תרגיל בית 1

מבוא לבינה מלאכותית ולמידת מכונה

שמות הסטודנטים המגישים:

מתן קוק פורית 318734944

רותם צלישר 203773601

הערות והוראות הגשה

1. במידה וזה יכול לסייע לכם בתשובותיכם, ניתן (ואף מומלץ) לעשות שימוש חוזר בקוד הנמצא במחברות ה- jupyter שנמסרו לכם בכיתה, או שכתבתם במסגרת מטלות בית קודמות.
2. כפתרון לעבודת בית זאת יש להגיש 3 קבצים נפרדים (לא מאוגדים כקובץ zip):
   א. קובץ עם התשובות לשאלות העיוניות בתרגיל זה (סרוק ברזולוציה גבוהה במידה והתשובות הן בכתב יד).
   ב. מחברת jupyter עם התשובה לשאלה 2. לאחר הרצת המחברת במלואה הגישו אותה בשתי גרסאות:
      i. קובץ ipynb כשהוא מוכן להרצה מחדש (ובנוסף גם קבצי דאטה, במידה ויש כאלה)
      ii. קובץ pdf (על קובץ זה להיות זהה בכל פרטיו לקובץ בסעיף i, פרט להיותו בפורמט שונה)

בהצלחה!

1. מסווג KNN

נתון מידע המכיל 8 דוגמאות מסווגות $\mathcal{D}=\{(\boldsymbol{x_n},y_n)\}_{n=1}^{8}$ כמפורט בטבלה הבאה, כאשר וקטור המאפיינים הוא בעל שני מימדים $\boldsymbol{x}=[x_1,x_2]\ \in\mathbb{R}^2$, התיוג הוא בינארי, וקטור המאפיינים של הדוגמה ה-n-ית מיוצג על ידי $\boldsymbol{x_n}=[x_{n,1},x_{n,2}]$, והסיווג הבינארי של הדוגמה ה-n-ית מיוצג כ- $y_n=1$ או $y_n=0$

n	$x_{n,1}$	$x_{n,2}$	y_n
1	8.1	-1.1	0
2	9.9	2.8	1
3	7.8	2.1	1
4	9.2	0.7	0
5	10.6	0.7	1
6	6.4	2.1	0
7	6.6	1.0	0
8	9.5	-1.1	0

הדוגמאות הנתונות ב-$\mathcal{D}$

א. ציירו את המידע על גרף דו-מימדי עם הצירים x_1 ו- x_2

ב. הוסיפו לגרף הנ"ל את הנקודה הלא מסווגת $\boldsymbol{x_9}=[8.0,1.5]$

ג. חשבו את המרחק האויקלידי של $\boldsymbol{x_9}$ מכל אחת מהנקודות המסווגות, והסבירו את אופן החישוב (די בהסבר עבור אחת מהדוגמאות המתויגות)

ד. סדרו את הנקודות המסווגות על-פי מרחקן האויקלידי מ $\boldsymbol{x_9}$

ה. חשבו את $\hat{y}_9$, סיווג KNN של $\boldsymbol{x_9}$, עבור כל אחד מהערכים $\text{K}=[1,2,3,4,5,6,7,8]$. הסבירו את אופן החישוב והתייחסו לסוגיה הבאה: האם יש מקום לבחון את כל ערכי K הנ"ל, או שישנם ערכים שאינם מתיישבים עם ההיגיון הפנימי של מסווג KNN?

ו. חיזרו על סעיפים ב-ה עבור $[7.0,0]=\boldsymbol{x_{10}}$. התייחסו בקצרה לסוגית מובהקות הסיווג: האם לדעתכם ניתן להסתמך על הסיווגים $\hat{y}_9$ ו-$\hat{y}_{10}$ באופן דומה?

