ד' - 25/3/20

הרצאה 1

נעבור על:
1. הסתברות וסטטיסטיקה.
2. התפלגות במימדים גבוהים.
3. חסמים ואי"ש שקשורים להסתברות: מרקוב, צ'בישב, הופדינג.

1. הסתברות וסטטיסטיקה

(1) הסתברות

מרחב מדגם הוא הזוג (Ω, D) כך ש:

- Ω מרחב המדגם: קב' כל התוצאות האפשריות $\omega\in\Omega$. יכול להיות בדיד או רציף.
  - A מאורע: תת-קב' של תוצאות: $A\subseteq\Omega$.
  - A^c מאורע משלים: $A^c=\Omega\setminus A$, המאורע שהמאורע A לא התרחש.
  - מאורעות זרים: A,B מאורעות זרים אם: $A\cap B=\emptyset$.
  - ★ אם A,B מאורעות, אזי גם: $A\cap B$, $A\cup B$.
- D פונק' הסתברות: $D:2^{\Omega}\to\mathbb{R}$ עבור Ω מרחב המדגם שלו היא מוגדרת, המקיימת:
  1. $D(\Omega)=1$.
  2. לכל $\omega\in\Omega$: $D(\omega)\in[0,1]$.
  3. לכל $A,B\subseteq\Omega$ מאורעות זרים: $D(A\cup B)=D(A)+D(B)$.

טענה:
תרגיל 1: $\forall$ לכל $A,B\subseteq\Omega$, הראה: $D(A\cup B)=D(A)+D(B)-D(A\cap B)$

הוכחה: נבחין שמתקיימים הבאים: $A\cup B=A\setminus B\cup B\setminus A\cup A\cap B$,

$A=A\setminus B\cup A\cap B$, $B=B\setminus A\cup A\cap B$

לכן: $D(A\cup B)=D(A\setminus B)+D(B\setminus A)+D(A\cap B)=$

$$=(D(A)-D(A\cap B))+(D(B)-\cancel{D(A\cap B)})+\cancel{D(A\cap B)}=D(A)+D(B)-D(A\cap B)$$

$\boxtimes$

- מאורעות בת"ל: $A,B\subseteq\Omega$ בת"ל אם: $D(A\cap B)=D(A)D(B)$ (התרחשות אחד מהם לא משפיעה על התרחשות השני).

טענה:
תרגיל 2: אם $A,B\subseteq\Omega$ בת"ל, אז גם A,B^c בת"ל.

הוכחה: $D(A)=D(A\cap B)+D(A\cap B^c)\Leftarrow A=(A\cup B)\cup(A\cup B^c)$

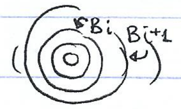
ולכן: $D(A\cap B^c)=D(A)-D(A\cap B)\overset{A,B\text{ בת"ל}}{=}D(A)-D(A)D(B)=$

$$=D(A)(1-D(B))=D(A)\cdot D(B^c)$$

$\boxtimes$

למה (חסם האיחוד): יהא (Ω, D) מרחב הסתברות, אזי D מקיימת
לכל רצף מאורעות (A_k): $D(\bigcup_{k=1}^{\infty} A_k) \le \sum_{k=1}^{\infty} D(A_k)$

הוכחה: נגדיר סדרת מאורעות חדשה (B_k) כך:



$B_1 = A_1$ ולכל $k \ge 2$: $B_k = A_k \setminus (\bigcup_{i=1}^{k-1} A_i)$.

נבחין שמתקיים: $\bigcup_{i=1}^{k} A_i = \dot{\bigcup}_{i=1}^{k} B_i$, וכן ש-$B_k$ זרים.

כמו כן: $B_k \subseteq A_k$, לכן: $D(B_k) \le D(A_k)$ לכל k. אזי:

$$D(\bigcup_{k=1}^{\infty} A_k) \underset{\text{שוויון איחודים } B_k, A_k}{=} D(\dot{\bigcup}_{k=1}^{\infty} B_k) \underset{B_k \text{ זרים}}{=} \sum_{k=1}^{\infty} D(B_k) \underset{B_k \text{ תת-קב' } A_k}{\le} \sum_{k=1}^{\infty} D(A_k)$$

■

X משתנה מקרי: בהינתן מרחב הסתברות (Ω, D), משתנה מקרי הוא $X: \Omega \to \mathbb{R}$.
(ולכן גם כל פונק' של X הוא משתנה מקרי (X^2 למשל)).

