Machine learning

Naïve Bayes

Exercise IV

פיתוח:
ד"ר יהונתן שלר
משה פרידמן

מהתפלגות לפונקציית צפיפות

מושגים - תזכורת

משתנה מקרי: פונקציה המתאימה כל אירוע אפשרי במרחב הסתברות לערך מספרי. אצלינו – מאפיין. דוגמאות:

- מ"מ בדיד: זריקת מטבע אקראית. נוצר מ"מ בדיד בינארי. התאמת צד מטבע לערך 0, וצדו השני לערך 1
- מ"מ רציף: גובהו של אדם שנבחר באקראי הוא גם כן משתנה מקרי.

מרחב המדגם Ω: קבוצת כל התוצאות האפשריות בניסוי. אצלינו – אוסף הערכים המאפיינים האפשריים. דוגמאות:

- זריקת מטבע אקראית. מרחב המדגם: $\{0,1\}$
- טמפרטורה של מים. מרחב המדגם $[0,100]$

מאורע / תצפית: תוצאה נצפת מסוימת בניסוי מסוים. אצלינו – ערך מאפיין של דוגמה ב-dataset. דוגמאות:

- התוצאה 3 בזריקת קוביה
- גובה 1.72 של סטודנט

מושגים - תזכורת

הסתברות מאורע: מידת הסבירות שמאורע מסוים יתרחש.

- ההסתברות של מאורע יכולה לקבל ערך מספרי שבין 0 ל-1.

פונקציית צפיפות הסתברות (של משתנה מקרי) [PDF]: פונקציה המתארת את צפיפות המשתנה בכל נקודה במרחב המדגם.

- במ"מ בדיד - הצפיפות בנקודה מסוימת היא בעצם ההסתברות של המאורע (פונקצית המסה). סך כל הערכים שבפונקציית הצפיפות = 1
- במ"מ רציף – פונקציית הצפיפות לא שווה להסתברות של קיום אירוע. אפשר לראות את ה-PDF במ"מ רציף כסבירות היחסית שערך שייך להסתברות. ערכיו אי שליליים, אך לא מוגבלים ל-1 (כמו במ"מ בדיד).

פונקצית ההתפלגות המצטברת (של משתנה מקרי) [CDF]: פונקציה הקובעת את ההסתברות למאורעות X<=a, (לכל a ממשי).

- נדרשת עבור מ"מ רציף

ההתפלגות (של משתנה מקרי): קובעת מהי פונקצית הצפיפות (ומהי ההסתברות של כל מאורע).

- במשתנה מקרי בדיד בעל אוכלוסיה סופית (או במדגם-train-set) נחשב את ההסתברות (הצפיפות) בנקודה מסוימת כמספר המופעים של האירוע לחלק לסך כמות האירועים.
- במשתנה מקרי רציף, נמדדת בד"כ כפונקציה של הממוצע וסטיית התקן (דוגמאות בהמשך).

תרגיל 2 – מ"מ בדיד - הסתברות בסיסית – קריאת נתונים מה-train-set (או המדגם)

תרגיל 2 – תמונות מכוניות – הסתברויות בסיסיות

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

שאלה: מהי ההסתברות להמצאות מכונית אדומה?

- תשובה: $p(Color = red) = \frac{5}{10} = 0.5$

שאלה: מהי ההסתברות להמצאות מכונית ספורט?

- תשובה: $p(Type = Sports) = \frac{6}{10} = 0.6$

מושגים - תזכורת

התפלגות אחידה: התפלגות בה הצפיפות (סבירות) לכל מאורע היא זהה.

- **התפלגות אחידה בדידה**: ההסתברות שווה ל-1 חלקי מספר הערכים האפשריים במרחב המדגם. לדוגמה: הסתברות 1/6 לקבלת הערך 4 בקובייה הוגנת.
- **התפלגות אחידה רציפה**. לדוגמה: נניח ש-מ"מ X מתפלג באופן אחיד בקטע [0, 1]. אז פונקציית ההתפלגות המצטברת שלו:

$$F(x) = \begin{cases} 0 & : \ x < 0 \\ x & : \ 0 \leq x < 1 \\ 1 & : \ x \geq 1. \end{cases}$$

התפלגות ברנולי: מ"מ בדיד בינארי, עם מרחב המדגם: {0,1}. 1 – מסמן הצלחה ו-0 מסמן כישלון. אם סיכוי ההצלחה הוא p, סיכוי הכישלון הוא q=1-p.

