

## Introduction to Data Science - Final Project

Roi Yozevitch & Idan Kaminetsky

### הקדמה:

בעקבות ניסיון העבר ומגבלות הקורונה, הציון המסכם בקורס יתבסס ברובו על עבודה הגמר הנוכחית. מדובר בעבודה ממשמעותית הדורשת לא מעט מכם הסטודנטים.

את העבודה יש להגיש בתור מחברות Jupyter (סיממת ipynb) וקובצי PDF באתר GitHub ולצרף תיעוד משמעותי. נקודות יינטנו (או יוחסרו) על תיעוד מעולה (גראוע).

### חשיבות מzd

- נא לקרוא את פרקים 4-1 בספר הקורס [Hands-On Machine Learning with Scikit-Learn](#)
- נא לעבור על הקורס של [git](#) באתר Udacity . זו מטלה חובה. חלק מהציון הסופי תלוי במספרי commits אשר תעשו.
- רשות - עברו על [כל הקוווט](#) "מבוא למדעי הנתונים" באתר קמפוס ולא רק על הפרקים העוסקים ברכישת נתוניים).

### הנחיות כלליות

1. לעובדה שני חלקים מרכזיים. החלק הראשון כולל שאלות תיאורתיות ושאלות תכניות פשוטות. במידה ועשיתם את שיעורי הבית, לא אמורה להיות לכם בעיה עם שאלות אלו. ההגשה של שאלות אלו היא פרטנית דרך GitHub. ז"א שכל סטודנט צריך להגיש בעצמו את הפתרונות של השאלה.
2. החלק השני (המרכזי) של העבודה מורכב מניתוח נתונים של data אותו אתם אמורים להשיג בעצמכם. את החלק הזה מגישים בזוגות. שימו לב לקובץ הרישום.

"בכל מערכת יחסים יש את הצד שאוהב והצד שמרשה שיאהבו אותו" [טולסטוי]

3. חשוב להסביר שני סטודנטים יכולים לקבל ציון שונה על עבודה שהם הגישו ביחד. הסיבה היא שהעבודה כוללת מבחן סופי בע"פ (דרך zoom) אשר בו שני הסטודנטים יctraco להגן על העבודה.

ניסיון העבר מלמד שכאשר סטודנט אחד עושה את כל העבודה, ניתן לגנות את זה بصورة יחסית פשוטה בבחינה בעל-פה.

### רכישת ה-data:

הדרך הטובה ביותר להשיג את ה-data היא על ידי תהליך של [web scrapping](#). הקורס "[מבוא למדעי הנתונים](#)" באתר קמפוס מסביר במשך שני פרקים איך לעשות זאת. בנוסף, יש מדריכי למידה מעולים בראשת. בכל מקרה, מי שלא מצליח, יכול להוריד קובץ נתונים אחד האתרים המרכזיים בראשת (לא לאתר Kaggle!) ולעבוד עליו. אתר מצוין בעברית הינו [מארג הנתונים הממשלתי](#) ויש לו מקבילות גם באנגלית. ה-data יכול לעשות גם בשאלת סיוג (קלסיפיקציה) וגם בשאלת חיזוי.

**משמעותו לב, מי שיוריד נתונים שלא דוד web scrapping, הציוו המקסימלי שהוא יוכל לקבל על העבודה הינו 90.**

## חלק א- שאלות תיאורטיות ותשובות פשוטות

הסתברות, חוק ביסיס:

1. א. בערך 1/125 מהלידות זה תאומים לא זחים ו-1/300 מהלידות זה תאומים זהים. לאלביס היה אח תאום שמת בילדת. מה ההסתברות שאלביס היה תאום זהה? (ניתן להניח שההסתברות להולדת בן ובת שווה ל-1/2).  
ב. יש שתי קערות של עוגיות. בקערה 1 יש 10 עוגיות שקדים ו-30 עוגיות שוקולד. בקערה 2 יש 20 עוגיות שקדים ו-20 עוגיות שוקולד. אריק בחר קערה **באקראי** ובחר ממנה עוגיה **באקראי**. העוגיה שנבחרה היא שוקולד. מה ההסתברות שאrik בחר את קערה 1?  
ג. בשנת 1995 חברת M&M הוסיפה את הצבע כחול. לפני השנה זו, התפלגות הצבעים בשקיית M&M הייתה נראית כך:

30% Brown, 20% Yellow, 20% Red, 10% Green, 10% Orange, 10% Tan

החל משנה 1995, ההתפלגות נראית כך:

24% Blue , 20% Green, 16% Orange, 14% Yellow, 13% Red, 13% Brown.