- אם Ω מרחב בדיד, אזי X **מ"מ בדיד** מוגדר: $D(\{X=x\}) = \sum_{\omega: X(\omega)=x} D(\omega)$
- אם Ω מרחב רציף, אזי X **מ"מ רציף** מוגדר: $D(X \in D) = \int_D f(x)dx$

עבור $f(x) \ge 0$ כלשהי, לכל $D \subseteq \mathbb{R}$, ובפרט: $D(X \in [a,b]) = \int_a^b f(x)dx$
עבור $a,b \in \mathbb{R}$, $a \le b$. בקורס זה נניח ש-f פונק' רגולרית.
ולכן $D(X=a) = \int_a^a f(x)dx = 0$ לכל $a \in \mathbb{R}$.

☆ הפונק' f (נקראת פונק' צפיפות ההסתברות (PDF)) צריכה לקיים:
1. $f(x) \ge 0$.
2. $\int_{-\infty}^{\infty} f(x)dx = 1$.
3. $f(x)$ פונק' צפיפות, לא הסתברות (ולכן יכולה לקיים $f(x)>1$).

E תוחלת: עבור X משתנה מקרי בדיד: $E[X] = \sum_{x \in X} x D(x)$
עבור X משתנה מקרי רציף: $E[X] = \int_{-\infty}^{\infty} x f(x)dx$

התוחלת מקיימת את התכונות הבאות:
1. $E[g(X)] = \sum g(x) D(x)$
2. $E[c] = c$: $c \in \mathbb{R}$
3. ליניאריות: $E[aX+bY] = aE[X] + bE[Y]$: $a,b \in \mathbb{R}$
4. אם X,Y בת"ל אזי: $E[XY] = E[X]E[Y]$

V שונות: עבור מ"מ X עם תוחלת $E[X] = \mu$, נגדיר את השונות שלו להיות: $V(X) = E[(X-\mu)^2]$

השונות הסטנדרטית σ שלו מוגדרת להיות: $\sigma = \sqrt{V(X)}$

תכונות השונות:

1. $V(X) \geq 0$, ו- $V(X)=0$ $\Leftrightarrow$ X קבוע
2. $V(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
3. $V(aX+b) = a^2 V(X)$
4. $V(X+Y) = \mathbb{E}[X+Y-\mathbb{E}(X)-(\mathbb{E}(Y))^2] =$

$$= V(X) + V(Y) + 2\underbrace{\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]}_{Cov(X,Y)} = V(X)+V(Y)+2Cov(X,Y)$$

5. אם $X_1,..,X_n$ מ"מ בת"ל אז: $V(\sum X_i) = \sum V(X_i)$

$Cov(X,Y)$: $Cov(X,Y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]$

תכונות: 1. סימטריה: $Cov(X,Y) = Cov(Y,X)$

2. $Cov(aX+b, cY+d) = a \cdot c \cdot Cov(X,Y)$

3. $Cov(X,X) = V(X)$

(2) סטטיסטיקה

במקרה כזה: $Var[\sum x_i] = \sum Var(x_i)$

(i.i.d = independent and identically distributed)

ננסה ללמוד את העולם בהתבסס על דוגמאות. לרוב נותנים לנו מ"מ X עם התפלגות בלתי-ידועה אותה נסמן ב- P_X , ו- n דגימות בת"ל ידועות מההתפלגות הזו: $X_1,..,X_n \overset{i.i.d}{\sim} P_X$

בהינתן המדגם הזה נרצה לשערך תכונות של ההתפלגות: (תכונה משוערכת תסומן עם כובע מעליה ($\hat{\square}$). σ = שונות, μ = ממוצע)

- Sample mean: $\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Sample variance: $\hat{\sigma}^2_X = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\hat{\mu}_X)^2$

למה $\frac{1}{n-1}$ ולא $\frac{1}{n}$? $\hat{\mu}, \hat{\sigma}^2$ הם בעצמם מ"מ (כפונק' של מ"מ)!
ולכן אפשר לחשב תוחלת עבורם. מתקיים: $\mathbb{E}[\hat{\sigma}^2_X] = V(X)$,
לעומת עם $\frac{1}{n}$ שמחזיר: $\frac{n-1}{n}V(X)$ ... לכן צריך את הפקטור,
כדי שיקרב טוב יותר לגודל האמיתי.

