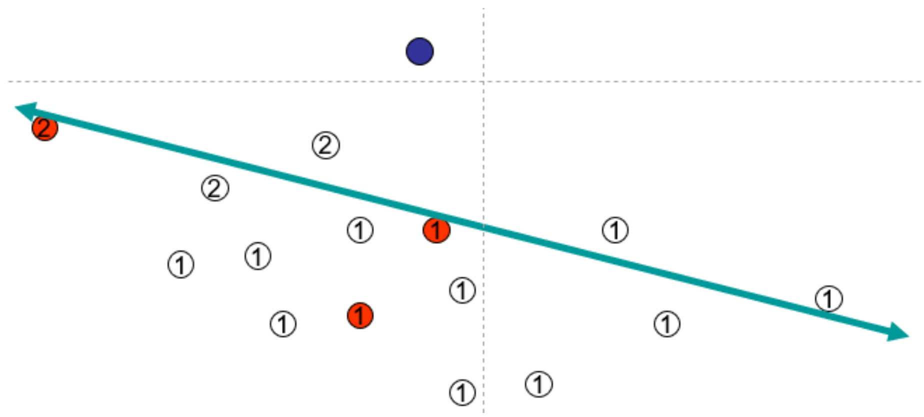
אלגוריתם 2 :

- נאתחל את כל הנקודות למשקל 1 , וכמו מקודם נדגום $|S'|$ נקודות, ונעביר קו:

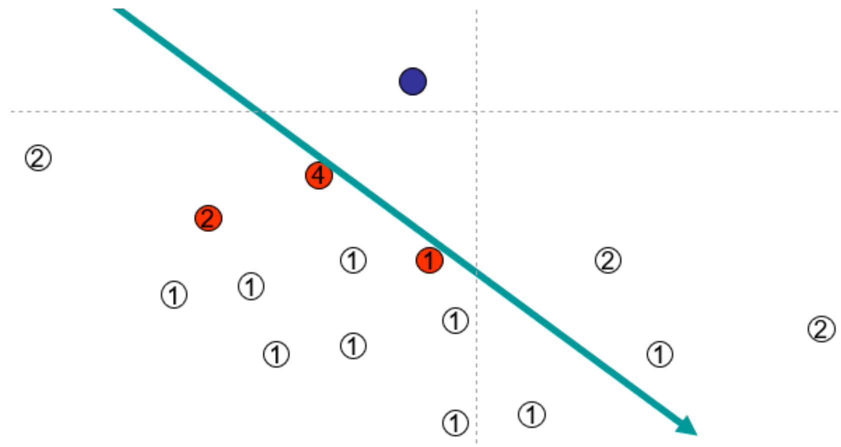


- כל נקודה שהיא טעות נכפיל את משקלה $\Leftrightarrow$ להגדיר שכל נקודה היא בעצם 2 נקודות
- נזרוק את המדגם הקודם

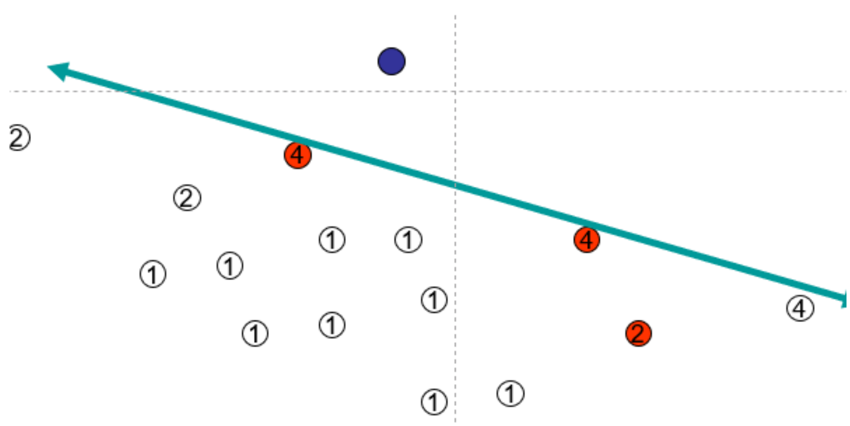
- נדגום שוב על פי המשקל, (כעת ההסבתרות לדגום את הנקודות מהטעות הקודמת עלתה), ונשמתש בהם על מנת להעביר קו:



- נמשיך , נכפיל את הטעויות, נזרוק, נדגום, נעביר קו:



- ושוב:



- אם אין טעויות, סיימנו
  - קלרקסון מוכיח שמבספר אטרציות נמוך נקבל SV שצודק בהסתברות מאוד גבוהה

Clarkson's second algorithm:

$O(\log n)$iterations, eac on sample of size $O\left(d^2\right) \simeq d^6 \log n$

facts - Suppoet vector (SV)set V:

- At each iteration ,the weight of one SV doubles . after kd iterations weight of suppoert vectors is at least $d \cdot 2^k$

Entire set - (all points), c is constant:

- At each iteration misclassifies $O\left(\frac{d}{m}\right) = O\left(\frac{1}{d}\right) < \frac{1}{cd}$
- after kd iteration, weight of the entrie set at most $n\left(1+\frac{1}{cd}\right)^{kd}$
  - n is the initial points weight or point's count (depend on implementation)
- The weight of the set of SV can't be greater than the weight of entire set

$$d \cdot 2^k < n \cdot e^{\frac{k}{d}} \iff k = O(\log n)$$

- ההסבר לכך הוא שבכל שלב אמנם אנחנו חוסמים בגדול הטעות מלמעלה, אבל אנחנו גם חוסמים מלמטה בלפחות 1 (אחרת היינו עוצרים), ולכן משקל ה sv גדל יותר מהר מהמשקל של כל העולם
- שיפור קל: אם נפלתי על דגימה עם אחוז טעות גבוה - נתעלם ממנה

אלגוריתם $window$

דוגמה: אני מאמן ואני בוחר שחקנים לקבוצת כדרוסל, ונניח שיש לי שני חוקים:

- יודע לשחק
- אדם גבוה (לא בהכרח יודע לשחק)

מגיע מישהו מבחוץ שלא מכיר את החוקים לבחירה, ומנסה ללמוד אותה, ולהבין מה הם החוקים.

דוגמה אחרת, דוגמת הבית קפה מהשיעור הראשון (למידה מי מזמין קפה/תה)

ונרצה להבין מה הם המאפיינים החשובים

	X_1	X_2	X_3	X_4	
Person	old/yound	m/f	Tall/short	Resident/tourist	Coffe/Tea
$A - y_1$	1	0	1	0	+
$B - y_2$	0	1	0	1	+
$C - y_3$	1	1	0	0	-
$D - y_4$	0	1	0	0	-

Algorithm:

1. Weights $w_1 = \ldots = w_n = 1$ (n = number of properties, r = number of important properties)
2. For y_j
   (a) guess + if: $\sum_{i=1}^{n} w_i x_i(y_j) \geq n$
   (b) guess – else

3. On a mistake for y_j:
    - A. guessed + but - was true:
        - for all i with $x_i(y_j) = 1$,
          set $w_i = 0$
    - B. guessed - but + was true:
        - for all i with $x_i(y_j) = 1$,
          set $w_i = 2w_i$
4. Iterate on all points until no mistake is found

Claim: At most $2 \cdot r \cdot \log n$ mistakes (r number of true properties)

Proof:

- Mistake B can happen at most $r \cdot \log(n)$ times.
    - After $\boldsymbol{r \cdot \log(n)}$ mistakes, every important x_i has weight n.
- Mistake A can happen at most $\boldsymbol{r \cdot \log(n)}$ times.
    - Mistake A removed at least n total weight (= sum of w_i's)
    - Mistake B adds total weight less than n $\Rightarrow$

$$\#\text{ of mistakes of type A} < \#\text{ of mistakes of type B} \leq \boldsymbol{r \cdot \log(n)}$$

