Machine learning

Naïve Bayes – part 2

* Smoothing
* Continuous features

Exercise V

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

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).

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:

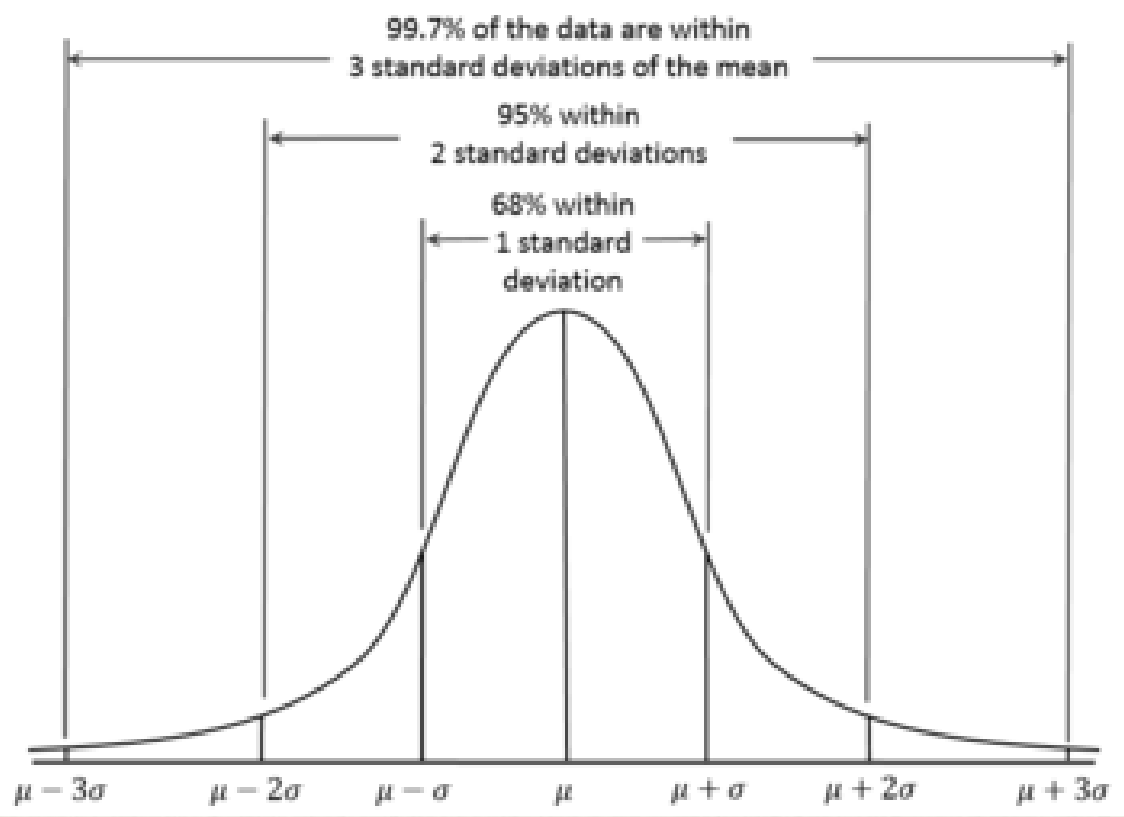
$$\mathrm{Var}(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \mu^2$$

- 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 נורמלית



התפלגות נורמלית: נקראת גם גאוסיאן (Gaussian) או עקומת פעמון.

- פונקציית צפיפות סמטרית.

התפלגות z: תת קבוצה של התפלגות נורמלית בו התוחלת/הממוצע=0 וסטיית התקן=1.

- כל התפלגות נורמלית ניתן להפוך להתפלגות z

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.

חזרה להתפלגות נורמלית

התפלגות נורמלית: נקראת גם גאוסיאן (Gaussian) או עקומת פעמון.

- פונקציית צפיפות סמטרית.

התפלגות z: תת קבוצה של התפלגות נורמלית בו התוחלת/הממוצע=0 וסטיית התקן=1.