2. מימוש בפייתון של מסווג KNN

בתרגיל זה תממשו מסווג KNN בפייתון ע"פ השלבים שתארנו בכיתה (ובאופן התואם את שלבי החישוב הידני שביצעתם בשאלה 1), תבחנו את ביצועיו על המידע של מאפייני זני האירוס אותו ראינו בהרצאה, ותבחרו ערך מתאים של K עבור מידע זה.

<u>הערה</u>: מטרותיו העיקריות של התרגיל הינן התנסות עם בעיית סיווג באופן כללי ומסווג KNN בפרט, ותרגול ראשוני של שימוש בפייתון ללמידת מכונה. לפיכך שימו דגש על קריאות הקוד שלכם ובהירות תפקידו של כל אחד מהחלקים בו, לפחות במידה דומה לשיקולי יעילות כגון זמן ריצה.

א. פיתחו את הקובץ המצורף HW1Q2.ipynb והוסיפו את הקוד הדרוש ע"פ ההנחיות בו. לאחר השלמת המחברת והרצתה מודפס בסופה גרף המסכם את דיוק המסווג על סדרת המבחן עבור ערכי K שונים. הסבירו את התוצאות המוצגות בגרף.

ב. בסעיף זה הנכם מתבקשים לחזור על התהליך שביצעתם בסעיף א 100 פעם, כשבכל פעם תוגרל חלוקה שונה של המידע לסדרות אימון ומבחן, ובסופו של דבר עליכם ליצור גרף **<u>יחיד</u>** הדומה במהותו לזה שבסעיף א, אך **<u>עבור כל K</u>** לדווח בו את **<u>ממוצע הדיוק</u>** (accuracy) שהתקבל על פני כל 100 ההרצות.
הסבירו את הגרף שקיבלתם, והשתמשו בו על מנת לקבוע ערך K מתאים למידע הנתון, שבו תמליצו להשתמש לסיווג של פרחים שאינם ידועים לכם בשלב האימון. נמקו בחירתכם.

שימו לב: בסעיף א וקטור הדיוקים שחישבתם הינו באורך Kmax (כפי שערכו נקבע בקוד המחברת). בסעיף ב הינכם נדרשים ראשית לחשב מטריצת דיוקים במימד 100 x Kmax, ואז לחשב את הממוצע עבור כל ערך של K ולקבל וקטור של ממוצעי דיוקים (שוב באורך Kmax).

3. מטבע מוטה

בשאלה זאת תפתרו את המקרה כללי של אחד התרגילים שפתרנו בכיתה. בנוסף לפתרון לשאלה העיונית, הנכם מתבקשים לצרף למחברת ה- Jupyter שתגישו במסגרת המטלה גם את הגרפים המתוארים למטה יחד עם הקוד המייצר אותם והסבר קצר של התוצאות המתוארות בהם.

נתון מטבע בעל שני צדדים המסומנים H ו-T, שנסמן כ-p את ההסתברות לקבלת H בהטלה בודדת שלו. המטבע הוטל N פעמים, מתוכם התקבל K פעמים הצד המסומן H, וביתר הפעמים התקבל הצד המסומן T.

א. רישמו את ההסתברות לקבלת הסדרה הנ"ל בתלות ב-p

ב. כתבו קוד המצייר את גרף הפונקציה שקיבלתם בסעיף א לכל הערכים בתחום $0 \leq p \leq 1$ עבור כל אחד מהמקרים הבאים:
- N=10, K=3
- N=50, K=15
- N=300, K=90

ג. חשבו את הנוסחה שערוך הסבירות המרבית של p בתלות בערכי N ו-K כלליים (באופן דומה לחישוב שביצענו בכיתה תוך שימוש בפונקציית ה log-likelihood). השתמשו בנוסחה בכדי לחשב את הערך המתקבל עבור על אחד מצמד הערכים בסעיף ב, והוסיפו את הערך שהתקבל לגרף הרלוונטי.

דונו בקצרה בתוצאות, והסבירו כיצד N ו-K משפיעים על שערוך ערכו של p בהסתמך על הדאטה.