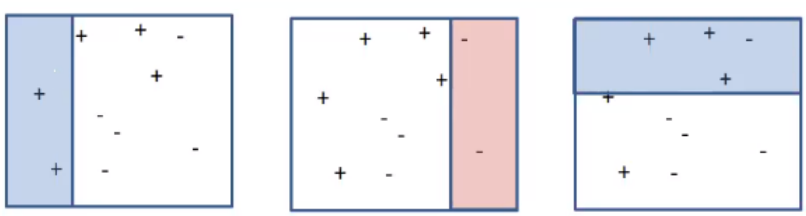
AdaBoost

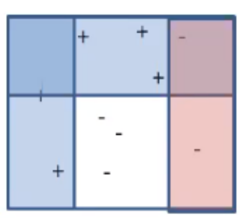
הרעיון - לקחת חוקים פשוטים ולבנות מהם חוק מורכב, נזכר במשפט:

Given a set of simple hypothese of vc-dimension d , the set of hypotheses formed by the intesection of s simples rules has vc-dimension of: $2d \cdot s \cdot \log(3s)$

הרעיון הוא: שאם לא בנית אוסף גדול מדי, אתה מקבל חסם די טוב, (אחרת הוא גדול מדי)

דוגמה:





בשורה העליונה כל חוק הוא טוב במידה מאוד מסוימת, בשורה התחתונה בנינו מקומבינציה שלהם (שני הכחולים פחות האדום) למידה מאוד מדויקת

נדגיש שלא נרצה לערבב יותר מדי חוקים, כי אז ה$vc-dim$ יעלה מאוד.

Given: wak classifiers $h_j : x \to \{-1,1\}$ in H

Answer:

F(x) $= \sum\limits_{t\in|H|} \alpha_t h_t(x)$

H(w) = sign$[F(x)]$=sign$\left[\sum\limits_{t\in|H|} \alpha_t h_t(x)\right]$

$Adaboost$ מקבל המון חוקים, ונותן משקל לכל חוק, שזה כמו לומר מה הם החוקים החשובים. איך הוא עושה זאת?

נסמן את $error_t = \epsilon_t$

1. Give every point $x_i$weight $D\left(X_i\right) = \frac{1}{m}$
2. Compute weigthed erro for each classifier

$$\epsilon_t\left(h\right) = \sum_{i=1}^{m} D_t h_j\left(x_i \neq y_i\right) \forall j \in [1,|H|]$$

3. selelct classifier with smallest error, $\alpha_j = \frac{1}{2}\ln\left(\frac{1-\epsilon_t(h_t)}{\epsilon_t(h_t)}\right)$
4. update point weights $D_{t+1}(x_i) = D_t\left(x_i\right) \cdot e^{-\alpha_j \cdot h_j(x_i) \cdot y_i}$

בשיעור הבא נראה איך הנוסחא הזו לעדכון מעלה חוקים חשובים, ומורידה חוקים שאינם.

שיעור 7

זה המשפט המרכזי שלנו:

Given a class H of hypotheses, random sample S of size m , if we find some hypothesis h in H that consistent with S, then for any $0<\delta<1$, its true rror on the space is:

$$e\left(h\right) = \frac{1}{m}\cdot\left(\log|H| + \log\left(\frac{1}{\delta}\right)\right)$$

with probability $1-\delta$, while $d = vc-dim$ of H

וע"פ המשפט של סאוור ניתן לשפר, עבור $d = vc-dim$ של H :

$$e\left(h\right) = \frac{1}{m}\cdot\left(\log m^d + \log\left(\frac{1}{\delta}\right)\right)$$

גרסה נוספת, עבור חוק עם טעות אמפירית קיימת:

if we find some hypothesis $h \in H$ with empirical error $\overline{e}\left(h\right)$, then with prob' $1-\delta$:

$$e\left(h\right) = \overline{e}\left(h\right) + \sqrt{\frac{\log|H| + \log\left(\frac{2}{\delta}\right)}{2m}}$$

ושוב בעזרת סאוור:

$$e\left(h\right)=\overline{e}\left(h\right)+\sqrt{\frac{\log m^{d}+\log\left(\frac{2}{\delta}\right)}{2m}}$$

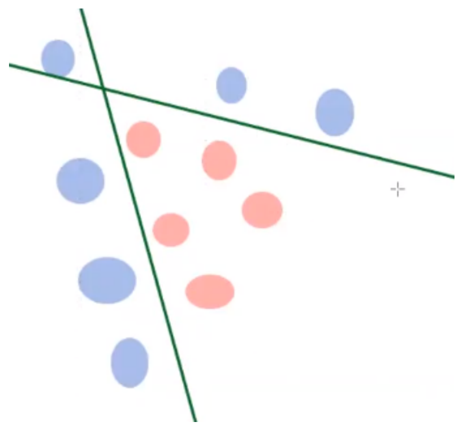
וזה מוביל אותנו ל $bias-variance\ tradeoff$, מה הטריידוף?

- עבור חוק פשוט כנראה ש$\overline{e}\left(h\right)$ יהיה גבוה ו $\sqrt{...}$ יהיה נמוך
- עבור חוק מסובך (עד כדי $over-fitting$) כנראה שה $\overline{e}(h)$ יהיה נמוך וה $\sqrt{..}$ יהיה גבוה

וזה נושא מרכזי בלמידת מכונה , מציאת האיזון המתאים.

$AdaBoost$

אלגוריתם שלוקח חוקים פשוטים ובונה חוק מסובך, דוגמה:



ה"בעיה" שאיחוד חוקים עלול להוביל לחסם גבוה, נזכר במשפט:

if we take the union of s hypotheses from a set H with VC-dim d, the resulting set has VC-dim

$$2ds\log(3s)$$

מה שנוכיח ש $AdaBoost$ יגיע לחוק עיקבי עם חסם טוב די מהר.

הגדרות:

Input:

- hypthesis: $h\in H\ \ h:S\to\{-1,1\}$
- samplle: m labelled points (x_i,y_i)
- Iterations: r

output:α_i and hypotheses $h_i\in H$

- Function: $F\left(x\right)=\sum_{i=1}^{r}\alpha_i\cdot h_i\left(x\right)$

- $\mathrm{H}(x) = sign\left(F(x)\right)$

Algorithem:

1. Initialize points $x_i$weight $D\left(X_i\right) = \frac{1}{m}$
2. For i = 1,...,r
   (a) Compute weigthed error for each classifier
   $$error\left(h\right) = \sum_{j=1}^{m} D_i\left(x_j\right) \cdot h_j\left(x_j \neq y_j\right)$$
   (b) let h_i be the hypothesis with lowest error
   (c) Compute: $\alpha_j = \frac{1}{2}\ln\left(\frac{1-error(h_i)}{error(h_i)}\right)$
   (d) update point weights $D_{i+1}\left(x_j\right) = \frac{1}{Z_i} \cdot D_i\left(x_j\right) e^{-\alpha_i \cdot h_i(x_j) \cdot y_j}$

- נשים לב שאם הטעות לחוק h_i היא $\frac{1}{2}$ אז $\alpha_j = 0$ (והחוק לא נחשב)
- בביטוי $D_i\left(x_j\right) e^{-\alpha_i \cdot h_i(x_j) \cdot y_j}$:
  - אם צדקנו נקבל בחזקה ביטוי חיובי ובגלל המינוס בחזקה המשקל של D_{i+1} ירד
  - אם טעינו נקבל בחזקה ביטוי שלילי ובגלל המינוס האקספ' **יגדל** והמשקל D_{i+1} יעלה
- Z_i הוא נרמול של כל הנקודות.: $Z_i = \sum_{j=1}^{m} D_i\left(x_j\right) e^{-\alpha_i \cdot h_i(x_j) \cdot y_j}$