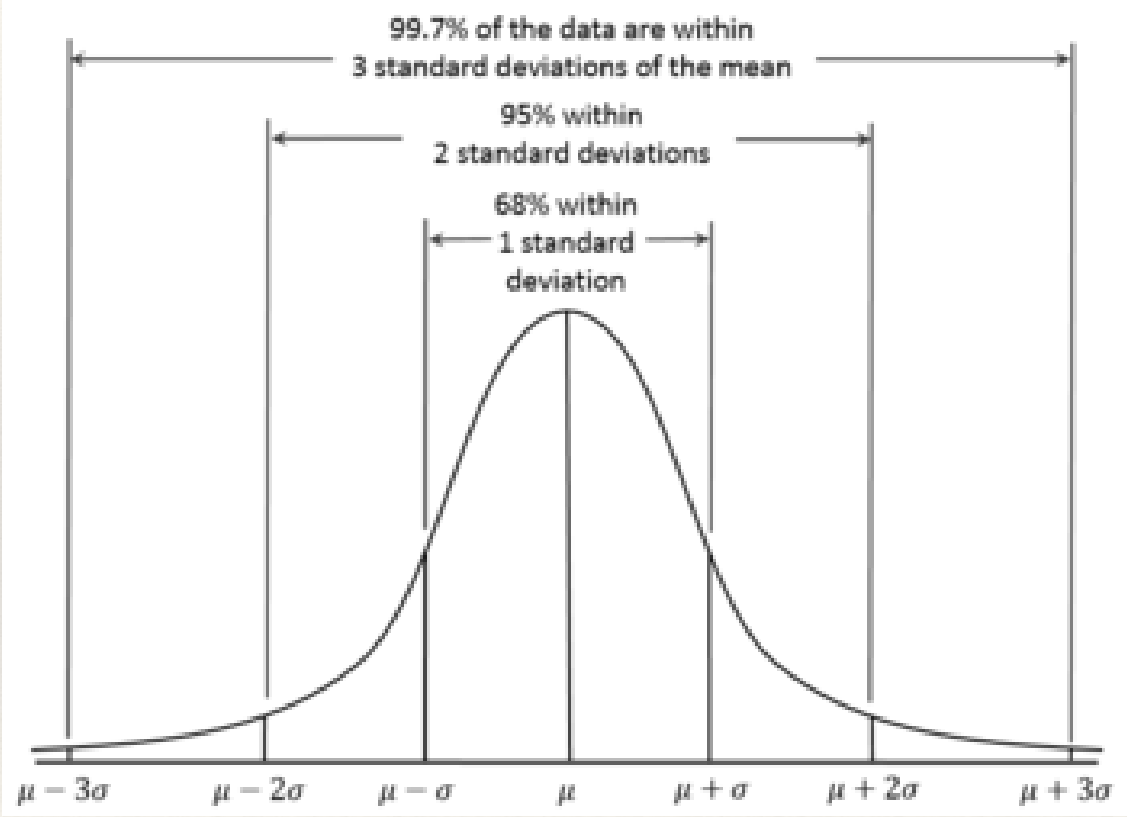
- כל התפלגות נורמלית ניתן להפוך להתפלגות z
- **מהתפלגות נורמלית להתפלגות z:** z-val=(x-μ)/σ

תוחלת/ממוצע ערך מאורע

פונקצית צפיפות:

$$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

סטיית תקן



הערה חשובה: אנחנו נחשב סטיית תקן במדגם ואת ה-t-val

תזכורת - *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

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

חוק בייס והנחת חוסר התלות - תזכורת

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 8 - סימולציית סיווג – מ"מ רציף

מסווג Gaussian Naïve Bayes עבור מ"מ רציף - תזכורת

- Train Naïve Bayes (examples)

  for each value y_k

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

  for each attribute X_i estimate $P(X_i|Y = y_k)$

  - class conditional mean μ_{ik}, standard deviation σ_{ik}

- Classify (X^{new})

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

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \mathcal{N}(X_i^{new}; \mu_{ik}, \sigma_{ik})$$

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

עבור Gaussian Naïve Bayes, נשתמש בפונקציית צפיפות:

$$P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} \; e^{\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

תרגיל 8 - סימולציית סיווג – מ"מ רציף

outlook	temperature	humidity	windy	play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

- נתונות 14 דוגמאות של טניסאי שהיה צריך להחליט אם לשחק טניס ביום מסוים.
- בנו מסווג שיחזה האם האדם ישחק טניס ביום עם התנאים הבאים:

outlook=overcast,
temperature=60,
humidity=62,
windy=false.