4. בעיית מונטי הול

בשאלה זאת תממשו את המשחק מונטי הול שדנו בו בכיתה (תיאור שלו מצורף למטה, ובנוסף ניתן לקרוא עליו כאן), תסמלצו אותו מספר רב של פעמים עבור טקטיקות משחק שונות, ותפרשו את התוצאות הסטטיסטיות של הסימולציה. כפתרון עליכם להגיש מחברת jupyter הכוללת את הקוד שכתבתם ואת הגרפים המתוארים למטה, והסבר קצר שלכם לתוצאות.

תיאור המשחק:
כפי שדנו בשיעור ,במשחק הקרוי 'מונטי הול" מציגים בפני שחקן שלוש דלתות, שמאחורי אחת מהן מצויה מכונית ומאחורי כל אחת משתי הדלתות האחרות ישנה עז. מטרת השחקן היא לבחור את הדלת שמאחוריה נמצאת המכונית - אם יצליח יזכה בה, ואם לא אז לא יקבל דבר. ע"פ כללי המשחק ראשית בוחר השחקן דלת באקראי. לאחר מכן המנחה, היודע מה יש מאחורי כל דלת, פותח אחת משתי הדלתות האחרות ומגלה מאחוריה עז. כעת ניתנת לשחקן האפשרות לדבוק בבחירתו המקורית, או להחליפה בדלת האחרת שעודנה סגורה.
נבחן את תוצאות המשחק עבור שלושה סוגי שחקנים:

- שחקן א תמיד דבק בבחירתו המקורית.
- שחקן ב תמיד מחליף את בחירתו לדלת שנותרה סגורה.
- שחקן ג דבק בבחירתו בסיכוי 0.5, ומחליפה בסיכוי 0.5.

א. כיתבו קוד המדמה את משחק מונטי הול ע"פ השלבים הבאים (מומלץ לממש כל שלב בפונקציה נפרדת):
- הגרילו מאחורי איזו דלת תימצא המכונית
- בחרו באופן אקראי את הדלת אותה בוחר השחקן
- קיבעו מהי הדלת אותה פותח המנחה (בהתאם לכללי המשחק)
- בהתאם לסוג השחקן (מבין השלושה שתוארו למעלה) קיבעו האם הוא בוחר להישאר עם בחירתו המקורית או להחליפה
- בידקו האם בחירתו הסופית של השחקן הובילה לזכייה במכונית, או לא

ב. שחקן א משחק 1000 פעמים ברציפות. הריצו את הקוד שכתבתם 1000 פעם ושימרו את התוצאות. ציירו גרף המתאר, לאחר כל משחק, את מספר הפעמים שהשחקן זכה במכונית. בציר האופקי ציינו מספר המשחקים ששוחקו ובציר האנכי את שיעור הזכיות במכונית עד לאותו המשחק (כלומר הערכים בציר האופקי יהיו בין 1 ל 1000, ובציר האנכי בין 0 ל-1).

ג. חיזרו על הנ"ל בנפרד עבור שחקן ב ושחקן ג. מי השחקן שהטקטיקה שלו הובילה לסיכוי הגבוה ביותר לזכות במכונית? הסבירו בקצרה את התוצאות.

ד. ענו בקצרה על השאלה הבאה: בהינתן התוצאות של שחקנים א ו-ב, האם ניתן היה להעריך את סיכויי ההצלחה של שחקן ג ללא הרצת הסימולציה עבורו?

5. התפלגות משותפת, שולית ומותנית של משתנים אקראיים בדידים

נתונה $P(X,Y)$ ההתפלגות המשותפת הבאה של שני משתנים אקראיים בדידים

P(X,Y)