★ משערכים שמקיימים את התכונה שתוחלתם שווה לשונות של המשתנה אותו הם מנסים לשערך, נקראים משערכים לא משוחדים. ($\hat{\sigma}^2$ לא משוחד, ה- $\frac{1}{n}$ לעומת זה כן)

2. התפלגות במימדים גבוהים.

(1) חזרנו להסתברות 🙂 עכשיו נחשוב על המ"מ כעל וקטורים ב-$\mathbb{R}^d$ במקום כעל סקלרים ב-$\mathbb{R}$.

- וקטור אקראי (מ"מ וקטורי): אוסף מ"מ סדור: $\overbrace{X=(X_1,\dots,X_n)^T}^{\text{שכולם מעל אותו מרחב הסתברות } (\Omega,P)}$
- התפלגות משותפת: של וקטור רנדומי מוגדרת ע"י ההסתברות שכל אחד מרכיבי הוקטור $X_1,\dots,X_n$ נופלים באינטרוול טווח ערכים (עבור מ"מ וקטורי רציף) או באינטרוול 00 ערכים (עבור מ"מ בדיד).
- פונק' הסתברות משותפת של וקטור אקראי: ההתפלגות המשותפת נקבעת ע"י פונק' ההסתברות המשותפת. ב-מ"מ אקראיים X,Y הם רציפים ומשותפים אם קיימת פונק' $0 \le f_{XY}:\mathbb{R}^2\to\mathbb{R}$, כך שלכל $A\in\mathbb{R}^2$ מתקיים: $P((X,Y)\in A)=\int_A f_{XY}(x,y)\,dx\,dy$. f_{XY} נקראת פונק' ~~צפיפות~~ ההסתברות המשותפת של X ו-Y.

★ המידע על כל רכיב $X_1,\dots,X_n$ בנפרד לא מספיק כדי להבין את התנהגות X כולו.

ההתפלגות הנורמלית:

א- משפט הגבול המרכזי: סכום מנורמל של מ"מים ב"ת מתכנס להתפלגות נורמלית. וכן יש לה תכונות נוחות.

נזכר בהתפלגות נורמלית במקרה הסקלרי: למ"מ X יש התפלגות נורמלית עם תוחלת μ ושונות σ^2 אם פונק' צפיפות ההתפלגות שלו היא:

$$f(x)=\frac{1}{\underbrace{\sqrt{2\pi\sigma^2}}_{\text{פקטור נירמול שדואג ש: } \int f=1}}\cdot e^{-(x-\mu)^2/2\sigma^2}$$

סימון: $X\sim N(\mu,\sigma^2)$.

★ שיא הגרף הוא בערך התוחלת, והפונק' סימטרית. לכן $\underline{\underline{\mu}}$ = פיק, שינוי $\underline{\underline{\sigma^2}}$ משנה את הרוחב (ככל ש-σ^2 גדל, הפונק' משתטחת ומתרחבת) והשטח (האינטגרל) נשאר 1.

במקרה הרב-מימדי (Multivariate normal distribution):

לוקטור הרנדומי X יש התפלגות נורמלית רב-מימדית עם תוחלת μ (וקטור) ומטריצת שונות Σ אם קיימת פונק' הסתברות משותפת מהצורה:

$$f(x)=\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\exp\left\{-\frac{\overbrace{(x-\mu)^T\Sigma^{-1}(x-\mu)}^{\text{תבנית ריבועית שמגדירה סקלר}}}{2}\right\}$$

סימון: $X\sim N(\mu,\Sigma)$

★ הוקטור μ שולט על מיקום פסגת העקומה, הצורה (שטוח/קמור מול צר/רחב) נשלטת ע"י המטריצה Σ: ע"ע גדולים = כובע רחב, ע"ע קטנים = כובע צר - כיוונים המאונכים (לפי ו"ע)

ההתפלגות השולית: של התפלגות משותפת $P_{X,Y}$:

$$P_X(x) = \int_y P_{X,Y}(x,y)\,dy$$

תרגיל 4: יהא $X \sim N(\mu,\Sigma)$ עבור $\mu=(\mu_1,\mu_2)$, $\Sigma = \begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix}$

נמצא את ה-PDF של ההתפלגות השולית של X_1:

פתרון:

$$f(x) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\cdot \exp\left\{-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right\} =$$ הצבת Σ וחישוב $|\Sigma|$

$$= \frac{1}{\sqrt{(2\pi)^2\sigma_1^2\sigma_2^2}}\cdot\exp\left\{-\frac{(x-\mu)^T\begin{pmatrix}\sigma_1^{-2}&0\\0&\sigma_2^{-2}\end{pmatrix}(x-\mu)}{2}\right\} =$$ חישוב התבנית הריבועית

$$= \frac{1}{\sqrt{(2\pi)^2\sigma_1^2\sigma_2^2}}\cdot\exp\left\{-\left(\frac{x_1-\mu_1}{\sqrt{2}\sigma_1}\right)^2-\left(\frac{x_2-\mu_2}{\sqrt{2}\sigma_2}\right)^2\right\} =$$ הפרדת ה-exp לרכיבים

$$= \frac{1}{\sqrt{(2\pi)^2\sigma_1^2\sigma_2^2}}\cdot\exp\left\{-\left(\frac{x_1-\mu_1}{\sqrt{2}\sigma_1}\right)^2\right\}\cdot\frac{1}{\sqrt{(2\pi)^2\sigma_1^2\sigma_2^2}}\cdot\exp\left\{-\left(\frac{x_2-\mu_2}{\sqrt{2}\sigma_2}\right)^2\right\} =$$

כעת:

$$I_{x_2} = \int_{-\infty}^{\infty}\frac{1}{(2\pi)^2\sigma_2^2}\cdot\exp\left\{-\left(\frac{x_2-\mu_2}{\sqrt{2}\sigma_2}\right)^2\right\}dx_2 = 1$$

כי זו ההתפלגות הנורמלית הסקלארית של x_2, לכן:

$$\int_{-\infty}^{\infty}\frac{1}{(2\pi)^n|\Sigma|}\exp\left\{-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right\}dx_2 =$$

$$= \frac{1}{\sqrt{(2\pi)\sigma_1^2}}\exp\left\{\left(\frac{x_1-\mu_1}{\sqrt{2}\sigma_1}\right)^2\right\}\cdot I_{x_2} = \frac{1}{\sqrt{(2\pi)\sigma_1^2}}\cdot\exp\left\{-\left(\frac{x_1-\mu_1}{\sqrt{2}\sigma_1}\right)^2\right\}$$

אז קיבלנו: $X_1 \sim N(\mu_1,\sigma_1^2)$ (במקרה הסקלרי 🙂) ■

• **Σ מטריצת ה-Cov**: יהא $X=(X_1,\dots,X_d)^T$ וקטור אקראי. נגדיר את מטריצת ה-Cov Σ כמטריצה $d\times d$ שמקיימת: $\Sigma_{i,j}=Cov(X_i,X_j)$.

כלומר: $\Sigma = E[(X-E[X])(X-E[X])^T]$. **תכונות**:

- איברי האלכסון של Σ הם $V(X_i)$. ($V=Cov(X,X)$ מהגדרת).
- Σ סימטרית (כי: Cov סימטרי), והיא ≥ 0 (שוב מתכונות Cov).

(2) **נקפוץ לסטטיסטיקה**:

נשערך את מטריצת ה-Cov Σ של מ"מ וקטורי, נקרא: V population cov. matrix

- במקרה הדו-מימדי: $X=(X_1,X_2)$ עבור m דגימות נסמן: $X_1^1,\dots,X_1^m$, $X_2^1,\dots,X_2^m$. נסדר אותן במטריצה $M\in\mathbb{R}^{d\times m}$, כאשר m זה הדגימות ו-d זה המימד. כאן $d=2$.

ניתן להכליל מעבר ל $d=2$

• Sample variance: ל-$i\in\{1,2\}$: $\hat{\sigma}^2 x_i = \frac{1}{m-1}\sum_{j=1}^{m}(x_{ij}-\hat{\mu}_i)^2$

• Sample covariance: $\hat{\sigma}(x_1,x_2) = \frac{1}{m-1}\sum_{j=1}^{m}(x_{1,j}-\hat{\mu}_1)(x_{2,j}-\hat{\mu}_2)$

המטריצה המשוערכת: $\hat{\Sigma}_{i,j}=\hat{\sigma}(x_i,x_j)$ או $\hat{C} = \frac{1}{m-1}\sum_{i=1}^{m}(x_i-\bar{X})(x_i-\bar{X})^T$

עבור $\bar{X}\in\mathbb{R}^{d\times 1}$ וקטור של הממוצעים.