תרגיל 8 - סימולציית סיווג – מ"מ רציף - פתרון

outlook	temperature	humidity	windy	play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

שלב האימון:

<u>א. חישוב הסתברויות priors</u> של המחלקות:

- p(yes) = 9 / (9+5) = 0.643
- p(no) = 5 / 14 = 0.357

תרגיל 8 - סימולציית סיווג – מ"מ רציף - פתרון

outlook	temperature	humidity	windy	play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

שלב האימון:

<u>א. חישוב הסתברויות priors</u> של המחלקות:

- p(yes) = 0.643
- p(no) = 0.357

תרגיל 8 - סימולציית סיווג – מ"מ רציף - פתרון

outlook	temperature	humidity	windy	play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

הווקטור החדש:

outlook=overcast,
temperature=60,
humidity=62,
windy=false.

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

שלב האימון (נחשב רק עבור הערכים שבווקטור החדש):

ב. חישוב ההסתברויות המותנות הבדידות:

P(outlook=overcast|yes)=
P(windy=false|yes)=

P(outlook=overcast|no)=
P(windy=false|no)=

תרגיל 8 - סימולציית סיווג – מ"מ רציף - פתרון

outlook	temperature	humidity	windy	play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

הווקטור החדש:

outlook=overcast,
temperature=60,
humidity=62,
windy=false.

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

שלב האימון (נחשב רק עבור הערכים שבווקטור החדש):

ב. חישוב ההסתברויות המותנות הבדידות:

P(outlook=overcast|yes)=4/9 ~ 0.44
P(windy=false|yes)=6/9 ~ 0.66

P(outlook=overcast|no)=0/5=0
P(windy=false|no)=2/5=0.4

תרגיל 8 - סימולציית סיווג – מ"מ רציף - פתרון

הווקטור החדש:
outlook=overcast,
temperature=60,
humidity=62,
windy=false.

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

שלב האימון:

ב. חישוב ההסתברויות המותנות הבדידות:

שימו לב, שיש לנו כאן הסתברות 0

- נשתמש בהחלקה עם הפרמטרים: p= 1/3, m=3

P(outlook=overcast|yes)=4/9 ~ 0.44
P(windy=false|yes)=6/9 ~ 0.66

P(outlook=overcast|no)=0/5=0
P(windy=false|no)=2/5=0.4

$$P(X = x \mid Y = y) = \frac{n_c + mp}{n + m}$$

$P(outlook = overcast \mid yes) =$

$P(outlook = overcast \mid no) =$

תזכורת - Smoothing solution

- Probability estimates are adjusted or *smoothed*.
- Assumes that each feature is given a prior probability, *p*, that is assumed to have been previously observed in a "virtual" sample of size *m*.
- Usually, in the binary case *p* is simply assumed to be 0.5

$$P(X = x \mid Y = y) = \frac{n_c + mp}{n + m}$$

- n = number of training examples for which Y = y
- n_c =number of examples where *X=x* and Y=y
- p = a prior estimation for P(X=x|Y=y)
- m = the equivalent sample size

תרגיל 8 - סימולציית סיווג – מ"מ רציף - פתרון

הווקטור החדש:
outlook=overcast,
temperature=60,
humidity=62,
windy=false.

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

שלב האימון:

ב. חישוב ההסתברויות המותנות הבדידות:

שימו לב, שיש לנו כאן הסתברות 0

- נשתמש בהחלקה עם הפרמטרים: p= 1/3, m=3

P(outlook=overcast|yes)=4/9 ~ 0.44
P(windy=false|yes)=6/9 ~ 0.66

P(outlook=overcast|no)=0/5=0
P(windy=false|no)=2/5=0.4

$$P(X = x \mid Y = y) = \frac{n_c + mp}{n + m}$$

$$P(outlook = overcast \mid yes) = \frac{4+1}{9+3} = \frac{5}{12} = 0.4167$$

$$P(outlook = overcast \mid no) = \frac{0+1}{5+3} = \frac{1}{8} = 0.125$$

תרגיל 8 - סימולציית סיווג – מ"מ רציף - פתרון

outlook	temperature	humidity	windy	play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

שלב האימון:

<u>ג. חישוב ההסתברויות המותנות הרציפות:</u>

(יש לחשב בנפרד עבור המחלקות השונות)

Temperature: μ_temp_yes= , σ_temp_yes=

μ_temp_no= , σ_temp_no=

Humidity: μ_hum_yes= , σ_temp_yes=

μ_hum_no= , σ_temp_no=

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

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

תרגיל 8 - סימולציית סיווג – מ"מ רציף - פתרון

outlook	temperature	humidity	windy	play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