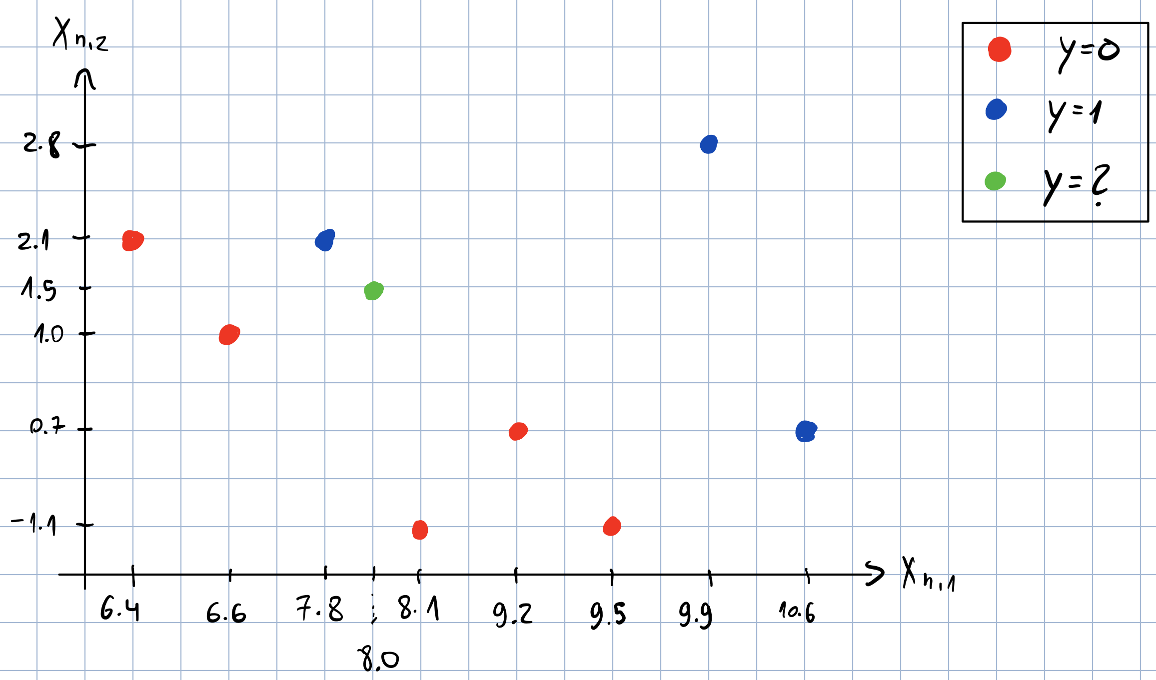
	0	0.01	0.02	0.03	0.1	0.1
Y	**1**	0.05	0.1	0.05	0.07	0.2
	2	0.1	0.05	0.03	0.05	0.04
		0	**1**	**2**	**3**	**4**

X

חשבו את

א. $P(X|Y=1)$
ב. $P(Y|X=3)$
ג. $P(Y|X>2)$
ד. $P(X)$
ה. $P(Y)$
ו. האם המשתנים האקראיים X ו Y הם בלתי תלויים סטטיסטית?