לחבר שלכם יש 2 שקיות M&M, אחת משנה 1994 ואחת משנה 1996 והוא לא מוכן לגנות לכם איזו שקייה שייכת לאיזו שנה. אבל הוא נותן לכם סוכריה אחת מכל שקייה. סוכריה אחת היא צהובה ואחת היא יוקה. מה הסיכוי שהסוכריה הצהובה הגיעה מהשקייה של 1994?

3. (20 נקודות) הלכת לדוקטור בעקבות ציפורי חודרנית. הדוקטור בחר באקרואי לבצע בדיקתدم הבודקת שפעת חזירים. ידוע סטטיסטי ששיעור שפעת זו פוגעת ב-1 מתוך 10,000 אנשים באוכלוסייה. הבדיקה מדויקת ב-99 אחוז במובן שהסתברות ל *false positive* היא 1%. הוא אומר שהבדיקה סיוגה בטעות אדם בריא כאדם חולה היא 1 אחוז. ההסתברות ל- *false negative* היא 0 – אין סיכוי שהבדיקה תגידי על אדם החולה בשפעת חזירים שהוא בריא. בבדיקה יצא חיובי (יש לך שפעת).

- מה ההסתברות שיש לך שפעת חזירים?
- נניח שהזורת מתאילנד לאחרונה ואתה יודע ש-1 מתוך 200 אנשים שזרו לאחרונה מתאילנד, שזרו עם שפעת חזירים. בהינתן אותה סיטואציה כמו בשאלת א, מה ההסתברות (המתוקנת) שיש לך שפעת חזירים?

4. בערך  $\frac{1}{300}$  מהלידות היא של תאומים זהים ו  $\frac{1}{125}$  מהלידות היא של תאומים לא זהים. לנסיך צ'ארלס היה אח תאום שמת בילדת. מה ההסתברות שהיא לו אח תאום זהה? (תאומים זהים חייבים להיות בני אותו המין)

#### קריאה/צפיה מומלצת:

[מבוא מקסים לחוק ביס](#) מאת אליעזר יודובסקי (מחבר הספר הארי פוטר והרצינליות)

### Random Variables:

- Roi is playing a dice game with Yael.

Roi will roll 2 six-sided dice, and if the sum of the dice is divisible by 3, he will win 6\$. If the sum is not divisible by 3, he will lose 3\$.

**What is Roi's expected value of playing this game?**

- Sharon has challenged Alex to a round of Marker Mixup. Marker Mixup is a game where there is a bag of 5 red markers numbered 1 through 5, and another bag with 5 green markers numbered 6 through 10.

Alex will grab 1 marker from each bag, and if the 2 markers add up to more than 12, he will win 5\$, 5. If the sum is exactly 12, he will break even, and If the sum is less than 12, he will lose 6\$.

**What is Alex's expected value of playing Marker Mixup?**

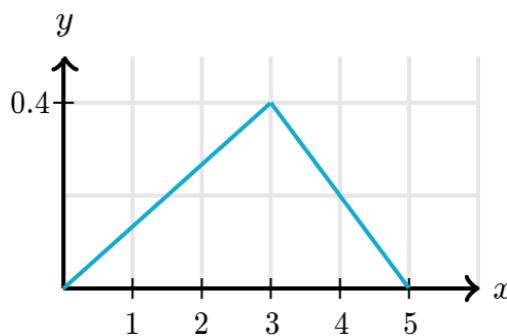
- A division of a company has 200 employees, 40%, percent of which are male. Each month, the company randomly selects 8 of these employees to have lunch with the CEO.

**What are the mean and standard deviation of the number of males selected each month?**

4. Different dealers may sell the same car for different prices. The sale prices for a particular car are normally distributed with a mean and standard deviation of 26,000\$ and 2,000\$, respectively. Suppose we select one of these cars at random. Let  $X$  = the sale price (in thousands of dollars) for the selected car.