Analysis - Note that:

$$\begin{aligned} D_{i+1}\left(x_j\right) &= \tfrac{1}{Z_i} \cdot D_i\left(x_j\right) e^{-\alpha_i \cdot h_i(x_j) \cdot y_j} = \tfrac{1}{Z_i \cdot Z_{i-1}} \cdot D_{i-1}\left(x_j\right) e^{-(\alpha_i \cdot h_i(x_j) \cdot y_j + \alpha_{i-1} \cdot h_{i-1}(x_j) \cdot y_j)} \\ &= ... = \tfrac{1}{Z_i \cdot Z_{i-1} \cdot ... \cdot Z_1} \cdot D_1\left(x_j\right) e^{-(\alpha_i \cdot h_i(x_j) \cdot y_j + \alpha_{i-1} \cdot h_{i-1}(x_j) \cdot y_j + ... + \alpha_1 \cdot h_1(x_j) \cdot y_j)} \end{aligned}$$

(הגדרת המשקל היא רקרוסיבית)

lets Define $Z^i = Z_i \cdot Z_{i-1} \cdot ... \cdot Z_1$, $D_1\left(x_j\right) = \frac{1}{m}$by defination and e's power is F(x), so:

$$D_{i+1}\left(x_j\right) = \frac{1}{Z^i} \cdot \frac{1}{m} e^{F(x_j) \cdot y_j}$$

note that $Z = Z^r$ so:

$$Z = \frac{1}{m} e^{-y_j F(x_j)}$$

החוק הסופי: $H(x_j) = sign\left[F(x_j)\right]$

claim 1: $\underbrace{error\,(H)}_{LHS} < \underbrace{Z}_{RHS}$

- Note: $\mathrm{F}(x_j) = sign\,[F(x_j)]\,|F\,(x_j)| = H\,(x_j) \cdot |F\,(x_j)|$
- if $H\,(x_j) \neq y_i \Rightarrow LHS = error\,(H) = 1 \leq Z = RHS = e^{+|F(x_j)|}$
- if $H\,(x_j) = y_i \Rightarrow LHS = error\,(H) = 0 \leq Z = RHS = e^{-|F(x_j)|}$
- What is Z_r?

$$\sum_{j=1}^{r} D_r\,(x_j)\, e^{-y_j \alpha_r h(x_j)}$$
$$= \sum_{j\in A} D_r\,(x_j)\, e^{-\alpha_r \cdot 1} + \sum_{j\in \overline{A}} D_r\,(x_j)\, e^{\alpha_r (-1)(-1) = \alpha_r}$$

  - A is the indices of points that h gets right
  - $\overline{A}$ is the indices of points that h gets wrong

- lets minimize Z_r:

$$(Z_r)' = \frac{dZ_r(\alpha_r, h_r)}{d\alpha_r} =$$
$$= \sum_{j\in A} -D_r\,(x_j)\, e^{-\alpha_r} + \sum_{j\in \overline{A}} D_r\,(x_j)\, e^{\alpha_r} = 0$$
$$\Rightarrow \underbrace{\sum_{j\in A} D_r\,(x_j)}_{1-error(h_r)} = \underbrace{\sum_{j\in \overline{A}} D_r\,(x_j)\, e^{2\alpha_r}}_{error(h_r)}$$
$$\Rightarrow \alpha_r = \frac{1}{2}\ln\left(\frac{1-error(h_r)}{error(h_r)}\right) = \frac{1}{2}\ln\left(\frac{1-\epsilon(h_r)}{\epsilon(h_r)}\right)$$

claim 2: Z decreases exponentially in r* (* iteration that we found a good rule)

$\sum_{j\in A} D_r\,(x_j) \rightarrow 1 - error\,(h_r) = 1 - \epsilon\,(h_r)$

$\sum_{j\in \overline{A}} D_r\,(x_j) \rightarrow error\,(h_r) = \epsilon\,(h_r)$

We can plug it back into the normalization term to get the minimum:

$$Z_r = \sum_{j\in A} D_r\,(x_j)\, e^{-\alpha_r} + \sum_{j\in \overline{A}} D_r\,(x_j)\, e^{\alpha_r} =$$
$$= (1 - \epsilon_r\,(h_r))\sqrt{\frac{\epsilon_r(h_r)}{1-\epsilon_r(h_r)}} + \epsilon_r\,(h_r)\sqrt{\frac{1-\epsilon_r(h_r)}{e_r(h_r)}}$$
$$= 2\sqrt{\epsilon\,(h_r)\,(1 - \epsilon\,(h_r))}$$

Change a variable:$\gamma = \frac{1}{2} - \epsilon\,(h_r)\,, \gamma_r \in \left(0, \frac{1}{2}\right)$

Then, we have the minimum to be:

$$Z_r = 2\sqrt{\epsilon\,(h_r)\,(1 - \epsilon\,(h_r))} =$$
$$= \sqrt{1 - 4\gamma_r^2} \leq e^{-2_r^2}$$

As long as $error\,(h_r) < \frac{1}{2} - c$, Z decreases exponentially So does $H\,(x_j)$

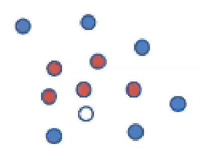
Therefore, after r steps, the error rate of the strong classifier is bounded on top by

$$Error\,(H) \leq Z = Z_1 \cdot \ldots \cdot Z_r \leq e^{-\left[2\sum\limits_{i=1}^{r}\gamma_r^2\right]}$$

שיעור 8

נעבור לאלגוריתם חדש- מציאת שכן קרוב.

נניח שיש לנו את סט הנקודות הבא



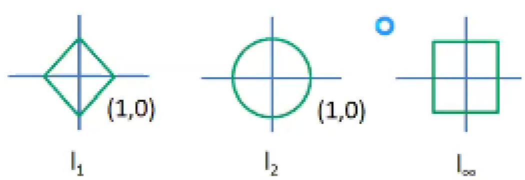
נרצה לשאול עבור קורדינטה מסוימת, מה השכן הכי קרוב. בשביל לענות על זה צריך לברר איך אנחנו מגדירים מרחק, ויש לנו מספר אפשרויות לכך:

- Eculidean distance (l_2): $d\,(x,y) = \|x-y\| = \sqrt{\sum\limits_{i=1}^{d}(x_i-y_i)^2}$

  for example:: $\begin{array}{c} x=(1,2,3) \\ y=(7,6,5) \end{array}$ then: $d\,(x,y) = \sqrt{6^2+4^2+2^2} = 7.48$

- Manhattan distance (l_1):$d\,(x,y) = \sum\limits_{i=1}^{d}|x_i-y_i| = 6+4+2 = 12$

- $L_p - distance$: $\sqrt[p]{\sum\limits_{i=1}^{d}|x_i-y_i|^p}$



בדוגמה רואים את הגרפים עבור $d = l_p = 1$

במרחק מנהטן(l_1) : נשים לב שעל הצירים יהיו לנו את הנקודות .. $(1,0)\,,(0,1)$ ונקודה כלשהי באמצע תראה : $(0.5,0.5)$, ולכן אנחנו מקבילים מעין מעוין.