1 א + ב



ג

$$d_{X_g,X_n} = \sqrt{(X_{g,1}-X_{n,1})^2+(X_{g,2}-X_{n,2})^2}$$

נוסחה כללית עבור מרחק אוקלידי
בין 2 נקודות

$d_{X_g,X_1} = 2.602$

$d_{X_g,X_2} = 2.302$

$d_{X_g,X_3} = 0.632$

$d_{X_g,X_4} = 1.44$

$d_{X_g,X_5} = 2.72$

$d_{X_g,X_6} = 1.709$

$d_{X_g,X_7} = 1.487$

$d_{X_g,X_8} = 3$

ד

X_3 מרחק קטן

X_4

X_7

X_6

X_2

X_1

X_5

X_8 מרחק גדול

ה $\hat{y}_9 = ?$, $k = [1,2,3,4,5,6,7,8]$

$k=1$: $\hat{y}_9 = y_3 = 1$

$k=2$: $\hat{y}_9 = (y_4 \text{ או } y_3) = 1$

$k=3$: $\hat{y}_9 = \underset{y_3}{1}, \underset{y_4}{0}, \underset{y_7}{0} = 0$

$k=4$: $\hat{y}_9 = 1,0,0,\underset{y_6}{0} = 0$

$k=5$: $\hat{y}_9 = 1,0,0,0,\underset{y_2}{1} = 0$

$k=6$: $\hat{y}_9 = 1,0,0,0,1,\underset{y_1}{0} = 0$

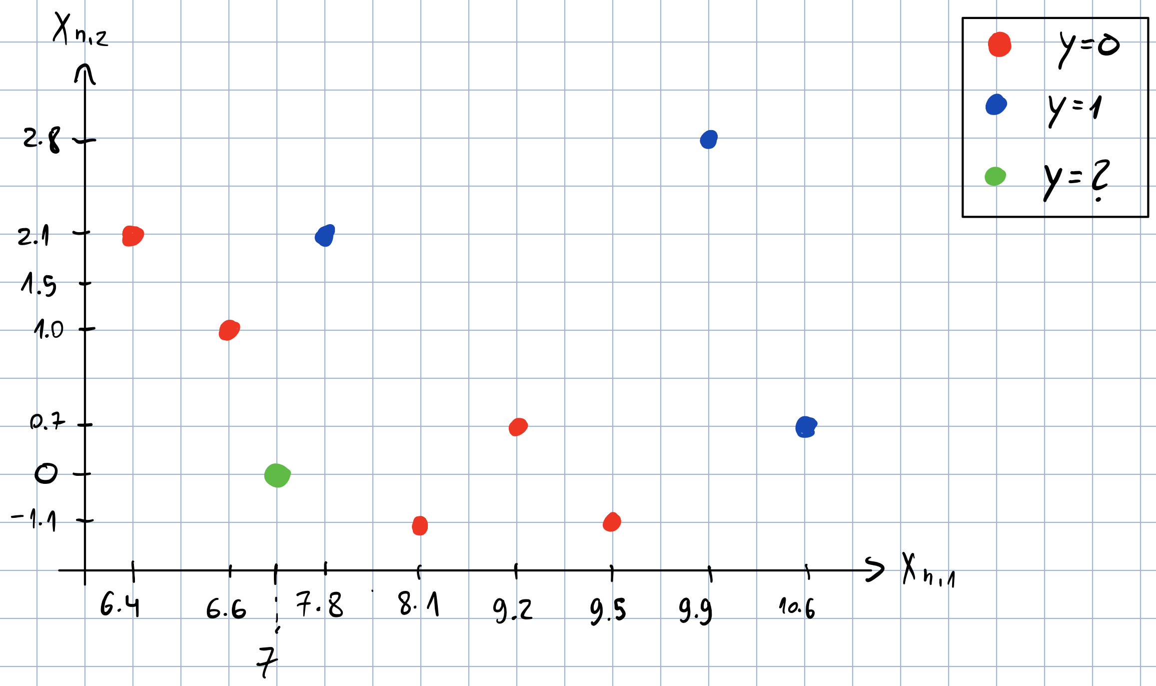
$k=7$: $\hat{y}_9 = 1,0,0,0,1,0,\underset{y_5}{1} = 0$

$k=8$: $\hat{y}_9 = 1,0,0,0,1,0,1,\underset{y_8}{0} = 0$

ניתן היה לא לבדוק את $k=8$ כי הוא יגרום לנקודה החדשה שתכנס לקבל סוג עפ״י הצבע הדומיננטי בכל המפה.

ניתן לא לבדוק את $k=1$, מפני שזה יגרום לכך שבצבעה חדש שיתויג על ידינו יקבל יותר מדי משקל בהשפעתו על הצבעה חדש שיגיע אחריו.