מושגים - תזכורת

תוחלת: מייצגת תוצאה "צפויה" (Expected) של ניסוי זהה החוזר על עצמו פעמים רבות.

- עבור משתנה מקרי בדיד: $\mu = E[X] = \sum_{x \in A} P(X = x)x$ ה: התוחלת של קובייה הוגנת = 3.5
- עבור משתנה מקרי רציף: $\mu = \int x\, f(x)\, dx$

שונות: מדד לפיזור ערכים באוכלוסייה נתונה ביחס לתוחלת שלה.

$$\mathrm{Var}(X) = \mathbb{E}((X-\mu)^2) = \mathbb{E}(X^2) - \mu^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$

- מ"מ בדיד, אם האוכלוסייה בגודל N:

$$\mathrm{Var}(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \mu^2$$

- מ"מ רציף:

$$\mathrm{Var}(X) = \sigma^2 = \int (x-\mu)^2\, f(x)\, dx = \int x^2\, f(x)\, dx - \mu^2$$

סטיית תקן: שורש השונות.

התפלגות במדגם

מדגם (sample): מדגם הוא קבוצת פרטים, המהווה מודל לאוכלוסייה, שאליה היא שייכת. אצלינו – ה-train-set.

ממוצע במדגם: $$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

סטיית התקן במדגם: $$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

התפלגות t: התפלגות המבוססת על מידע שנאסף במדגם.

- התפלגות t שואפת להתפלגות z, כאשר גודל המדגם שואף לאינסוף.

תרגיל 3 - התפלגות מותנית - תרגיל בסיסי

$$P(X = x \mid Y = y) = P(X = x, Y = y)/P(Y = y)$$

תרגיל 3 - התפלגות מותנית - תרגיל בסיסי

$$P(X = x \mid Y = y) = P(X = x, Y = y)/P(Y = y)$$

	$Y = 0$	$Y = 1$	
X=0	1/9	2/9	1/3
X=1	2/9	4/9	2/3
	1/3	2/3	1

תזכורת - $P(X = x \mid Y = y) = P(X = x, Y = y)/P(Y = y)$

א. מהי ההסתברות של $P(X = 1, Y = 0)$, איך היא נקראת?

ב. מהי ההסתברות של $P(Y = 1)$? איך היא נקראת?

ג. מהי ההסתברות של $P(X = 0 \mid Y = 1)$, איך היא נקראת?

תרגיל 3 - התפלגות מותנית - תרגיל בסיסי - פתרון

	$Y=0$	$Y=1$	
X=0	1/9	2/9	1/3
X=1	2/9	4/9	2/3
	1/3	2/3	1

תזכורת - $P(X = x \mid Y = y) = P(X = x, Y = y)/P(Y = y)$

א. מהי ההסתברות של $P(X = 1, Y = 0)$, איך היא נקראת?

ב. מהי ההסתברות של $P(Y = 1)$? איך היא נקראת?

ג. מהי ההסתברות של $P(X = 0 \mid Y = 1)$, איך היא נקראת?

פתרון:

א. התפלגות משותפת. $P(X = 1, Y = 0)=2/9$

ב. ההתפלגות הבלתי תלויה. $P(Y = 1)=2/3$

ג. התפלגות מותנית. $P(X = 0 \mid Y = 1)=(2/9)/(2/3)=1/3$

תרגיל 4 – חוק בייס והנחת חוסר התלות

תרגיל 4

	Age	Hobby	Weather	Buy Computer?
1	Young	Sport	Cold	“Yes”
2	Young	Sport	Cold	“Yes”
3	Young	Sport	Cold	“No”
4	Old	Sport	Hot	“Yes”
5	Old	Sport	Hot	“Yes”
6	Old	Paint	Hot	“No”
7	Old	Paint	Cold	“Yes”
8	Old	Paint	Cold	“Yes”
9	Young	Paint	Hot	“No”
10	Young	Sport	Hot	“No”