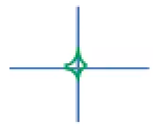
במרחק אוקלידי (l_2) : הוא מעגל היחידה על ראשית הצירים, ולמעשה הוא גרף שמראה את כל הנקודות שבמרחק אוקלידי (l_2) מראשית הצירים

נשים לב שעליה ב p "דוחפת" את הגבולות החוצה וזו האינטואיציה עבור $p \to \infty$ שהגרף שלו נראה כמו ריבוע. פורמלית:

$$Frechet\ distance = l_\infty = \max_i |x_i - y_i|$$

In our example: 7-1=6

ישנה טענה ־ מאמר שליעד כותב ־ על לקחת $p < 1$, אם נמשיך עם האינטואיציה שפיתחנו ־ נקבל שהקווים ”נדחפים פנימה“ ונקבל את הצורה הבאה:



$l_{1/2}$

- metrica definaton:

  distances obey:

  - Symmetric: $f(x,y) = f(y,x)$
  - $d(x,y) = 0 \iff x = y$
  - Triangle inequality, for every x,y,z $f(x,y) \leq f(x,z) + f(z,y)$

- l_p with $(p < 1)$ dosen't satisfy triange inequality

  - for example:

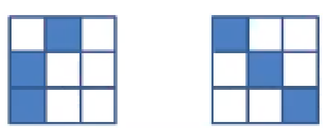
$$x = (0,0) \qquad y = (4,4) \qquad z = (2,0)$$

$$p = \frac{1}{2} \qquad f(x,y) = \left(4^{\frac{1}{2}} + 4^{\frac{1}{2}}\right)^2 = 16$$

$$f(x,z) = \left(2^{\frac{1}{2}} + 0^{\frac{1}{2}}\right)^2 = 2$$

$$f(z,y) = \left(2^{\frac{1}{2}} + 4^{\frac{1}{2}}\right)^2 = 11.65$$

Earthmover distance:



אם נשאל מה המרחק בין התמונות ־ אז עבור חישוב סך ההזזות (l_1) נקבל שהמרחק הוא 4 , דרך אחרת לפתור את הבעיה היא דרך בעיה מוכרת:

Assignment problem:

- Input:
  - N Workers
  - N tasks
  - Matrix: Cost of each worker doing task
- OUtput: 1 to 1 assingmnet of each worker to each task Minimize total cost

Min sum total matching

DNA

Suppose we had two DNA sequences:

1. GCGCAATG
2. GCCCTAGCG

how we can achieve the 2 sequence from the 1 sequence, through these actions: Insetion, deletion,Subtitution.

This distance name is: Levenstein distnace

Dna -dynamic program

	-	G	C	G	C	A	A
-	0	1	2	3	4	5	6
G	1	0	1	2			
G	2	1	1	1			
C	3	2	1	2			
T	4	3	2	2			
A	5	4	3	3			
G	6	5	4	3			

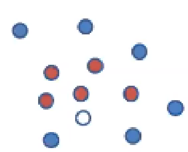
t

s

Alogrithem:

$$A(i,j) = \min\left\{A(i-1,j)+1, A(i,j-1)+1, A(i-1,j-1)+A(i-1,j-1)+\mathbb{I}\left(s[i]==t[i]\right)\right\}$$

Generalizion Bounds

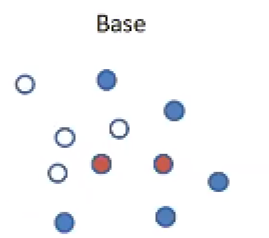


- 1-NNS - infinite vc-dim':

  - Take the smaple to be the vase set
  - The nearest nghgbor n the base set of every point in the sample is itself

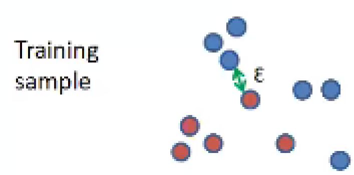
- K-NNS
  - Take K neareset neighbor , majority
  - Good for de-noinsing

- Condensing
  - Base set is bounded by some s
  - Rule: NN in the base set
  - Vc-dimension : $\mathrm{O}(s)$



שיעור 9

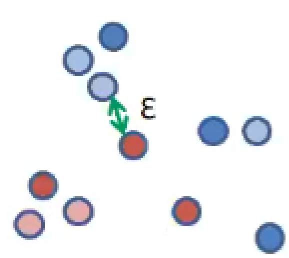
NN Condesing



let ε be the minimum distance from Red to Blue sets

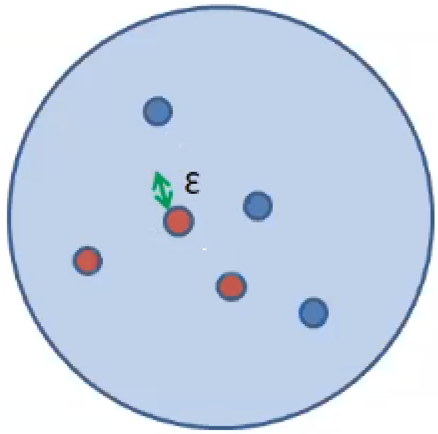
Defintaion: $\varepsilon - net$ of S is subset is a T satisifing the following:

- **Packing**: for every p, q in T: $d(p,q) \geq \varepsilon$
- **Convering**: for very p in S: $d(p,T) < \varepsilon$

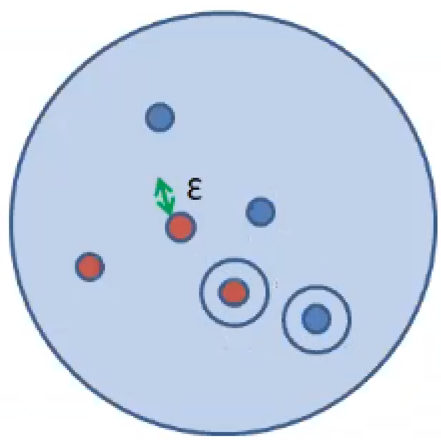


Take T as base for K-nn. 1-NN consistent on S

How big is T?



1. Big ball contains al points - radius 1
2. Points of T are at distance greater then ε
3. if I draw balls of radius $\frac{\varepsilon}{2}$ around points of T, these balls do not intesect



Can now bound the size of T Volume of big ball?

$$V(d,1) = \frac{\pi^{\frac{2}{2}} \cdot 1^d}{\Gamma(d)}$$

voulime of small balls:

$$V\left(d, \frac{2}{\varepsilon}\right) = \frac{\pi^{\frac{2}{2}} \cdot \left(\frac{\varepsilon}{2}\right)^d}{\Gamma(d)}$$

then:

$$\text{Size of T} < \frac{V(d,1)}{V(d,\varepsilon)} = \left(\frac{2}{\varepsilon}\right)^2$$

אז מה למדנו פה?

let ε be the **margin** distance from Red to Blue sets.

Given a sample S with margin ε , Extract from S an $\varepsilon - net$ T , and use T as he base for 1-NN classifier.

