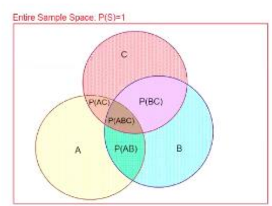
למידה חישובית ממידע: תרגול 3 – MAP Classifier and Bayesian Learning

אלגוריתמים הסתברותיים

האלגוריתם האינטואיטיבי ביותר הוא להחזיר את קלאס הרוב, או במילים אחרות להחזיר את הקלאס הסביר ביותר לדאטה אימון. היום נלמד אלגוריתמים הסתברותיים – אלגוריתמים שמשתמשים בטכניקות הסתברותיות כדי לסווג את הדגימה החדשה לקלאס.

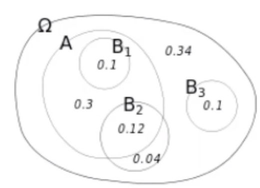
חזרה קצרה בהסתברות:

- מרחב מדגם – הוא קבוצה של מאורעות events, שזו רשימה של כל התוצאות האפשריות של המאורע. מרחב הדגימות יכול להיות גם רציף (גבהים וכו')
- מאורע – תת קבוצה של מרחב המדגם. זריקה של שתי קוביות שהסכום על הקוביות הוא 7.
- משתנה מקרי – פונקציה ממרחב המדגם ומחזירה ערך רציף. למשל סכום של שתי קוביות. עליו ניתן לשאול מה ההסתברות שאיקס שווה לתוצאה מסוימת. למשל מה ההסתברות שאיקס = 1 עבור איקס המתאר את סכום על שתי הקוביות (ההסתברות היא 0).
- על משתנים מקריים אפשר לשאול מה התוחלת שמסומנת ב-miu ביוונית (ה-expected value),

  עבור משתנה בדיד התוחלת הינה
  $$E[X] = \sum_x xp(x)$$
  ועבור משתנה רציף
  $$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$
  כאשר f היא פונקציית הצפיפות.
- שונות – מסומנת על ידי סיגמה בריבוע. מוגדרת להיות: $\sigma^2 = var(X) = E[(x-\mu)^2]$
- סטיית תקן – היא שורש השונות $\sigma = \sqrt{var(X)} = \sqrt{E[(x-\mu)^2]}$
- 

  - $P(A \cup B) = ?$
  $$P(A) + P(B) - P(A \cap B)$$
  - $P(A \cup B \cup C) = ?$
  $$\begin{aligned} &P\big((A \cup B) \cup C\big) = \\ &P(A \cup B) + P(C) - P\big((A \cup B) \cap C\big) = \\ &P(A) + P(B) - P(A \cap B) + P(C) - P\big((A \cap C) \cup (B \cap C)\big) = \\ &P(A) + P(B) - P(A \cap B) + P(C) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) = \\ &P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \end{aligned}$$

חזרה קצרה על הסתברות מותנית



- <u>Conditional probability</u>:
  - $P(A|B) = \frac{P(A \cap B)}{P(B)}$
  - $P(A|B_1) = ?$
    - $\frac{P(A \cap B_1)}{P(B_1)} = \frac{0.1}{0.1} = 1$
  - $P(A|B_2) = ?$
    - $\frac{P(A \cap B_2)}{P(B_2)} = \frac{0.12}{0.16} = 0.75$

<u>דוגמה עבור הסתברות מותנית:</u>

נתון כי ההסתברות לעבור את המבחן היא P(pass) = 90% ולהיכשל = P(fail) 10%. בנוסף ידוע לנו כי:

- $P(Learn\ for\ the\ test|Pass) = 90\%$
- $P(Didn't\ learn|Pass) = 10\%$
- $P(Learn\ for\ the\ test|Fail) = 5\%$
- $P(Didn't\ learn|Fail) = 95\%$

<u>מה ההסתברות שלמדת למבחן?</u> 0.815 להלן החישובים הדרושים לפתרון

- $P(Pass \cap Learn\ for\ the\ test) = P(Pass) \times P(Learn\ for\ the\ test|Pass) = 90\% \times 90\% = 81\%$
- $P(Pass \cap Didn't\ learn) = P(Pass) \times P(Didn't\ learn|Pass) = 90\% \times 10\% = 9\%$
- $P(Fail \cap Learn\ for\ the\ test) = P(Fail) \times P(Learn\ for\ the\ test|Fail) = 10\% \times 5\% = 0.5\%$
- $P(Fail \cap Didn't\ learn) = P(Fail) \times P(Didn't\ learn|Fail) = 10\% \times 95\% = 9.5\%$
- $P(Learn\ for\ the\ test) = P(Pass \cap Learn\ for\ the\ test) + P(Fail \cap Learn\ for\ the\ test) = 81\% + 0.5\% = 81.5\%$

מתבקש לשאול, ולהלן החישובים

- מה ההסתברות שסטודנט עבר בהינתן שלמד?
- מה ההסתברות שסטודנט נכשל בהינתן שלמד?
- מה ההסתברות שסטודנט עבר בהינתן כי לא למד?
- ומה ההסתברות שסטודנט נכשל בהינתן כי לא למד?