נתונה קבוצת האימון הבאה:

נדרשת הערכת Yes/No עבור:

Age = “Young”
Hobby = “Paint”
Weather = “Cold”

תרגיל 4 - פתרון

- אנו צריכים לחשב את ההסתברויות הבאות
- P(“yes” | Age = “Young”, Hobby = “Paint”, Weather = “Cold”)
- P(“no” | Age = “Young”, Hobby = “Paint”, Weather = “Cold”)

- ונבחר את ההערכה עם ההסתברות הגבוהה יותר

MAP = Maximum a posteriori (estimation)

- בהינתן ווקטור לסיווג – $(x_1,x_2,x_3.....x_n)$, נעריך את ההסתברות עבור כל סיווג c_i השייך לקבוצה C ונבחר את הסיווג עם ההסתברות הגבוהה ביותר.
- $P(c_1 | x_1,x_2,x_3.....x_n)$
- $P(c_2 | x_1,x_2,x_3.....x_n)$
- $P(c_3 | x_1,x_2,x_3.....x_n)$
- ...

כלומר, נבחר את הקטגוריה c, המקיימת

$$h_{MAP} = \arg\max_{c \in C} P(c \mid X)$$

תרגיל 4 - פתרון

$$h_{MAP} = \arg\max_{c \in C} P(c \mid X)$$

- אנו צריכים לחשב את ההסתברויות הבאות
- P(“yes” | Age = “Young”, Hobby = “Paint”, Weather = “Cold”)
- P(“no” | Age = “Young”, Hobby = “Paint”, Weather = “Cold”)

- ונבחר את ההערכה עם ההסתברות הגבוהה יותר
- נשתמש בהנחת בייס לאי תלות בין המאפיינים

חוק בייס והנחת חוסר התלות

Class prior

Likelihood probability

חוק בייס:

$$P(c \mid x_1, x_2, \ldots\ldots x_D) = \frac{P(c)P(x_1, x_2, \ldots\ldots x_D \mid c)}{P(x_1, x_2, \ldots\ldots x_D)}$$

a posteriori probability

Feature (predictor) priors

בגלל הנחת חוסר התלות בין המאפיינים:

$$P(x_1, x_2, \ldots\ldots x_D \mid c) = P(x_1 \mid c)P(x_2 \mid c)P(x_3 \mid c)\ldots P(x_D \mid c) = \prod_{i=1}^{D} P(x_i \mid c)$$

תרגיל 4 - פתרון

נפצל לשתי טבלאות:
טבלת YES
טבלת NO

	Age	Hobby	Weather	Buy Computer?
1	Young	Sport	Cold	“Yes”
2	Young	Sport	Cold	“Yes”
3	Old	Sport	Hot	“Yes”
4	Old	Sport	Hot	“Yes”
5	Old	Paint	Cold	“Yes”
6	Old	Paint	Cold	“Yes”

	Age	Hobby	Weather	Buy Computer?
1	Young	Sport	Cold	“No”
2	Old	Paint	Hot	“No”
3	Young	Paint	Hot	“No”
4	Young	Sport	Hot	“No”

תרגיל 4 – פתרון
נחשב את ההסתברות ל- "yes"

$$P(c \mid x_1, x_2, \ldots x_D) = \frac{P(c)P(x_1, x_2, \ldots x_D \mid c)}{P(x_1, x_2, \ldots x_D)}$$

$$P(\text{“yes”} \mid \text{Age} = \text{“Young”}, \text{Hobby} = \text{“Paint”}, \text{Weather} = \text{“Cold”}) =$$

$$P(\text{"yes"}) \times P(\text{Age} = \text{“Young”}, \text{Hobby} = \text{“Paint”}, \text{Weather} = \text{“Cold”} \mid \text{"yes"})/K$$

$$P(\text{"yes"}) \times P(\text{Age} = \text{“Young”} \mid \text{"yes"}) \times P(\text{Hobby} = \text{“Paint”} \mid \text{"yes"}) \times P(\text{Weather} = \text{“Cold”} \mid \text{"yes"})/K$$