1. 1-NN is consisitent on S
2. VC-dim of the set of classifiers $O\left(\left(\frac{2}{\varepsilon}\right)^d\right)$

הורדת מימד

Johonson-Lindestrauss lemma '84:

Given a set V of n Euclidean vectors in d-dim' space R^d , and any $0 < \varepsilon < 1$

There exists a linear function $f : V \to R^k$ Such that for all $y, z \in V$, $|V| = n$

$$(1-\varepsilon)\|y-z\| < \|f(y) - f(z)\| < (1+\varepsilon)\|y-z\|$$

For $k = \frac{8 \ln n}{\varepsilon^2}$

example:

$$\begin{pmatrix} 8 & 1 & 6 & 3 \end{pmatrix} \begin{pmatrix} -1 & 1 \\ 1 & -1 \\ 1 & 1 \\ -1 & 1 \end{pmatrix} = \sqrt{2}\begin{pmatrix} -4, 16 \end{pmatrix}$$

$$\begin{pmatrix} 8 & 1 & 6 & 3 \end{pmatrix} \begin{pmatrix} N(0,1) & N(0,1) \\ N(0,1) & N(0,1) \\ N(0,1) & N(0,1) \\ N(0,1) & N(0,1) \end{pmatrix} = \sqrt{2}\begin{pmatrix} ?, ? \end{pmatrix}$$

JL-transform

- Take f to be defined by matrix X, where:
  - $x_{i,j}$ is ab bernoulli random varibel with $p = 0.5$ in $\{-1, 1\}$
  - if w=y-z:
    - let $w' = f(w) = f(y-z) = f(y) - f(z)$
    - so $\|w'\| = \|f(y) - f(z)\|$
    - Suffices to show that $\|w\| \approx \|w'\|$
    - withou loss of genrality assume $\|w\| = 1$

  1. $w' = \left(\sum_j^d x_{1j} w_j, \sum_j^d x_{2j} w_j, \ldots\right)$
  2. $\|w'\|^2 = \left(\sum_j^d x_{1j} w_j\right)^2 + \left(\sum_j^d x_{2j} w_j\right)^2 + ..$
  3. $E\left[\left(\sum_j^d x_{1j} w_j\right)^2\right] = E\left[\sum_j^d x_{1j}^2 w_j^2 + 2\sum_{i \neq p} x_{1i} x_{1p} w_i w_p\right]$
  4. $\sum_j^d w_j^2 + 0 = \|w\|^2$

- $E\left[\|w'\|^2 = k\|w\|^2\right]$

1. ההכפלה על פי הגדרה .2 הנורמה בריבוע. .3 התוחלת + חוקי התוחלת. .4 הערכים הם $\{-1,1\}$ + תוחלת 0

$JL - transform\ 2$

- Take f to be defined by matrix X, where:
  - $x_{i,j} \backsim N(0,1)$
  - $w' = \left(\sum_i^d w_i x_{1i}, \sum_i^d w_i x_{2i,\dots}\right)$
  - $(w'_1)^2 = \left(\sum_i w_i x_{1i}\right)^2$
  - Noramal distribution has the followung properties:
    - $aN(0,1) = N\left(0,a^2\right) = N\left(0,a^2\right)$
    - $N\left(0,a^2\right) + N\left(0,b^2\right) = N\left(0,a^2+b^2\right)$
  - $\sum_i w_i x_i \backsim N\left(0,\|w\|^2\right) = \|w\| N(0,1)$
  - $(w'_1)^2 \backsim \left(\|w\|^2 N(0,1)^2\right) = \chi^2$ (1 degree of freedom)
  - $\|w'\| = \sum_i w_i'^2 = \chi^2$ (k degree of freedom)

$$Pr\left(\|w'\|^2\right) > (1+\varepsilon) E\left[\|w'\|^2\right] < e^{\left(\frac{k}{2}\left(\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3}\right)\right)}$$

  - Take $k \geq \frac{4\ln n}{\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3}}$

שיעור 10

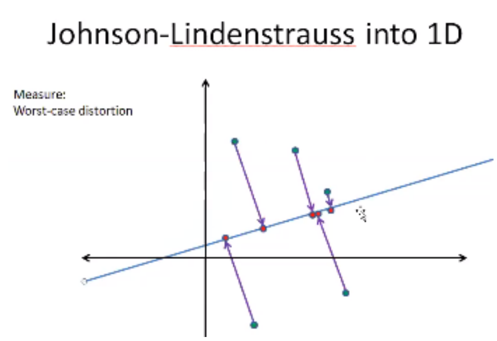
מאגרים $kaggle$, uci

Given a set of Euclidean vectors $V \subseteq R^m$ of size n, ther exits a linear function $f : V \to R^k$ for $k = \frac{8\ln n}{\varepsilon^2}$

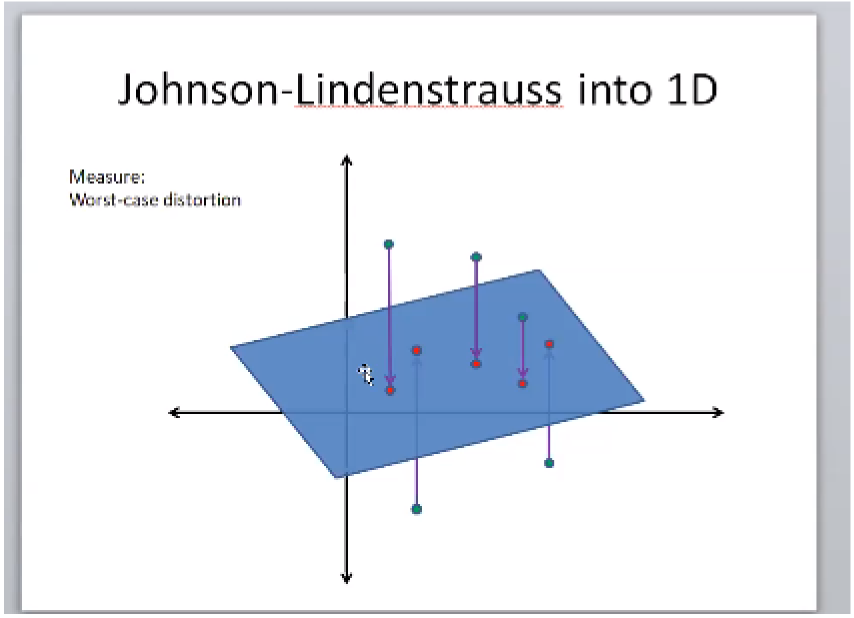
such that for all $y, z \in V$

$$(1-\varepsilon)\|y-z\| \leq \|f(y) - f(z)\| \leq (1+\varepsilon)\|y-z\|$$

הדגמה גיאומטרית:



שיכון לתוך מישור של 2 מימדים:



נרצה למצוא חסם k יותר טוב:

Lower-bound for JL

- How many dimensions are necessary to achieve distortion $(1+\varepsilon)$?
  - Can we better than $k = \frac{(8lnn)}{\varepsilon^2}$?
- Easy to show that l_2 requires $\Omega(\log n)$ dimensions
  - Volume argument on the basis vectors

$$(\frac{1}{\sqrt{2}}, 0, 0), (0, \frac{1}{\sqrt{2}}, 0), (0, 0, \frac{1}{\sqrt{2}})$$

  - Reduce dimension, contraction at most $\frac{1}{2}$
  - So every point has distance $1/2$ to every other point $\Rightarrow$ is the center of non-intersection balls of radius $\frac{1}{4}$.

$$\frac{V(d,1)}{V(d,\frac{1}{4})} = \frac{1}{(1/4)^d} = 4^d$$

- so must take $d > \log n$

–

- How many dimensions are necessary to achieve distortion $(1+\epsilon)$?
  - Dependence on logn unavoidable
  - What about dependence on ϵ^2? Noga Alon showed a lower-bound of:

$$k = \Omega\left(\frac{logn}{\epsilon^2 \log \frac{1}{\varepsilon}}\right)$$

- This leaves a gap of $\log\left(\frac{1}{\varepsilon}\right)$ between upper-bound and lower-bound

- Resolving this gap was a long-standing open problem...
- How many dimensions are necessary to achieve distortion $(1+\varepsilon)$?
- In 2017, Jelani Nelson and Kasper Green-Larson showed a lower-bound of:

$$k = \Omega\left(\frac{\log n}{\epsilon^2}\right)$$

- This resolved the gap, and JL is tight up to constants

Lower-bounds for other l_p

- How many dimensions are necessary to achieve distortion $(1+\epsilon)$, or even 2?
- l_1 requires $\Theta(n)$ dimensions
  - Andoni, Charikar, Neimann, Nguyen
- l_∞ requires $\Theta(n)$ dimensions
  - Proof:....
- Open problem for other values of p.
  - Known upper-bound $O(n^2)$