או שאפשר לחשוב עלי כמטריצה: $\tilde{X}\in\mathbb{R}^{d\times m}$ כאשר: $\hat{C}=\frac{\tilde{X}\tilde{X}^T}{m-1}\in\mathbb{R}^{d\times d}$

תרגיל 5: $X^T=\begin{pmatrix}150 & 45\\ 170 & 74\\ 184 & 79\end{pmatrix}$, 3 דגימות. נחשב את $\hat{C}$:

height ← → weight

פתרון: $\hat{\mu}_1=168$, $\hat{\mu}_2=66$. אז: $\bar{X}=\begin{pmatrix}168 & 66\\ 168 & 66\\ 168 & 66\end{pmatrix}$ ולכן:

$$X^T_{\text{ממורכז}} = X^T-\bar{X} = \begin{pmatrix}-18 & -21\\ 2 & 8\\ 16 & 13\end{pmatrix} \Rightarrow \hat{C} = \frac{X_{\text{ממורכז}}X^T_{\text{ממורכז}}}{3-1} = \begin{bmatrix}292 & 301\\ 301 & 337\end{bmatrix}$$

(3) איך טרנספורמציות ליניאריות משפיעות על datasets?

נקח 2 מ"מ ממורכזים (תוחלת=0) ושונות זהה (נסמנה σ^2) והם לא מתואמים (אין ביניהם קשר..). אם נצייר מ' נקודות כך תראה הצפיפות: מעגלים ממורכזים ב-0 וצפיפות קטנה כשמתרחקים

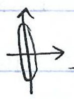
המטריצה תהיה: $C=\begin{pmatrix}\sigma_1^2 & \overset{cov}{0}\\ 0 & \sigma_2^2\end{pmatrix}$

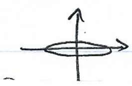
נפעיל עליה טרנספורמציות ליניאריות:

1. Scaling (מתיחה של הצירים): $S=\begin{pmatrix}s_1 & 0\\ 0 & s_2\end{pmatrix}$. הכפלה של C ב-S:

$$\frac{SX(SX)^T}{n-1} = S\underbrace{\frac{XX^T}{n-1}}_{C}S^T = SCS^T = \begin{pmatrix}(s_1\sigma_1)^2 & 0\\ 0 & (s_2\sigma_2)^2\end{pmatrix}$$

← אלכסונית ולא סקלרית (אם $s_1\neq s_2$)

★ אם $s_1<s_2$ נקבל אליפסה ממורכזת  צפיפות גבוהה ב-(0,0) שקטנה ככל שמתרחקים - קטנה אטי בציר ה-y ומהר בציר ה-X.

★ אם $s_1>s_2$: .

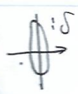
המשתנים עדיין לא מתואמים

2. סיבוב: $T=RS$ עבור $R=\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}$

אז:

$$\frac{(RSX)(RSX)^T}{n-1} = RS\frac{XX^T}{n-1}S^TR^T = RSCS^TR^T =$$

$$= R\begin{pmatrix}(s_1\sigma)^2 & 0\\ 0 & (s_2\sigma)^2\end{pmatrix}R^T = \sigma^2\times\begin{pmatrix}s_1^2\cos^2\theta+s_2^2\sin^2\theta & \sin\theta\cos\theta(s_1^2-s_2^2)\\ \sin\theta\cos\theta(s_1^2-s_2^2) & s_1^2\sin^2\theta+s_2^2\cos^2\theta\end{pmatrix}$$

★ אם $s_1=s_2$: עשינו Scaling אחיד (כיווץ/ניפוח) והאיברים שאינם על האלכסון מתאפסים. לכן המשתנים לא מתואמים במצב הזה. ונקבל דומה ל: 

★ אם $s_1 \neq s_2$ אז ה- Scaling אסימטרי ולכן ייראה
כשהסיבוב נובע מהזווית θ - הציר הארוך $\underbrace{\text{בזווית}}_{}$ 45° מטה מהקצר,
והציר הקצר מאונך לציר השני. הכיוונים הללו הם הו"ע של המטריצה הזו.