$$K = P(\text{Age} = \text{“Young”}, \text{Hobby} = \text{“Paint”}, \text{Weather} = \text{“Cold”})$$

$$P(Age = "Young" \mid "yes") = P(Age = "Young" \wedge "yes") / P("yes") = \frac{2}{6} = \frac{1}{3}$$

$$P(Hobby = "Paint" \mid "yes") = P(Hobby = "Paint" \wedge "yes") / P("yes") = \frac{2}{6} = \frac{1}{3}$$

$$P(Weather = "cold" \mid "yes") = P(Weather = "cold" \wedge "yes") / P("yes") = \frac{4}{6} = \frac{2}{3}$$

Prior class distribution: $P("yes") = 0.6$

$$(\text{"yes"}) \times P(\text{Age} = \text{“Young”} \mid \text{"yes"}) \times P(\text{Hobby} = \text{“Paint”} \mid \text{"yes"}) \times P(\text{Weather} = \text{“Cold”} \mid \text{"yes"})/K$$

$$= \frac{6}{10} \times \frac{1}{3} \times \frac{1}{3} \times \frac{2}{3} = \frac{12}{270} = 0.04$$

$$P(x_1, x_2, \ldots x_D \mid c) = P(x_1 \mid c)P(x_2 \mid c)P(x_3 \mid c) \ldots P(x_D \mid c) = \prod_{i=1}^{D} P(x_i \mid c)$$

תרגיל 4 – פתרון
נחשב את ההסתברות ל- "no"

P("no" | Age = "Young", Hobby = "Paint", Weather = "Cold") =
P("no") × P(Age = "Young", Hobby = "Paint", Weather = "Cold"|"no")/K
P("no") × P(Age = "Young"|"no") × P(Hobby = "Paint"|"no") × P(Weather = "Cold"|"no")/K

K = P(Age = "Young", Hobby = "Paint", Weather = "Cold")

$$P(Age = "Young" | "no") = P(Age = "Young" \wedge "no") / P("no") = \frac{3}{4}$$

$$P(Hobby = "Paint" | "no") = P(Hobby = "Paint" \wedge "no") / P("no") = \frac{2}{4}$$

$$P(Weather = "cold" | "no") = P(Weather = "cold" \wedge "no") / P("no") = \frac{1}{4}$$

Prior class distribution: $P("no") = 0.4$

$$("no") \times P(Age = \text{“Young”} | "no") \times P(Hobby = \text{“Paint”} | "no") \times P(Weather = \text{“Cold”} | "no")/K$$

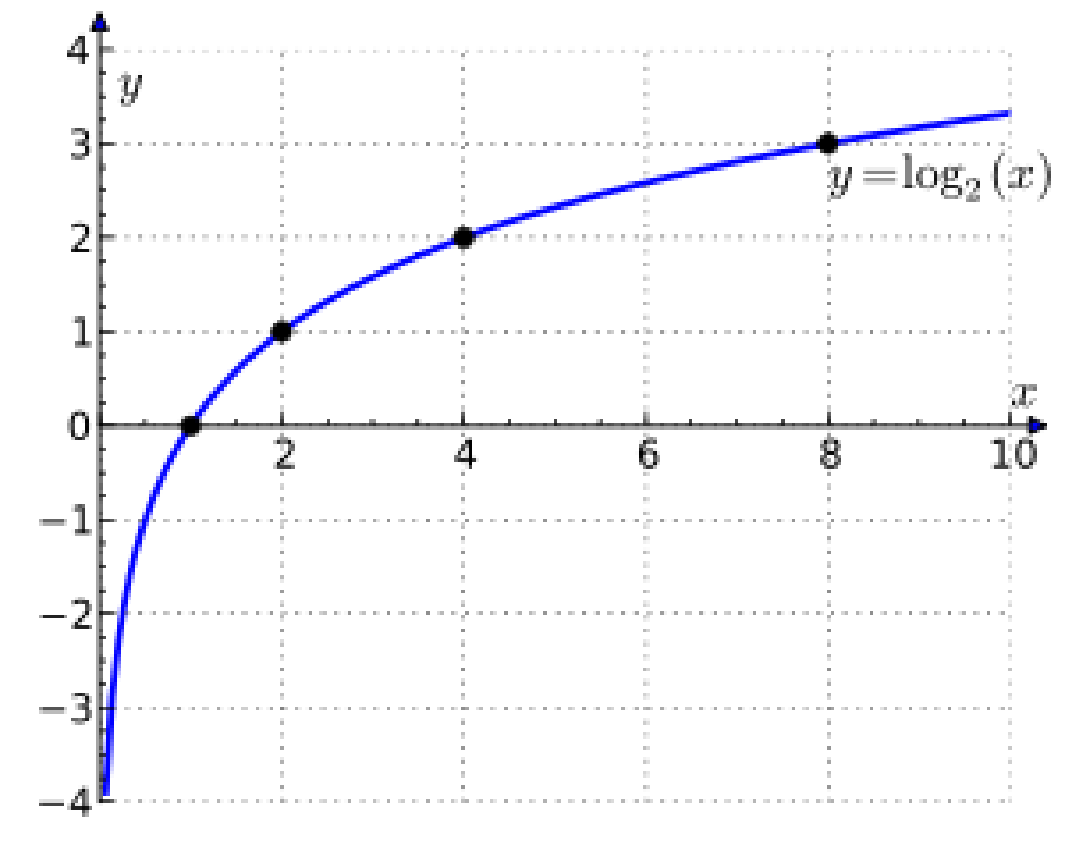
$$= \frac{4}{10} \times \frac{3}{4} \times \frac{2}{4} \times \frac{1}{4} = \frac{24}{640} = 0.0375$$