PCA - Principle Component Analysis

Given point set X (n points in d dimensions), example with 2 points and 4 dim':

$$x_1 = \begin{pmatrix} 1 \\ 4 \\ 2 \\ -1 \end{pmatrix}, x_2 = \begin{pmatrix} 2 \\ 3 \\ 1 \\ 4 \end{pmatrix}$$

and the set P_k of all $n \times n$ matrices of rank at most $k \ll n$,

$$\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} \Rightarrow rank = 1 < 2$$

how we choose the matrices:

Given point set X (n points in d dimensions = all points)

and the set P_k of all $n \times d$ matrices of rank at most k, find

$$\min_{P \in P_k} \|PX - X\|_F^2$$

where

$$\|A\|_p = \sqrt{\sum_i \sum_j |a_{i,j}|}$$

frobenius norm

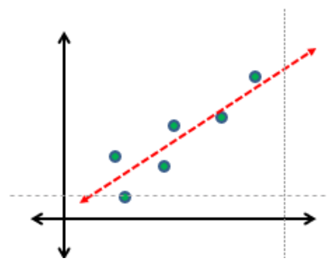
(how to pick the matrix?) Solution: compute eigenvectors and eigenvalue of X. choose the k eigenvaectors with highest eigenvalues.

Multiclass

- So far, we've done classification with only two classes {0,1} or {-1,1}.
- What about multiclass? K classes, {1,2,3}
- Reduce to two classes:
  - One-vs-one problems: k^2
  - One-vs-all problems: k

Regression

- Labels in real range Example: {0,M} or {-M,M} for some M
- Predict labels of unknown points



- In 1801, Giuseppe Piazzi observed the dwarf planet Ceres before its transit to the sun
  - Where will it appear after the transit?
- Gauss: Linear regression

Process:

- Predict labels of unseen points with continuous labels, for example [0,M]
- Want
  - Simple function $h \in H$
  - Small loss: $L\left(l\left(x_i\right), h\left(x_i\right)\right) = \left|l\left(x_i\right) - h\left(x_i\right)\right|^p$
  - $p \geq 1$. $p = 2$ is linear regression.
  - Minimize generalization error
  $$e\left(h\right) = E\left[L\left(l\left(x_i\right), h\left(x_i\right)\right)\right]$$
  where $x \sim D$.

Generalization bounds for regression

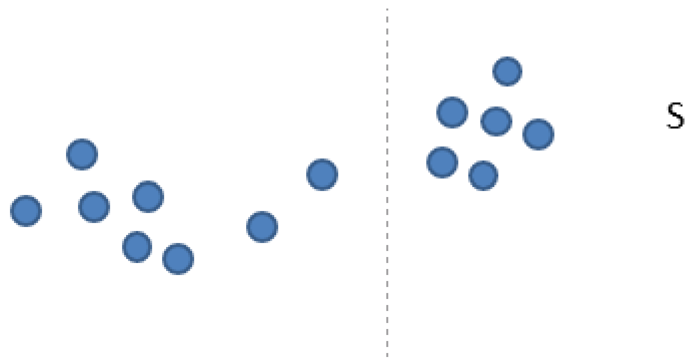
- Theorem: Let H be a finite hypothesis set, and let the range of labels be $[0, M]$. Let S be a sample of size m. Then for any $0 < \delta < 1$ with probability at least $1 - \delta$, the following holds:
- For all $h \in H$:
$$e\left(h\right) \leq \hat{e}\left(h\right) + M\sqrt{\frac{\log\left|H\right| + \log\frac{2}{\delta}}{2m}}$$

שיעור 11

clustering

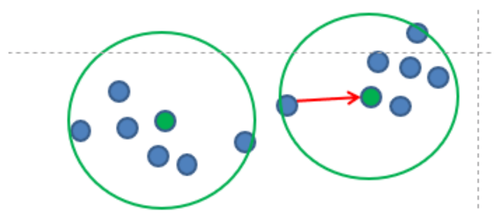
- Problem: Given a sample S of m points, cluster S into k groups.

  Question of measure: What's a good clustering?

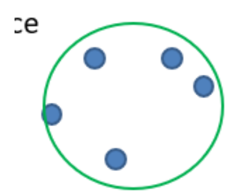


k-center clustering

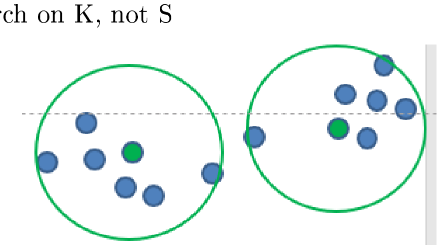
- Problem: Given a sample S of m points, cluster S into k groups
- k-center: Choose a set of k centers $K \subset S$ that minimize $\max_{v \in S} d(v, K)$
  - Distance from v to closest point in K



- This is the discrete version (K subset of S)
  - Non-discrete: centers chosen from ambient Euclidean space

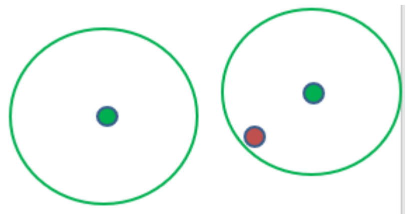


- Goals for clustering:
  - Learning: A new point can be assigned a group
  - Compression: May only need to retain K, not S
  - Run-time: nearest neighbor search on K, not S



- Learning bounds via compression.
  - Example: m^k possible rules. If m points are shattered,

$$m^k \geq 2^m$$
$$k \log m \geq m$$
$$m = O\left(k \log k\right)$$



- Exact algorithm for discrete k-center:
  - Brute force:
  - try all sets of k in $O(m^k)$ time Can we do better?
- Finding optimal k-center clustering is NP-hard
  - When either k or dimension d is large
  - Also NP-hard to approximate radius within factor $2 - \epsilon$

- Proof (of both statements): reduction from Minimum Dominating Set
  - Given graph $G = (V, E)$, find minimum sized subset $V' \subset V$
  - Any vertex $v \in V - V'$ is adjacent to some vertex in V'

k-center approximation

- K-center is hard to solve exactly
- Two possible polynomial-time approximation algorithms for the discrete case:
  - Optimal radius, but more centers
    - Can't do better than logn-approximation
  - k centers, but larger radius
    - Can't do better than twice the radius

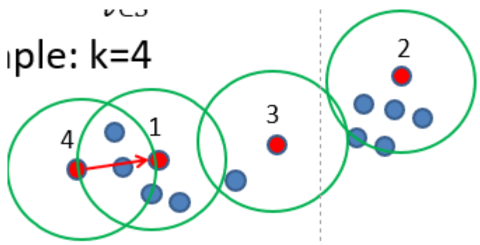
algorithms

Second approximation algorithms:

- k centers, but larger radius
- Greedy algorithm:
  - Choose the farthest uncovered point

$$\max_{(v\in S)} d(v,K) = \max_{v\in S}\min_{w\in k} d\,(v,w)$$

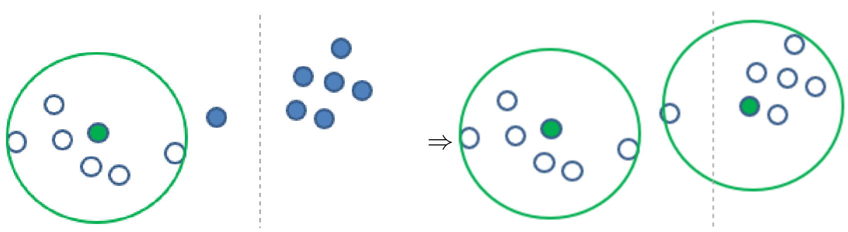
  - Example:$k=4$



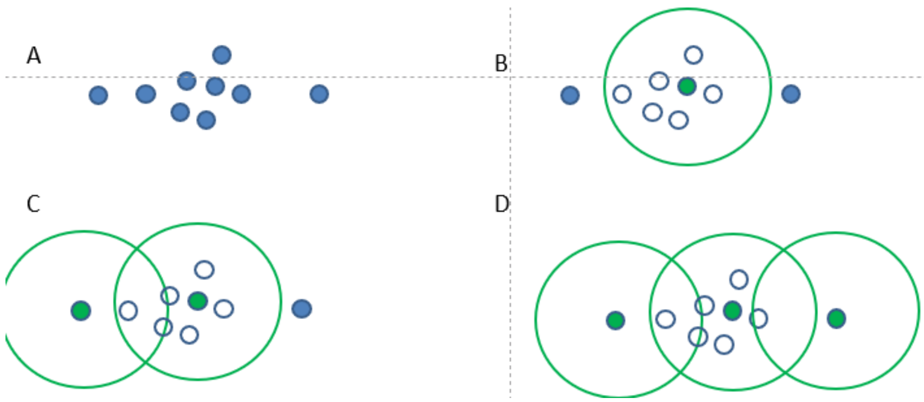
- Claim: radius r at most twice optimal
  - No center is found in more than one ball – distance is greater than r
  - k balls of radius $^{r}/_{2}$ needed to cover them.

First approximation algorithm:

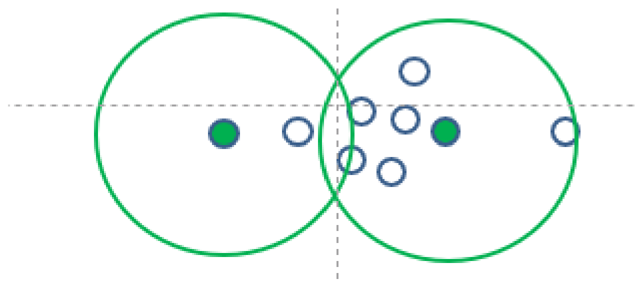
- Optimal radius, but (k ln m) centers
- Greedy algorithm:
  - Assume optimal radius r is known (m^2 guesses)
  - Take center which covers most points
  - Remove covers points and repeat



- Greedy not necessarily optimal
  - Greedy:



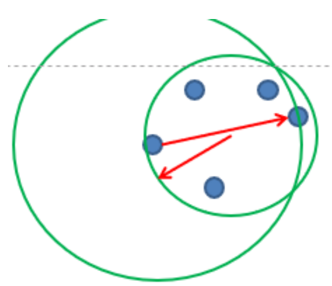
- but Optimal:



- Standard greedy analysis:
  - k centers cover S, and also any subset of S
  - Some center covers a $\frac{1}{k}$ fraction of S or its subset
- At every step, a $1/k$ fraction of points are covered.
  - After i iterations, $(1-\frac{`1}{k})^i m \leq e^{-\frac{i}{k}} m$ points remain
- Max i = k ln m iterations.
  - So $(k \ln m)$ centers, instead of k

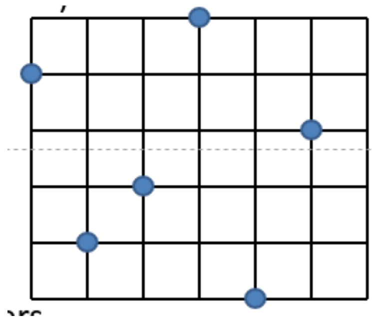
<u>What about the non-discrete case? (Centers come from ambient space, not S)</u>

- Can just solve the discrete case Lose a factor 2 in the radius.



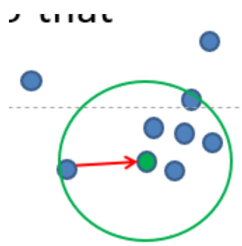
- Can also take an ϵ-grid
  - $(\frac{\sqrt{d}}{\varepsilon})^d$gridpoints which can be centers Can

- preprocess: reduce dimension to $d = \frac{\log n}{\epsilon^2}$ But runtime still large



איך בודקים מהו פתרון טוב?

- Problem: Given a sample S of m points, cluster S into k groups
- k-center: Choose a set of k centers $K \subset S$ that minimize $\max\limits_{(v \in S)} d(v, K)$
  - Distance from v to closest point in K
  - Problem: Not robust to outliers



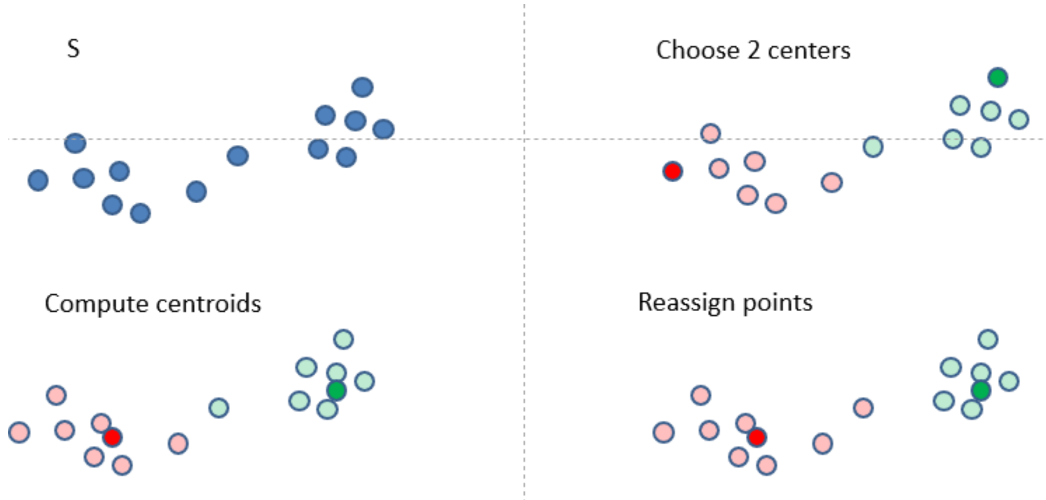
- Other measures k-median:
  - Choose K to minimize $\frac{1}{m} \sum\limits_{v \in S} d(v, K)$
  - k-means: Choose K to minimize $\frac{1}{m} \sum\limits_{v \in S} d^2(v, K)$

k-means algorithm

Algorithm for non-discrete k-means Variant of Lloyd's algorithm

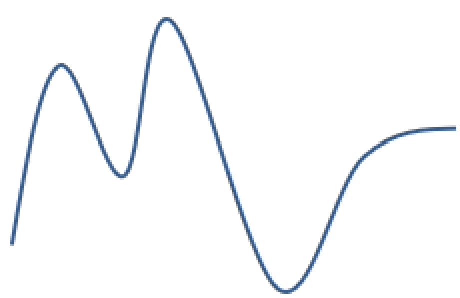
1. Pick k centers arbitrarily
   - Many papers on this choice - seeding
2. Assign each point in S to closest center
3. For each cluster C, compute centroid
   - Centroid $= \frac{1}{|C|} \sum\limits_{x \in C} x$
4. Assign each point in S to closest centroid
5. Repeat 3,4 until no change

Example:



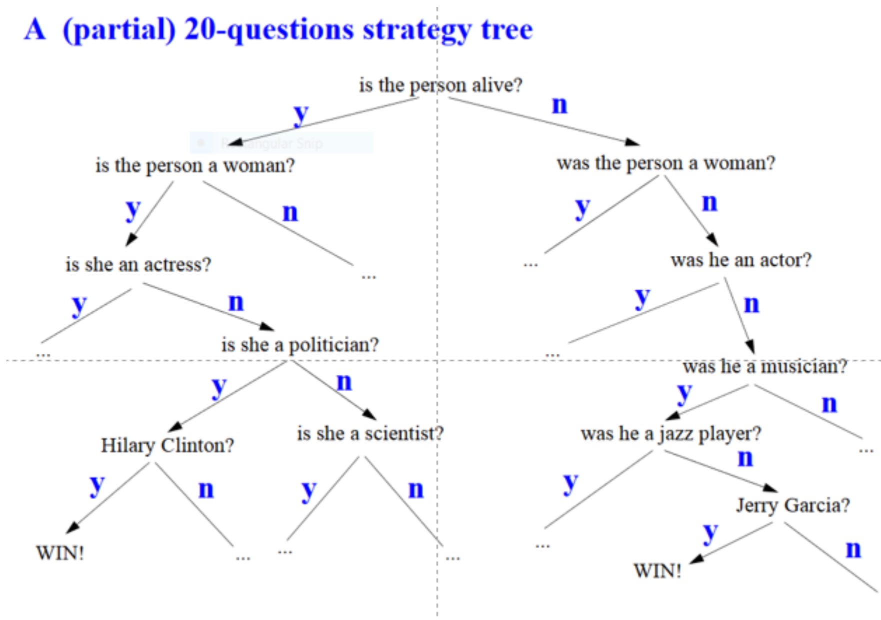
<u>k-means</u>

- What guarantees does this algorithm have?
- Iterations: worst-case $2^{\sqrt{m}}$
  - Average case much better
- Quality of solution: local minima
  - Why seeding is important



שיעור 12

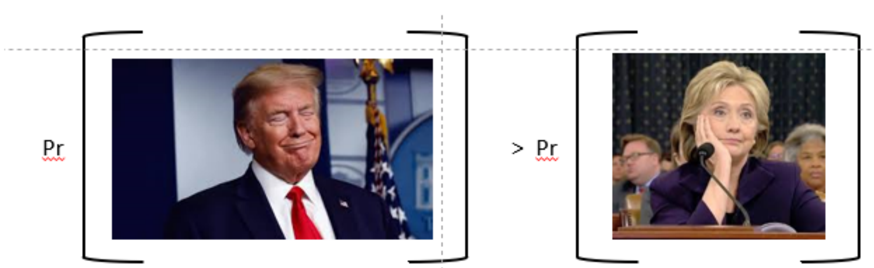
Decision trees



- A binary tree with 20 levels has
  - $2^{20} = 1,048,576$ leaves.
  - That's much larger than the number of people I know of.
- So you should be able to win the game every time
  - At least if you have a good splitting strategy

A different game

- What if my goal is to ask the smallest number of questions on average?
  - Minimize expected number of guesses: $Pr\left[x_i\right] * \text{number\_of\_guess}\left[x_i\right]$
- For example
  - in our game above, suppose some people are much more likely to be chosen than others...



  - Up next!
    - Long detour into information theory.

Information thoery

- Familiar compression schemes:
  - zip, 7z, rar, jpeg.
- How compressed can a file get?
  - The field of information theory studies this.
- Let an **item** be (for example) a letter in a long text...
- Founding theorem of information theory:
- Shannon (1948): If item i has frequency w_i then the optimal compression of a message of length m is of size (bits) - entropy function:

$$m \sum_i w_i \log_2 \frac{1}{w_i}$$

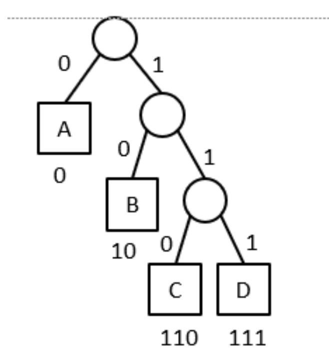
- Fine print: assumes Each item is encoded in a finite string representation of each item cannot change during encoding.

Entropy function

- Measures how random a set is
- Entropy function for item/event e_i: $\sum_i Pr\left(e_i\right) \log \left\lceil \frac{1}{pr(e_i)} \right\rceil$
  - For example, a fair coin has entropy 1, and a coin that always falls on heads has entropy 0.
  - Binary entropy function for item/events e_1, e_2

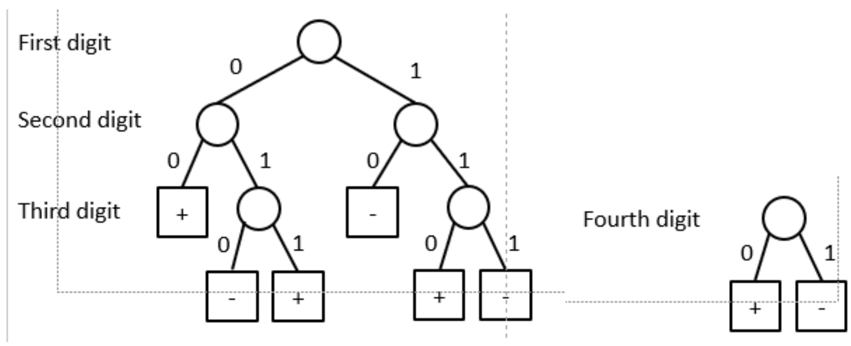
$$\begin{aligned} & Pr\left[e_i\right] \log \left\lceil \frac{1}{Pr(e_i)} \right\rceil + Pr\left[e_2\right] \log \left\lceil \frac{1}{Pr(e_2)} + \right\rceil \\ = & Pr\left[e_i\right] \log \left\lceil \frac{1}{Pr(e_i)} \right\rceil + \left(1 - Pr\left[e_1\right]\right) \log \left\lceil \frac{1}{1 - Pr[e_1]} \right\rceil + \end{aligned}$$

- Shannon's theorem is non-constructive
  - He didn't give a code that met his bounds
- Huffman (1952): bottom-up construction realizes optimal bound
  - Give shorter codes to more frequently occurring items
  - First example, BAABAC = 010000010010
  - Second example BAABAC = 1000100110



Learning with decision trees

- We want to learn with decision trees.
  - Example: sample of binary vectors $\left[\begin{array}{c|c} 0000 & + \\ 0001 & + \\ 0101 & - \\ 0110 & + \\ 1001 & - \\ 1011 & - \\ 1100 & + \\ 1111 & - \end{array}\right]$
- Split rule: value in each dimension



- What's the VC-dimension of decision trees on binary d-dimensional vectors ?
  - T^k – family of decision trees with k nodes
  - At each node, choose dimension to test, or mark it as a leaf in {+,-} – $d+2$ decisions
  - $(d+2)^k$ trees in T^k
  - $VC-dim(T^k) = O(k \log d)$
- So we want small k.
- Problem: find decision tree which is consistent with the data and has the smallest number of nodes.
  - NP-hard.
- Heuristics used instead.
  - ID3
  - Pruning

ID3 algorithm

- Top-down:
  - Start with root as a + leaf.
  - Function **Cost** measures quality of solution

  - Split leaf with rule that minimizes Cost
  - Repeat until consistent
- What's a good Cost function
  - Minimize sum of entropy of leaves
- No actual guarantees on tree size

Pruning

- Another way to achieve small tree
  - Can do it after brute-force or ID3
- Remove nodes starting at the bottom. Options:
  - Do nothing
  - Make into leaf in {0,1}
  - Replace with subtree

Random forest

- State-of-the-art
- Compute many decision trees
  - ID3 maybe
- Take majority decision