3. חסמים ואי"ש שקשורים להסתברות

רוצים לתת אינדיקציה לכמה טוב השערוך שלנו: כמה הם מתפלגים טוב סביב הערך אותו נרצה לשערך. נצפה לראות שהשונות סביב הערך האמתי תקטן ככל שנדגום יותר דגימות, השאלה היא איך זה מגדל?

- נניח ש: $\hat{p} = \frac{\#\text{מופעים}}{n}$, אז: $E[\hat{p}] = p$,

אך לזה מתווספת שגיאה כלשהי: ε, בגלל שאנחנו דוגמים דגימה סופית.

- "גם שהמדגם שדגמנו לא מייצג טוב את p האמיתית. - לא ט' סופי של דגימות.

⇐ מגדירים פרמטרים: • Accuracy (ε): מתאר את השגיאה. ($\varepsilon \in (0,1)$)
• confidence (δ): $\delta \in (0,1)$

ואנחנו רוצים למצוא כמה דגימות נחוצות לנו כדי להבטיח שהשערוך שלנו יהיה בעל שגיאה אבסולוטית $\le \varepsilon$, בהסתברות $\ge 1-\delta$: נקרא (PAC): probably approximatly correct

נרצה לתת חסמים על גודל הדגימה כדי שהאמור למעלה יתקיים.
ניזכר בכמה חסמים:

(1) אי"ש מרקוב: לכל מ"מ אי"ש X עם $E[X]<\infty$ ו- $0<t\in\mathbb{R}$ מתקיים:
$$P[X \ge t] \le \frac{E[X]}{t}$$

הוכחה: תהא $f(x)$ פונק' הצפיפות של X. X אי"ש, ולכן:
$$E[X] \overset{\text{הגדרה}}{=} \int_{x\in\mathbb{R}} f(x)x\,dx \underset{X\text{ אי"ש}}{=} \underbrace{\int_{x=0}^{t} f(x)x\,dx}_{\ge 0} + \int_{x=t}^{\infty} f(x)x\,dx \ge \int_{x=t}^{\infty} f(x)x\,dx \ge$$
$$\ge \int_{x=t}^{\infty} f(x)t\,dt = t\int_{x=t}^{\infty} f(x)\,dx = t\cdot P[X\ge t]$$

ולכן: $P[X\ge t] = \frac{E[X]}{t}$, כנדרש. ■

מסקנה מאי"ש מרקוב: עבור m הגרלות i.i.d: $X_1,\dots,X_m$ (כאשר X מ"מ אי"ש עם $E[X]<\infty$), נגדיר את ממוצע המדגם: $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$,
אז עבור כל $t\in\mathbb{R}$, מתקיים: $P[\bar{X}\ge t] \le \frac{E[X]}{t}$.

הוכחה: $E[\bar{X}] = E[\frac{1}{m}\sum_{i=1}^{m} X_i] = \frac{1}{m}\sum_{i=1}^{m} E[X_i] = \frac{1}{m}\sum_{i=1}^{m} E[X] = E[X]$

(רציפה) (לינאריות) (לכל X_iים יש אותה תוחלת (כל X_i הוא עותק של X))

ומא"ש מרקוב: $P[\bar{X} \geq t] \leq \frac{E[\bar{X}]}{t} = \frac{E[X]}{t}$ ▨

(ב) א"ש צ'בישב: יהא X מ"מ עם $E[X]<\infty$, $Var[X]<\infty$ אז עבור כל $0<t\in\mathbb{R}$ מתקיים: $P[|X-E[X]| \geq t] \leq \frac{Var(X)}{t^2}$

הוכחה: נבחין ש: $P[|X-E(X)| \geq t] = P[(X-E(X))^2 \geq t^2]$ (הם מאורעות שקולים ולכן קורי אותה ההסתברות)

מכיוון ש: $E[X]\in\mathbb{R}$: $<\infty$ אז $X-E[X]$ מ"מ, ולכן גם $Y=(X-E[X])^2$ מ"מ (כפונק' של מ"מ) והוא גם אי שלילי ומתקיים: $E[Y] = E[(X-E(X))^2] = Var[X] < \infty$ נתון

ולכן ניתן להשתמש בא"ש מרקוב:

$$P[|X-E[X]| \geq t] = P[(X-E[X])^2 \geq t^2] \leq \frac{E[(X-E(X))^2]}{t^2} \underset{\text{הגדרת Var}}{=} \frac{Var[X]}{t^2}$$