1 → ב



ג

$$d_{X_{10},X_n} = \sqrt{(X_{9,1}-X_{n,1})^2+(X_{9,2}-X_{n,2})^2}$$

נוסחה כללית עבור מרחק אוקלידי
בין 2 נקודות

$d_{X_{10},X_1} = 1.56$	$d_{X_{10},X_5} = 3.67$
$d_{X_{10},X_2} = 4.03$	$d_{X_{10},X_6} = 2.18$
$d_{X_{10},X_3} = 2.247$	$d_{X_{10},X_7} = 1.08$
$d_{X_{10},X_4} = 2.308$	$d_{X_{10},X_8} = 2.731$

ד

X_7 מרחק קטן
X_1
X_6
X_3
X_4
X_8
X_5
X_3 מרחק גדול

ה

$\hat{y}_9 = ?$, $k = [1, 2, 3, 4, 5, 6, 7, 8]$

k=1: $\hat{y}_9 = y_7 = 0$

k=2: $\hat{y}_9 = ($ y_7 או y_1 $) = 0$

k=3: $\hat{y}_9 = \underset{y_7}{0}, \underset{y_1}{0}, \underset{y_6}{0} = 0$

k=4: $\hat{y}_9 = 0, 0, 0, \underset{y_3}{1} = 0$

k=5: $\hat{y}_9 = 0, 0, 0, 1, \underset{y_4}{0} = 0$

k=6: $\hat{y}_9 = 0, 0, 0, 1, 0, \underset{y_8}{0} = 0$

k=7: $\hat{y}_9 = 0, 0, 0, 1, 0, 0, \underset{y_5}{1} = 0$

k=8: $\hat{y}_9 = 0, 0, 0, 1, 0, 0, 1, \underset{y_3}{1} = 0$

לא ניתן להסתמך על הסיווגים באופן דומה. ניתן לראות בבירור ש $\hat{y}_{10}$ נמצא במפה במקום שיותר קל לסיווג לעומת $\hat{y}_9$.

($\hat{y}_{10}$ נמצא "עמוק" יותר באדומים")

3

א

$$P(X=H)=p \quad , \quad P(X=T)=1-p$$

N הטלות:

k פעמים ← H

N-k פעמים ← T

הטלות בלתי תלויות

$$P(\underbrace{HHH\ldots H}_{k\text{ פעמים}}\ \underbrace{TTT\ldots T}_{N-k\text{ פעמים}}) = \underbrace{P(H)\cdot P(H)\cdots}_{k\text{ פעמים}}\ \underbrace{P(T)\cdot P(T)\cdots P(T)}_{N-k\text{ פעמים}} =$$

$$= P(H)^k \cdot P(T)^{N-k} = p^k \cdot (1-p)^{N-k} = p^k - p^N$$

ב

$$p_{ML} = \arg\max\left(p^k\cdot(1-p)^{N-k}\right) = \arg\max\left[\ln\left(p^k\cdot(1-p)^{N-k}\right)\right] =$$

$$= \frac{d}{dp}\left[\ln\left(p^k\cdot(1-p)^{N-k}\right)\right] = \frac{d}{dp}\left[\ln(p^k) + \ln\left((1-p)^{N-k}\right)\right] =$$

$$= \frac{d}{dp}\left[k\cdot\ln(p) + (N-k)\cdot\ln(1-p)\right] \Rightarrow \frac{k}{p} + \frac{(N-k)\cdot(-1)}{1-p} = 0$$

$$k - \cancel{kp} + \cancel{kp} - Np = 0$$

$$p = \frac{k}{N}$$

זה ה p שיביא אותנו להסתברות המקסימלית עבור סידרת ההטלות הנתונה.

ג

ככל שיש לנו יותר דאטה כך גדלה הוודאות שלנו לגבי ערכי p אפשריים.