תרגיל 4 – פתרון בחירת הקטגוריה

- P(“yes” | Age = “Young”, Hobby = “Paint”, Weather = “Cold”) =0.04
- P(“no” | Age = “Young”, Hobby = “Paint”, Weather = “Cold”) = 0.0375

- Predict - YES

פונקצית לוג



תכונות:

- פונקציית לוג של שברים תהיה שלילית, אך היא שומרת על הסדר, והיא גם מונוטונית עולה.
- log(x*y)=log(x)+log(y).

לכן, נרצה לעבוד עם חיבור לוגים, במקום מכפלת שברים (של הסתברויות).

מדוע?

תרגיל 4ב - סימולציית סיווג – מ"מ בדיד - פתרון

שלב הסיווג (המשך):

- במקום חישוב מכפלת הסתברויות יכולנו גם להוציא log

("yes") × P(Age = "Young"|"yes") × P(Hobby = "Paint"|"yes") × P(Weather = "Cold"|"yes")/K =

$$= \frac{6}{10} \times \frac{1}{3} \times \frac{1}{3} \times \frac{2}{3} = \frac{12}{270} = 0.04$$

וכעת נשתמש ב-log

log(0.6)+log(0.333)+log(0.333)+ log(0.666)=**-1.352**

("no") × P(Age = "Young"|"no") × P(Hobby = "Paint"|"no") × P(Weather = "Cold"|"no")/K =

$$= \frac{4}{10} \times \frac{3}{4} \times \frac{2}{4} \times \frac{1}{4} = \frac{24}{640} = 0.0375$$

וכעת נשתמש ב-log

log(0.4)+log(0.75)+log(0.5)+log(0.25)=-1.4259

תרגיל 5 - סימולציית סיווג – מ"מ בדיד

מסווג Naïve Bayes עבור מ"מ בדיד - תזכורת

- Train Naïve Bayes (given data for X and Y)

  for each* value y_k

  estimate $\pi_k \equiv P(Y = y_k)$

  for each* value x_{ij} of each attribute X_i

  estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \theta_{ijk}$$

* probabilities must sum to 1, so need estimate only n-1 of these...

תרגיל 5 - סימולציית סיווג – מ"מ בדיד

- נתונות 10 דוגמאות של רכבים, צריך לבנות מסווג שיחליט אם רכב מסוג Red, Domestic, SUV יש סבירות גבוהה שיגנב. להלן הנתונים

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

תרגיל 5 - סימולציית סיווג – מ"מ בדיד
מה עושים?