▨

כעת, השונות לא לינארית. למה היא שווה? $X=\sum_{i=1}^{m} X_i \overset{iid}{\sim} D$ מ"מ מעל התפלגות D.

$$Var[\bar{X}] = Var\left[\frac{1}{m}\sum_{i=1}^{m} X_i\right] = \frac{1}{m^2} Var\left[\sum_{i=1}^{m} X_i\right] \underset{X_i \overset{iid}{\sim} D}{=} \frac{1}{m^2}\sum_{i=1}^{m} Var[X_i] \underset{\substack{\text{לכל } X_i \text{ אותה התפלגות}\\ \text{ולכן אותו ה-Var}}}{=} \frac{1}{m^2}\sum_{i=1}^{m} Var[X] = \frac{1}{m} Var[X]$$

⇐ נניח ש: $E[\bar{X}] = E[X]$ אך $Var[\bar{X}] = \frac{1}{m}Var[X]$ ולכן: $Var[\bar{X}] \xrightarrow[m\to\infty]{} 0$

מסקנה מא"ש צ'בישב: יהא X מ"מ עם $E[X], Var[X] < \infty$, ויהיו $X_1,..,X_m \overset{i.i.d}{\sim} D$ m עותקים של X. נסמן: $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$ ממוצע המדגם ויהא $0<t\in\mathbb{R}$.

אז: $P[|\bar{X}-E[\bar{X}]| \geq t] = P[|\bar{X}-E(X)| \geq t] \leq \frac{Var[X]}{mt^2}$

דוגמא: (שערוך ה- bias & coin prediction = הסטייה מ- $p=\frac{1}{2}$)

נגדיר את הדוגמא: X - מ"מ ברנולי ($\Omega=\{0,1\}$)

ונסמן: $S = Z_1,..,Z_m \overset{i.i.d}{\sim} X$ דגימה. נסמן את ההסתברות (פונק') לשלוף את S: D_Z^m

נגדיר אלג' למידה A: פונקציה שמקבלת S ומחזירה $\hat{p}$: משערך ל- bias

בנוסף יש $\varepsilon, \delta > 0$.

נרצה ש- $\hat{p}$ יסטה מ- p לכל היותר ε. כמו כן נרצה לשלוט על הסיכוי שנשלוף מדגם S לא מייצג, כלומר נחסום את ההסתברות למאורע שנשלוף מדגם לא מי"צ: $B := \{S \mid |\hat{p}-p| > \varepsilon\}$ ב- δ: $P[B] < \delta$.

נגדיר את זה באופן מתמטי -

$\underbrace{\hat{p}: \bigcup_{m=1}^{\infty}\{0,1\}^m \to [0,1]}$

algo A mistake bound learning: למשימת coin prediction הוא פונק'

$\hat{p}(S)$ כאשר: $S \in \bigcup_{m=1}^{\infty}\{0,1\}^m$ ו- $\hat{p}(S)\in[0,1]$ המקיימת:

לכל $\varepsilon, \delta \le 1$ ק"מ: $0 \le m_A(\varepsilon,\delta) \in \mathbb{R}$ כך שבהינתן מדגם S

של m מספרים (כאשר $m \ge m_A$) שהוגרלו לפי ההתפלגות D_p^m

אז ההסתברות ש- $\hat{p}(S)$ לא יהיה ב: $[p-\varepsilon, p+\varepsilon]$ חסומה ע"

$D_p^m[\underbrace{|\hat{p}-p|>\varepsilon}_{\text{מאורע}}]<\delta$ לכל $0\le p\le 1$. ⊛

לחילופין, אם נקח S כאשר $|S|=m$ ו- $m<m_A$, אז קיים $0\le p\le 1$

כלשהו כך ש: $D_p^m[|\hat{p}-p|>\varepsilon]>\delta$

⋆ הפונק' $m_A(\varepsilon,\delta):[0,1]\times[0,1]\to\mathbb{N}$ נקראת sample complexity של אלגו' הלמידה A.

מה זה אומר? בהינתן $m_A<m$ דגימות, לא חשוב מה ה- p, נוכל לשערך

$\underbrace{\text{אותו}}_{\text{בקירוב } \varepsilon}$ בהסתברות $(1-\delta)$. כמו כן בהינתן $m_A>m$ דגימות, יהיה

קיים לפחות p אחד כך שהשערוך שלנו לא יהיה טוב מספיק (לא יהיה בהסתברות

$1-\delta$ בדיוק של $\pm\varepsilon$ מסביב לערך p האמיתי).