- $P(Pass|Learn\ for\ the\ test) = \frac{P(Pass \cap Learn\ for\ the\ test)}{P(Learn\ for\ the\ test)} = \frac{81\%}{81.5\%} = 99\%$
- $P(Fail|Learn\ for\ the\ test) = \frac{P(Fail \cap Learn\ for\ the\ test)}{P(Learn\ for\ the\ test)} = \frac{0.5\%}{81.5\%} = 1\%$
- $P(Pass|Didn't\ learn) = \frac{P(Pass \cap Didn't\ learn)}{P(Didn't\ learn)} = \frac{9\%}{18.5\%} = 49\%$
- $P(Fail|Didn't\ learn) = \frac{P(Fail \cap Didn't\ learn)}{P(Didn't\ learn)} = \frac{9.5\%}{18.5\%} = 51\%$

<u>נגדיר מאורעות בלתי תלויים:</u>

אם ההסתברות לחיתוך שווה למכפלת ההסתברויות אזי המאורעות בלתי תלויים.

המסקנה: זה שאחד מהם קרה, לא גורע ולא מוסיף למאורע השני.

- <u>Independent events</u>
  - If $P(A \cap B) = P(A)P(B)$ then A & B are independent
  - From conditional probability we get:

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

$$\Downarrow$$

$$P(A \cap B) = P(A|B)P(B)$$

  - If A & B are independent:

$$P(A)P(B) = P(A \cap B) = P(A|B)P(B)$$
$$P(A) = P(A|B)$$

* And also $P(B) = P(B|A)$

כעת, נוכל להתקדם למושגים שיותר רלוונטיים עבור Bayes.

הסתברות פריורית / Prior Probability

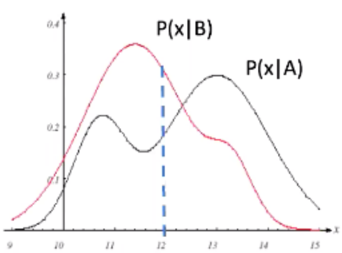
זו היא ההסתברות של המקרים בדאטה שיש לנו, האלגוריתם הנאיבי שתואר בעמוד הקודם משתמש בהסתברות prior. כלומר, החלוקה הפנימית של מאורעות, למשל 40% בנים ו60% בנות.

הסתברות הלייקליהוד / Likelihood Probability

זו היא הסתברות שמותנית על ידי קלאס: ההסתברות של דגימה x בהינתן הקלאס. למשל עבור דגימה x=12 ושני קלאסים אפשריים A ו-B: האדומה היא ההסתברות של B והשחורה היא A. נחשב את ערך ההסתברות בשני הגרפים עבור x=12 ונראה כי ההסתברות יותר גבוהה עבור קלאס B ולכן נסווג B.

אם נחזור לדוגמת עבר/נכשל, הסתברות זו היא כמו לשאול מה ההסתברות שמישהו למד למבחן בהינתן שעבר.

אבל, אנחנו רוצים לדעת מה ההסתברות לעבור או להיכשל בהינתן שלמדנו. אז אנחנו צריכים דרך לעבור מה-likelihood probability ל-posterior probability.



If x=12, we'll predict B, because P(x|B)>P(x|A)

הסתברות הפוסטריור וחוק בייס / Posterior Probability and Bayes rule

חוק בייס מניב עבורנו את ההסתברות הפוסטריור = מה ההסתברות לקלאס, בהינתן האינסטנס.

כדי לחשב את הפוסטריור נצטרך להשתמש בהסתברות הלייקליהוד והפריור (במונה) ובהסתברות המכנה (אווידנס).

$$P(A|x) = \frac{P(x|A)P(A)}{P(x)}$$

מסווג שיסווג A אם מתקיים P(A | x) > P (B | x), הוא מסווג שמביא למקסימום את הסתברות הפוסטריור – MAP Classifier. הקלסיפיקציה תלויה בפריור והלייקליהוד, האווידנס לא יעניין אותנו מפני שהוא במכנה והוא שווה בכל החישובים שלנו.