Find  $P(26 < X < 30)$ ,

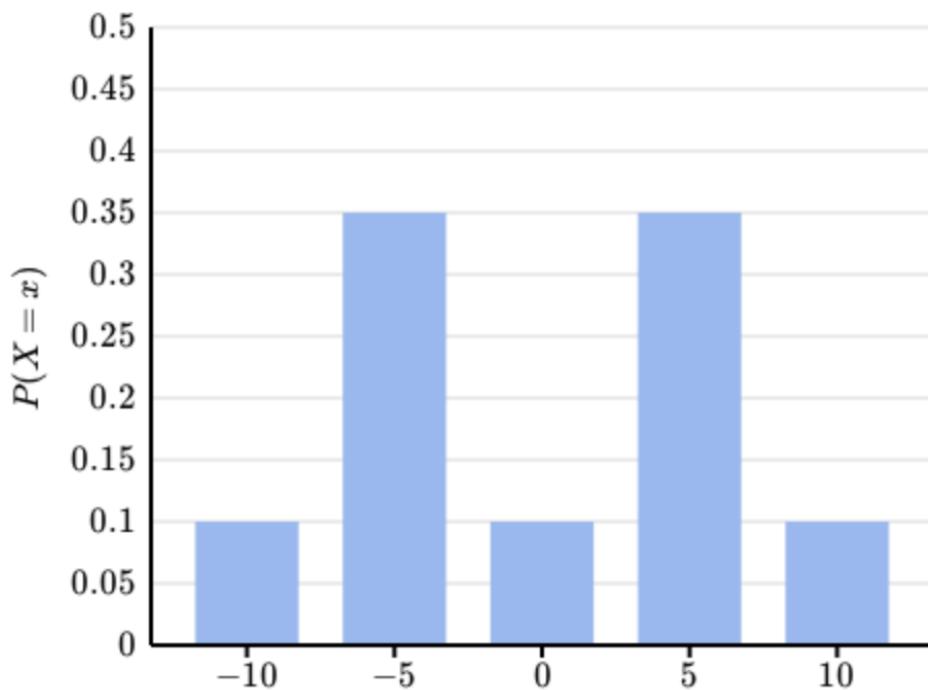
5. Given the following distribution, what is  $P(x > 3)$ ?



6. A company has 500 employees, and 60% of them have children. Suppose that we randomly select 4 of these employees.

What is the probability that exactly 3 of the 4 employees selected have children?

7. Look at the next Graph. What is the expected value of  $X$ ?



### קריאה מומלצת:

מאמר קצר באתר Medium

מאמר נרחב יותר בנושא.

### תרגילי תכניות pandas

1. כתוב תוכנית בשפת פיתון אשר מקבלת מספר בסיסי 10 ומדפיסה את המספר בכל הבסיסים האחרים (בסיס 2, בסיס 8 ובסיס 16).
2. לבעיה זו מצב DATASET של סרטים.

```
▶ import pandas as pd  
cast = pd.read_csv('data/cast.csv')  
cast.head()
```

2]:

	title	year	name	type	character	n
0	Suuri illusioni	1985	Homo \$	actor	Guests	22.0
1	Gangsta Rap: The Glockumentary	2007	Too \$hort	actor	Himself	NaN
2	Menace II Society	1993	Too \$hort	actor	Lew-Loc	27.0
3	Porndogs: The Adventures of Sadie	2009	Too \$hort	actor	Bosco	3.0
4	Stop Pepper Palmer	2014	Too \$hort	actor	Himself	NaN

טענו את ה-DATASET למחברת וענו על השאלות הבאות :

1. How many movies have the title "Hamlet"?
2. List all of the "Treasure Island" movies from earliest to most recent.
3. How many roles were credited in the silent 1921 version of Hamlet?
4. Use groupby() to plot the number of "Hamlet" films made each decade
5. How many leading (n=1) roles were available to actors, and how many to actresses, in each year of the 1950s?
6. List the 10 actors/actresses that have the most leading roles (n=1) since the 1990's.
7. List, in order by year, each of the films in which Frank Oz has played more than 1 role