4

ג ניתן לראות שב 1000 משחקים רנדומים שחקן מספר
2 זכה הכי הרבה פעמים במכונית, אחריו שחקן מספר 3
ובמקום השלישי והאחרון יש את שחקן מספר 1. זה מסתדר
עם התיאוריה שהוסברה בכיתה, הטקטיקה שאומרת שלאחר
שהמנחה פותח דלת עם עז, יש להחליף דלת היא הטקטיקה
המוצלחת ביותר. מפני שהמנחה מגלה לנו מידע נוסף
על המשחק שעוזר לשחקן לבחור במכונית. אם בבחירה
הראשונה השחקן בחר בעז, אז המנחה שולל את הדלת השניה
עם העז ולשחקן יש ניצחון "וודאי" אם יחליף דלת.
השחקן שפעל בטקטיקה זו הכי הרבה פעמים הוא זה שזכה
בהכי הרבה משחקים מתוך האלף הרנדומים.

ד ענינו במחברת ה Jupyter

5 c

		X					
		0	1	2	3	4	P(y)
y	0	0.01	0.02	0.03	0.1	0.1	0.26
	1	0.05	0.1	0.05	0.07	0.2	0.47
	2	0.1	0.05	0.03	0.05	0.04	0.27
	P(x)	0.16	0.17	0.11	0.22	0.34	

$$P(x \mid y=1) = \frac{P(x, y=1)}{P(y=1)} = \begin{cases} x=0 \rightarrow P(x \mid y=1) = 5/47 \\ x=1 \rightarrow P(x \mid y=1) = 10/47 \\ x=2 \rightarrow P(x \mid y=1) = 5/47 \\ x=3 \rightarrow P(x \mid y=1) = 7/47 \\ x=4 \rightarrow P(x \mid y=1) = 20/47 \end{cases}$$

ג

	0	1	2	3	4	P(y)
0	0.01	0.02	0.03	0.1	0.1	0.26
1	0.05	0.1	0.05	0.07	0.2	0.47
2	0.1	0.05	0.03	0.05	0.04	0.27
P(x)	0.16	0.17	0.11	0.22	0.34	

$$P(y \mid x=3) = \frac{P(x=3,y)}{P(x=3)} = \begin{cases} y=0 \rightarrow P(y \mid x=3) = 10/22 \\ y=1 \rightarrow P(y \mid x=3) = 7/22 \\ y=2 \rightarrow P(y \mid x=3) = 5/22 \end{cases}$$

ד

	0	1	2	3	4	P(y)
0	0.01	0.02	0.03	0.1	0.1	0.26
1	0.05	0.1	0.05	0.07	0.2	0.47
2	0.1	0.05	0.03	0.05	0.04	0.27
P(x)	0.16	0.17	0.11	0.22	0.34	

$$P(y \mid x>2) = \frac{p(x>2,y)}{p(x>2)} = \begin{cases} y=0 \rightarrow p(y \mid x>2) = \frac{0.1+0.1}{0.22+0.34} = \frac{20}{56} \\ y=0 \rightarrow p(y \mid x>2) = \frac{0.07+0.2}{0.56} = \frac{27}{56} \\ y=0 \rightarrow p(y \mid x>2) = = \frac{0.05+0.04}{0.56} = \frac{9}{56} \end{cases}$$

ק

	0	1	2	3	4	P(y)
0	0.01	0.02	0.03	0.1	0.1	0.26
1	0.05	0.1	0.05	0.07	0.2	0.47
2	0.1	0.05	0.03	0.05	0.04	0.27
P(x)	0.16	0.17	0.11	0.22	0.34	

$$P(x) = P(X, Y=y) =$$

ה

	0	1	2	3	4	P(y)
0	0.01	0.02	0.03	0.1	0.1	0.26
1	0.05	0.1	0.05	0.07	0.2	0.47
2	0.1	0.05	0.03	0.05	0.04	0.27
P(x)	0.16	0.17	0.11	0.22	0.34	

$$P(y) = P(X=x, Y) =$$

ו

אם X,Y בת"ס אז צריך להתקיים $P(x)\cdot P(y) = P(x,y)$

עבור כל ערכי X, Y.

אבל ניתן לראות שהתנאי לא מתקיים עבור $X=0, Y=0$

$$P_x(0)\cdot P_y(0) = 0.0416 \neq P(X=0, Y=0) = 0.01$$

לכן X,Y אינם בת"ס