- אין נתון כזה בטבלה, צריך לחשב הסתברויות:

- נחפש לחשב את $P(yes \mid Red \,\&\, SUV \,\&\, Domestic)$

$P(no \mid Red \,\&\, SUV \,\&\, Domestic)$.

ונמצא מה יותר סביר..

איך עושים זאת?

תרגיל 5 - סימולציית סיווג – מ"מ בדיד
מה עושים?

P(yes | Red & SUV & Domestic) = P(yes) * P(red | yes) * P(SUV | yes) * P(Domestic | Yes)

ובאופן דומה לגבי ההסתברות ל"לא".

P(yes) = 5/10 = 0.5

P(red|yes) = 3/5 = 0.6

P(suv|yes) = 1/5 = 0.2

P(domestic|yes) = 2/5 = 0.4

P(no) = 5/10 = 0.5

P(red|no) = 2/5 = 0.4

P(suv|no) = 3/5 = 0.6

P(domestic|no) = 3/5 = 0.6

תרגיל 5 - סימולציית סיווג – מ"מ בדיד
מה עושים? המשך..

P(yes | Red & SUV & Domestic) = P(yes) * P(red | yes) * P(SUV | yes) * P(Domestic | Yes) = 0.5 * .6 * .2 * .4 = 0.024

ובאופן דומה לגבי ההסתברות ל"לא".

P(no | Red & SUV & Domestic) = P(no) * P(red | no) * P(SUV | no) * P(Domestic | no) = 0.5 * .4 * .6 * .6 = 0.072

ומכאן ש:

P(no) > P(yes) and thus classified as “no”

תרגיל 5ב - סימולציית סיווג – מ"מ רציף - פתרון

שלב הסיווג (המשך):

- במקום חישוב מכפלת הסתברויות יכולנו גם להוציא log

P(yes | Red & SUV & Domestic) = P(yes) * P(red | yes) * P(SUV | yes) * P(Domestic | Yes)=

= 0.5 * 0.6 * 0.2 * 0.4 = 0.024

וכעת נשתמש ב-log

log(0.5)+log(0.6)+log(0.2)+ log(0.4) = -1.619

P(no | Red & SUV & Domestic) = P(no) * P(red | no) * P(SUV | no) * P(Domestic | no) =

= 0.5 * 0.4 *0.6 * 0.6 = **0.072**

וכעת נשתמש ב-log

log(0.5)+log(0.4)+log(0.6)+ log(0.6) = **-1.142**

תרגול שערוך המודל (evaluation)

תרגיל 6 – שיערוך המודל

- נדרשתם לכתוב מסווג לזיהוי סטודנטים שיסיימו בהצטיינות קורס "למידת מכונה".
- אימנתם את המסווג מבדיקת המסווג על נתוני הסטודנטים משנת תש"פ (2019-2020) ובדקתם על נתוני תשפ"א (2020-2021).
- גיליתם שמתוך 215 סטודנטים שלקחו את הקורס, המסווג אמר על 55 שיסיימו בהצטיינות, אך בפועל רק 40 מתוכם אכן סיימו בהצטיינות והיו עוד 5 אחרים שהמסווג פספס.
- בנו confusion matrix למסווג הנ"ל
- חשבו עבורו accuracy , ו- error-rate
- חשבו precision ו- recall (לערך החיובי של הקטגוריה)

	Predicted Yes	Predicted No
Actual Yes	40	5
Actual No	15	155

Accuracy = (40+155)/215 = *90.7%*

error-rate = (5+15)/215 = *9.3%*

Precision = *40/(40+15)* = *72.7%*

Recall = *40/(40+5)* =*88.9%*

תרגיל 7 – שיערוך המודל

- חשב confusion matrix למסווג הר"מ וחשב עבורו accuracy, precision, error_rate ו recall.

- Y Actual

1	0	1	0	1	0	1	0	1	0

- Y Predicted

1	1	1	1	1	1	1	0	0	0

	Predicted Yes	Predicted No
Actual Yes	4	1
Actual No	3	2

Accuracy = (4+2)/10 = 60%
Error Rate = (1+3)/10 = 40%
Precision = 4/(4+3) = 57.1%
Recall = 4/(4+1) =80%