### כפייה מומלצת

סרטונו מקסים של טרייקים ב-pandas

## חלק ב – מדעי הנתונים

במקרה זוגית זו (כל זוג סטודנטים מגיש), יש צורך למצוא נתונים בראשת, לרכוש אותם (מומלץ דרך ספריית BeautifulSoup) ואז לנסות ליצור מודל של ה-data, (איזו עמודה אנחנו רוצים לנחש). ניתן לחקח כל מידע טבלאי (לא תמונות, לא וידאו, לא אודיו) וניתן לבחור בעיית סיווג או חיזוי. **בנוסף**, במידה ובחרתם בעיית חיזוי, חפשו ב-Kaggle [בעיית סיווג](#), הורידו את ה-data ופתרו אותה. במידה ובחרתם בעיית חיזוי, חפשו ב-Kaggle [בעיית חיזוי](#), הורידו את ה-data ופתרו אותה. המטרה היא להכיר לכם את שני סוגי הבעיה והשיטות להתמודד איתן.

### **נקודות חשובות:**

1. האם ה-data שלכם הינו מסווג סיווג או חיזוי?
2. מה פונקציית המטרה שלכם (איזו עמודה אתם מנסים לנחש)?
3. איך אתם יכולים להבטיח שאין לכם **data leakage**?
4. מה אומר מודל 0? מה אומר מודל קצר יותר מתחכם?
5. האם יש לכם רעיון אפרורי לחבר בין המשתנים? מה אמורה (לפי דעתכם) להיות פונקציית המטרה?
6. איך ויזואלייזציה יכולה לעזור לכם להבין את ה-data? במידה והחלטתם לצרף תרשימים לעובדה שלכם, מה אתם לומדים ממנו? (חשוב מאד!)
7. לפי מה בחרתם את המאפיינים של המודל? האם הורדתם מאפיינים לא רלוונטיים? לפי מה בחרתם?
8. מה המודל המתאים ביותר לבעה שלכם? אבקש להשתמש גם במודלים שלמדנו (knn, linear regression) וגם במודלים אחרים לא למדנו (פחות מודל אחד). שימו לב שכאשר אתם עובדים עם מודל מסוים, בדקו את התיעוד הרלוונטי עליו בספרייה sklearn. נסו להבין את הפרמטרים השונים.
9. מה מודל השגיאה שלכם? שימו לב שעבור בעיות מסוימות שונים, קיימים מודלי שגיאה שונים (למשל, דיקוק בעיות סיווג מול MSE בעיות חיזוי). האם דיקוק מספיק? במידה ולא, מה למדתם מעוקמת AUC?
10. חשוב מאד לחלק את ה-data לסט אימון (train) וסט בבחינה (test) ולבדוק על סט הבחינה רק כאשר סיימתם להתאים את הפרמטרים שלכם. עשו זאת גם אם אתם משתמשים ב-cross-validation
11. האם יש לכם מאפיינים קטגוריאליים ב-data? איך אתם מטפלים בהם?

### Additional Sources

כל חומר העזר (וסיכוםי סטודנטים) יופיע בקישור [זהה](#).

הנתקן

.1 תשליך  
לכט

נניח ש 125 - 1 מינימום קבוצה של 300 נסיבות כ' 16'  
 גודלה של קבוצה מוגדרת כ' 16'.  
 סט  $\frac{1}{2}$  . 75% מילוי הינו מילוי (50%).  
 אם נוציאו 16' מילוי מילוי (50%) נישר 16'  
 סט  $\frac{1}{2}$ .

$$P\left(\frac{\text{טלאי}}{\text{טלאי}} \mid \text{טלאי טלאי}\right) = \frac{P\left(\frac{\text{טלאי}}{\text{טלאי}} \mid \text{טלאי}\right)}{P\left(\frac{\text{טלאי}}{\text{טלאי}}\right)}$$