שלב האימון:

<u>ג. חישוב ההסתברויות המותנות הרציפות:</u>

(יש לחשב בנפרד עבור המחלקות השונות)

Temperature:

μ_temp_yes=73, σ_temp_yes=6.2

μ_temp_no=74.6, σ_temp_no=8.0

Humidity:

μ_hum_yes=79.1, σ_temp_yes=10.2

μ_hum_no=86.2, σ_temp_no=9.7

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

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

תרגיל 8 - סימולציית סיווג – מ"מ רציף - פתרון

$$P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} \; e^{\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

$f(temperature = 60 \mid yes) =$

$f(temperature = 60 \mid no) =$

$f(humidity = 62 \mid yes) =$

$f(humidity = 62 \mid no) =$

שלב הסיווג – חישוב פונקצית הצפיפות עבור המ"מ הרציפים:

temperature=60 humidity=62

חישבנו:

μ_temp_yes=73, σ_temp_yes=6.2

μ_temp_no=74.6, σ_temp_no=8.0

μ_hum_yes=79.1, σ_temp_yes=10.2

μ_hum_no=86.2, σ_temp_no=9.7

תרגיל 8 - סימולציית סיווג – מ"מ רציף - פתרון

$$P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} e^{\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

שלב הסיווג – חישוב פונקצית הצפיפות עבור המ"מ הרציפים:

temperature=60 humidity=62

חישבנו:

μ_temp_yes=73, σ_temp_yes=6.2

μ_temp_no=74.6, σ_temp_no=8.0

μ_hum_yes=79.1, σ_temp_yes=10.2

μ_hum_no=86.2, σ_temp_no=9.7

$$f(temperature = 60 \mid yes) = \frac{1}{6.2\sqrt{2\pi}} e^{-\frac{(60-73)^2}{2\ (6.2)^2}} = 0.071$$

$$f(temperature = 60 \mid no) = \frac{1}{8\sqrt{2\pi}} e^{-\frac{(60-74.6)^2}{2\ 8^2}} = 0.0094$$

$$f(humidity = 62 \mid yes) = \frac{1}{10.2\sqrt{2\pi}} e^{-\frac{(62-79.1)^2}{2\ (10.2)^2}} = 0.0096$$

$$f(humidity = 62 \mid no) = \frac{1}{9.7\sqrt{2\pi}} e^{-\frac{(62-86.2)^2}{2\ (9.7)^2}} = 0.0018$$

תרגיל 8 - סימולציית סיווג – פתרון

Classify (X^{new})

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

הווקטור החדש:

outlook=overcast,
temperature=60,
humidity=62,
windy=false.

שלב הסיווג (המשך):

חישבנו ושלפנו את ההסתברויות המותנות ואת ה-priors:

$f(temperature = 60 \mid yes) = 0.071$

$f(temperature = 60 \mid no) = 0.0094$

$f(humidity = 62 \mid yes) = 0.0096$

$f(humidity = 62 \mid no) = 0.0018$

- p(yes) = 0.643
- p(no) = 0.357

P(outlook=overcast|no)= *0.125*
P(windy=false|no)=0.4

P(outlook=overcast|yes)~ *0.4167*
P(windy=false|yes)~ 0.66

כעת נחשב את ההסתברויות:

P(yes | outlook=overcast..) =

P(no | outlook=overcast..) =

תרגיל 8 - סימולציית סיווג – פתרון

Classify (X^{new})

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

הווקטור החדש:

outlook=overcast,
temperature=60,
humidity=62,
windy=false.

שלב הסיווג (המשך):

חישבנו ושלפנו את ההסתברויות המותנות ואת ה-priors:

$f(temperature = 60 \mid yes) = 0.071$

$f(temperature = 60 \mid no) = 0.0094$

$f(humidity = 62 \mid yes) = 0.0096$

$f(humidity = 62 \mid no) = 0.0018$

- p(yes) = 0.643
- p(no) = 0.357

P(outlook=overcast|no)= *0.125*
P(windy=false|no)=0.4

P(outlook=overcast|yes)~ *0.4167*
P(windy=false|yes)~ 0.66

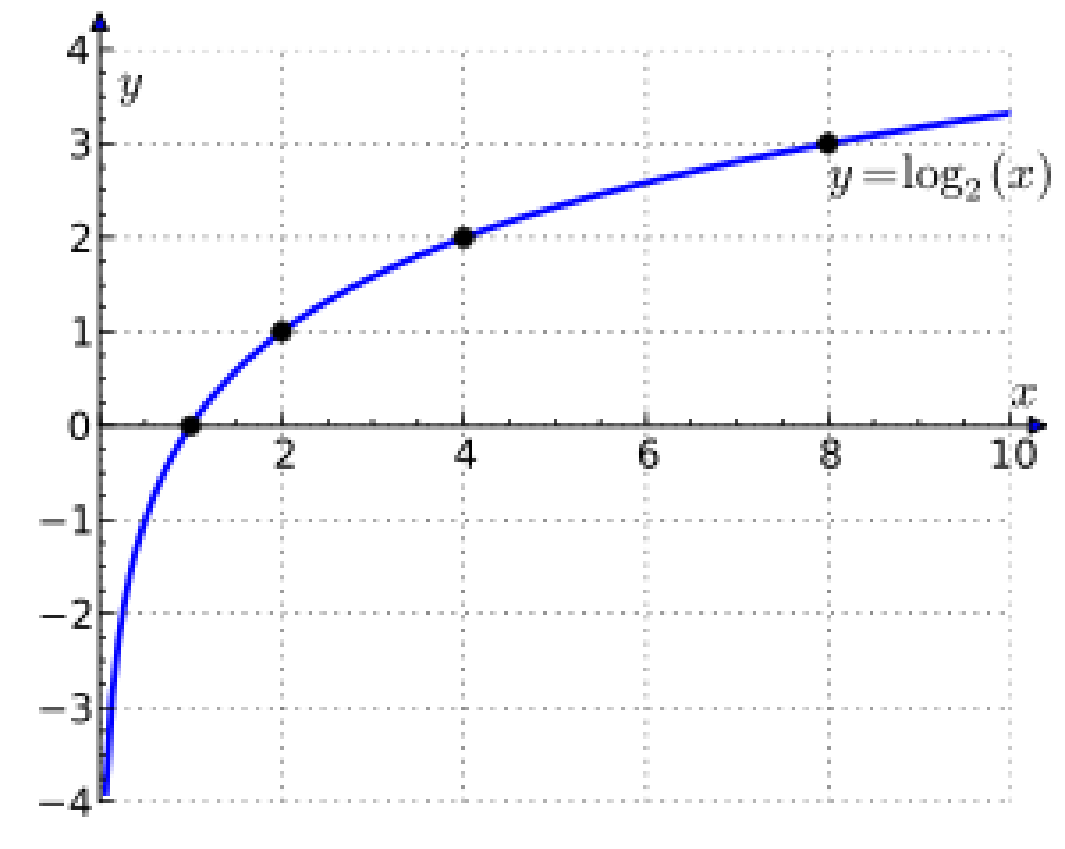
כעת נחשב את ההסתברויות:

P(yes|outlook=overcast..) = p(yes) * p(outlook=overcast|yes)*.. = 0.643*0.4167*0.667*0.071*0.0096= **0.00012**

P(no|outlook=overcast..) = p(no) * p(outlook=overcast|no)*.. = 0.357*0.125*0.4*0.0094*0.0018= 0.000000302

P(yes) is more probable..

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 8 - סימולציית סיווג – פתרון

Classify (X^{new})

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

הווקטור החדש:

outlook=overcast,
temperature=60,
humidity=62,
windy=false.

שלב הסיווג (המשך):

חישבנו ושלפנו את ההסתברויות המותנות ואת ה-priors:

$f(temperature = 60 \mid yes) = 0.071$

$f(temperature = 60 \mid no) = 0.0094$

$f(humidity = 62 \mid yes) = 0.0096$

$f(humidity = 62 \mid no) = 0.0018$

- p(yes) = 0.643
- p(no) = 0.357

P(outlook=overcast|no)= *0.125*
P(windy=false|no)=0.4

P(outlook=overcast|yes)~ *0.4167*
P(windy=false|yes)~ 0.66

כעת נחשב את ההסתברויות:

P(yes|outlook=overcast..) = p(yes) * p(outlook=overcast|yes)*..=log(0.643)+ log(0.4167)+log(0.667)+log(0.071)+log(0.0096) = **-13.003**

P(no|outlook=overcast..) = p(no) * p(outlook=overcast|no)*.. =log(0.357)+ log(0.125)+log(0.4)+log(0.0094)+log(0.0018) = -21.66

P(yes) is more probable..