נדגיש שהמצוי של δ לא גלוי מאז שייצרנו את המטבעות. נחשוב על זה כך:

נחשוב על יצירה של דגימות (סדרה של 0,1 באורך m) ונעצור על p מסוים

סדרה היא "טובה" - typical אם $m\hat{p}(S)$ נמצא בין $[m(p-\varepsilon), m(p+\varepsilon)]$, אחרת

היא "רעה" - atypical. אם p ספציפי. אז אפשר להגדיר $m_A^{(p)}$

ספציפי עבור המטבע p. לכן: $m_A(\varepsilon,\delta)=\max_p\{m_A^{(p)}(\varepsilon,\delta)\}$

נשתמש בחסמים: ראינו שהמ"מ: $\hat{p}=\frac{1}{m}\sum_{i=1}^{m} z_m$, מקיים שהתוחלת: $E[\hat{p}]=p$

ולכן $\hat{p}$ הוא unbiased estimator (תכונה טובה). אבל השונות?

1. נחסום ע"י א"ש מרקוב: נגדיר את המ"מ האקראי $X=|\hat{p}-p|\ge 0$. נקבל מא"ש מרקוב:

$$D^m[|\hat{p}-p|\ge\varepsilon]\le\frac{1}{\sqrt{4m\varepsilon^2}}<\delta \Longrightarrow m(\varepsilon,\delta)\le\left\lceil\frac{1}{4\varepsilon^2}\cdot\frac{1}{\delta^2}\right\rceil$$

⊛

2. נחסום ע"י א"ש צ'בישב: מהמסקנה של א"ש צ'בישב- $P[|\bar{X}-E[\bar{X}]|\ge t]\le\frac{Var[X]}{mt^2}$

ולכן: $D^m[|\hat{p}-p|\ge\varepsilon]\le\frac{p(1-p)}{m\varepsilon^2}$ (var של משתנה ברנולי) $\le\frac{1/4}{m\varepsilon^2}=\frac{1}{4m\varepsilon^2}<\delta$

⊛

אז: $m(\varepsilon,\delta)\le\left\lceil\frac{1}{4\varepsilon^2}\cdot\frac{1}{\delta}\right\rceil$ - קיבלנו שזה חסם הדוק יותר מא"ש מרקוב!

חסום ב- $\frac{1}{\delta}$ ולא ב- $\frac{1}{\delta^2}$.

3. נשתמש בעזר תכונה של מ"מ ברנולי כדי למצוא חסם הדוק יותר. התכונה היא שמ"מ ברנולי חסום מ-2 הכיוונים.

(3) חסם הופדינג: $X_1,\dots,X_m$ מ"מ אקראיים בת"ל וחסומים, המקיימים: $a_i \le X_i \le b_i$, ונגדיר: $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_m$. אזי:

$$\mathbb{P}[|\bar{X}-\mathbb{E}[\bar{X}]| \geq \varepsilon] \leq 2\cdot \exp\left(\frac{-2m^2\varepsilon^2}{\sum_{i=1}^{m}(b_i-a_i)^2}\right)$$

מסקנה מחסם הופדינג: נסמן $\mu=\mathbb{E}[X_i]$ ו- $a\le X_i \le b$ עבור ממשיים $X_1,\dots,X_m$ שנגדיר iid מ- X מ"מ. אם $\bar{X}=\frac{1}{m}\sum_{i=1}^{m}X_i$, אזי:

$$\mathbb{P}[|\bar{X}-\mu| \geq \varepsilon] \leq 2\cdot \exp\left(\frac{-2m\varepsilon^2}{(b-a)^2}\right)$$

נפעיל את החסם הזה:

$$D^m[|\hat{p}-p| \geq \varepsilon] \leq 2\cdot \exp(-2m\varepsilon^2)$$

כי $a=0, b=1$. קיבלנו:

$$m(\varepsilon,\delta) \leq \left\lceil \frac{1}{2\varepsilon^2}\cdot \log\left(\frac{2}{\delta}\right)\right\rceil$$

וזה חסם אפילו טוב יותר - לוג של $\frac{1}{\delta}$!