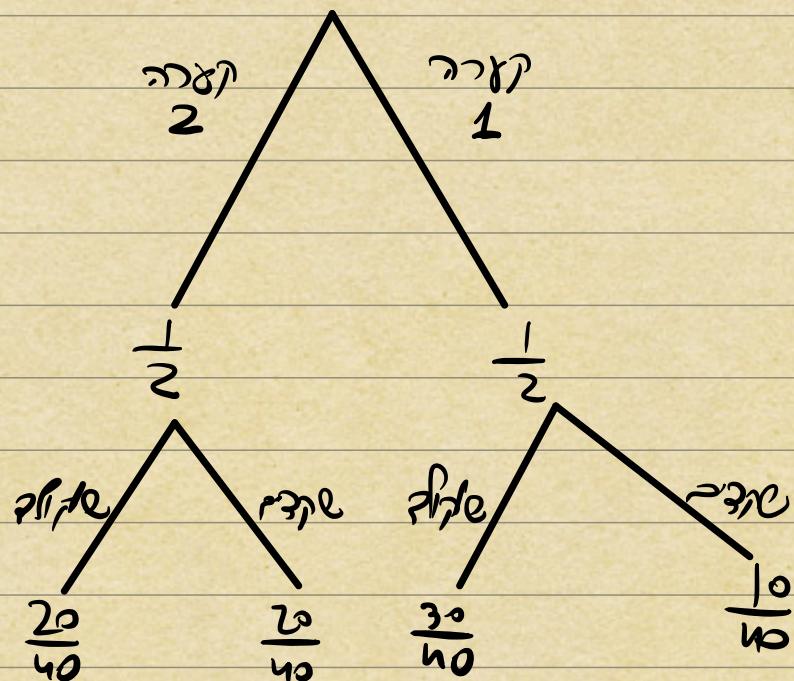
$$= \frac{\frac{1}{300}}{\frac{1}{300} + \frac{1}{2} \cdot \frac{1}{125}} = \frac{\frac{1}{300}}{\frac{1}{300} + \frac{1}{250}} = \frac{5}{11}$$

: 220

לנתקן ב' מילוי מילוי מילוי

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

5



ההסתה ששני הרים יישמרו  
ההסתה ששני הרים יישמרו ← 1  
ההסתה ששני הרים יישמרו ← 2 ← 2

$$P(\text{ההסתה ששני הרים יישמרו}) = \frac{1}{2} \cdot \frac{20}{40} + \frac{1}{2} \cdot \frac{30}{50} = \frac{5}{8}$$

1. חישוב הסתברות ששני הרים יישמרו

ההסתה ששני הרים יישמרו ← 1

$$P(\text{ההסתה ששני הרים יישמרו} \mid \text{ההסתה ששני הרים יישמו}) = \frac{P(\text{ההסתה ששני הרים יישמו} \mid \text{ההסתה ששני הרים יישמו})}{P(\text{ההסתה ששני הרים יישמו})}$$

$$= \frac{\frac{1}{2} \cdot \frac{30}{50}}{\frac{5}{8}} = \frac{\frac{3}{8}}{\frac{5}{8}} = \frac{3}{5}$$

דף

2 של

: 1995 ינואר

1/2	2/3	3/4	4/5	5/6	6/7
10%	10%	10%	20%	20%	30%
$\frac{10}{100}$	$\frac{10}{100}$	$\frac{10}{100}$	$\frac{20}{100}$	$\frac{20}{100}$	$\frac{30}{100}$

: lag 1995

1/2	2/3	3/4	4/5	5/6	6/7
20%	16%	15%	13%	14%	13%
$\frac{20}{100}$	$\frac{16}{100}$	$\frac{15}{100}$	$\frac{13}{100}$	$\frac{14}{100}$	$\frac{13}{100}$

לפחות אחד מ-6 מילים נערך ב-5 מילים  
כלשה. עלינו לרשום 11 מילים  
לפחות. לאן ניתן לרשום 11 מילים  
לפחות? לאן ניתן לרשום 11 מילים  
לפחות?

$$P(\text{פחות } 5 \text{ מילים}) = \frac{1}{2} \cdot \frac{20}{100} + \frac{1}{2} \cdot \frac{14}{100} = \frac{12}{100}$$

: מושג לאנרגיה הניתנת על ידי נזק האנרגיה

$$P(\text{לאנרגיה} - N) / \text{טבון} = \frac{P(\text{לאנרגיה} - N)}{P(\text{טבון})}$$

$$= \frac{\frac{1}{2} \cdot \frac{20}{100}}{\frac{12}{100}} = \frac{b}{12}$$

3 rule

א ב כ ד ג ה י ז ק מ נ ס ע ו ח י צ ו ש ת

א ב כ ד ג ה י ז ק מ נ ס ע ו ח י צ ו ש ת

א ב כ ד ג ה י ז ק מ נ ס ע ו ח י צ ו ש ת

א ב כ ד ג ה י ז ק מ נ ס ע ו ח י צ ו ש ת

	$\bar{B}$	B	
$\frac{1}{10,000}$	0 (False Negative)	$\frac{1}{5000}$ (True Positive)	A
$\frac{9999}{10,000}$	$\frac{9899}{10,000}$ (True negative)	$\frac{1}{100}$ (False positive)	$\bar{A}$
1	$\frac{9899}{10,000}$	$\frac{1}{10,000}$	

הה הינה איזה גורם שגורם לטעות?

: אנו

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{10,000}}{\frac{1}{100}} = \frac{1}{100}$$

. ב

( $\frac{1}{20} = \frac{50}{10,000}$ ) *ובכן אם לא תרמז 200 מקרים מתקיימים*  
*בנוסף מתקיימים 50, 10,000 מקרים*  
*ולפיכך מתקיימים 200 מקרים*  
*הנכונים מתקיימים 197 מקרים*

	$\bar{B}$ <small>(False negative)</small>	$B$ <small>(True positive)</small>	A
$\bar{B}$	$0$	$\frac{1}{200}$	$\frac{1}{200}$
$B$	$\frac{197}{200}$ <small>(True negative)</small>	$\frac{1}{100}$ <small>(False positive)</small>	$\bar{A}$
1	$\frac{197}{200}$	$\frac{3}{200}$	

$$P(\bar{A}/B) = \frac{P(\bar{A} \cap B)}{P(B)} = \frac{\frac{1}{200}}{\frac{3}{200}} = \frac{1}{3}$$

h re

T 1 IN CN

## Random Variables

$\frac{1}{3}$  : 3 - 2 first round 100%  
 $\cdot \frac{1}{3}$  : 3 - 2 first of same 100%

: 1st round 100% P

$$6\$ \cdot \frac{1}{3} - 3\$ \cdot \frac{2}{3} = 0\$$$

: 12 - 2nd round 100%  
 ②

$$6 \cdot \frac{1}{25} = \frac{6}{25} \left\{ \begin{array}{ll} 5, 8 & 3, 10 \\ 5, 9 & 4, 9 \\ 5, 10 & 4, 10 \end{array} \right.$$

$$6 \cdot \frac{1}{25} = \frac{6}{25} \left\{ \begin{array}{ll} 5, 7 & 2, 10 \\ 3, 9 & 4, 8 \end{array} \right.$$

: 12 - 3rd round 100%

$$1 - \frac{6}{25} - \frac{6}{25} = \frac{18}{25}$$

מבחן הילbert

$$5 \$ \cdot \frac{6}{25} - 6 \$ \cdot \frac{15}{25} = -2.4 \$$$

⑤ נספחים נבדקה:

$$0.4 \cdot 200 = 80$$

:  $p = 0.4$  מבחן הילbert מודולרי

$$\mu = np = 0.4 \cdot 8 = 3.2 \leftarrow \text{mean}$$

$$std = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.4(1-0.4)}{8}} \approx 0.2$$

$$std = 2000 \text{ ₪} \quad \text{mean} = 26000 \text{ ₪}$$

⑤

$$P(26 < X < 30) = P\left(\frac{26-26}{2} < \frac{X-\mu}{\sigma} < \frac{30-26}{2}\right)$$

$$P(26 < X < 30) = P(-0 < Z < 2) = 0.48$$

$$P(26 < X < 30) = 48\%$$

10%

: probabile reale reale reale ⑤

$$\frac{0.6}{2} = 0.3$$

: Probabile reale reale ⑥

: probabile reale reale reale reale

$$\frac{\frac{60}{100} \cdot 500}{500} = \frac{6}{10} = 0.6$$

: Probabile reale

$$\rho \left( \begin{array}{c} \text{probabile} \\ \text{reale} \end{array} \right) = \left( \begin{array}{c} 0.6 \\ 0.4 \end{array} \right) \cdot 0.6^3 \cdot (0.4)^1$$

$$= 0.6 \cdot 0.216 \cdot 0.4 = \frac{216}{625}$$

⑦

$$-10 \cdot 0.1 + (-5) \cdot 0.35 \times 0.1 + 5 \cdot 0.35 + 10 \cdot 0.1$$

$$= 0$$