- We will classify A if

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

$$P(x|A)P(A) > P(x|B)P(B)$$

- Note that P(x) is removed from both denominators simply because it is the same

שגיאה מינימלית במסווג מסוג MAP:

But, we classify B only if $P(B|x) > P(A|x)$, and therefore the probability of the error is minimal

$$P(error|x) = \min[P(A|x), P(B|x)]$$

-יש פה קאץ', זה לא שאנחנו ממזערים את הטעות הגלובלית העולמית של הפרדיקציה, אלא בהתייחס להסתברויות שלמדנו מהדאטה הטעות היא מינימלית. כלומר אם ההסתברויות של הדאטה אכן מייצגות את העולם האמיתי, אכן השגיאה היא מינימלית ביחס לעולם האמיתי.

הגדרת פונקציית ה-loss: אם בחרתי את Ai בהינתן שהקלאס הנכון היה Aj ניתן דוגמה של 0/1 loss

0-1 loss (the simplest one):

$$\lambda_{ij} = \lambda(Choose\ A_i|A_j) = \begin{cases} 1, & if\ i \neq j \\ 0, & if\ i = j \end{cases}$$

בהינתן פונקציית ה-loss נגדיר את הסיכון ה-risk: זהו בעצם הסיכון לבחור קלאס מסוים, למשל Ai בהינתן האינסטנס, זהו ההפסד כפול ההסתברות של הקלאסים. אנחנו עושים סכום על כמות הקלאסים k, של הלוס על כל אחד של הפוסטריורים של הקלאסים.

<u>דוגמה לסיכון במקרה שפונקציית הלוס הינה loss 0/1 :</u>

$$R(Choose\ A_i|x) = \sum_{j=1}^{k} \lambda_{ij} P(A_j|x) = \sum_{j \neq i} P(A_j|x) = 1 - P(A_i|x)$$

נוכל להגדיר פונקציות לוס הרבה יותר מורכבות.

מסווג שרוצה למזער את הסיכון יבחר Ai כך ש : $P(A_i|x) > P(A_j|x)\ \forall j \neq i$

הסתברות הפוסטריור המלאה (לפני שנשמיט את P(x)) הינה על פי בייס : $g_i(x) = P(A_i|x) = \frac{P(x|A_i)P(A_i)}{\sum_{j=1}^{k} P(x|A_j)\,P(A_j)} = P(x)$

ולאחר שנשמיט את P(x) (המכנה) יתקיים : $g_i(x) = P(x|A_i)P(A_i)$

<u>בייס עבור קלאסים רבים</u>

כדי להפוך את תהליך הסיווג ליותר יעיל נוכל להשתמש ב-()ln: $g_i(x) = \ln\big(P(x|A_i)P(A_i)\big) = \ln\big(P(x|A_i)\big) + \ln\big(P(A_i)\big)$

שימוש ב-ln עוזר להפחית את ההכפלות במספרים נמוכים (בין 0-1) (חוקי לוגריתמים, הופכת מכפלות לסכומים) וכן עוזר להתמודד יותר טוב עם פונקציית הצפיפות הנורמלית ((e^f(x)). אנחנו יכולים להשתמש ב-ln מפני שהיא פונקציה מונוטונית עולה.

<u>**היפותזת מקסימום לייקליהוד ML-Hypothesis**</u>

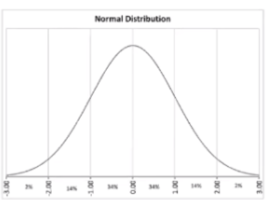
- נרצה לבחור את ההיפותזה שיש לה את ההיפותזה הכי סבירה (בעלת ההסתברות המקסימלית) בהינתן הדאטה , ובכתיב הסתברות נחפש אחר ההסתברות הפוסטריורית הבאה המקסימלית : $P(h|D) = P(D|h)P(h)$
- נוכל להניח שלכל ההיפותזות במרחב ההיפותזות יש את אותו הפריור P(h), ולכן נוכל למצוא את ההיפותזה הכי סבירה h על פי ה-maximum likelihood :

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}}\, p(D|h)$$

לכן למעשה ההיפותזה שממקסמת את הלייקליהוד היא ההיפותזה בעלת ההסתברות המקסימלית בהינתן הדאטה.

- נרצה למצוא את h-ML. נוכל להניח שכל האינסטנסים הם בלתי תלויים זה בזה ועל כן מתקיים :

$$P(D|h) = \prod_i P(y_i|h)$$



- **הנחה:** אם הטעות בהיפותזה מתפלגת באופן נורמלי עם ממוצע / תוחלת 0 $e_i \sim N(o, \sigma)$, אזי נוכל לומר שההסתברות ש- h(xi) = yi (שההיפותזה תדע לחזות בדיוק את ה-target value עבור אינסטנס מסויים xi) היא אותה ההסתברות ש- ei = 0 (שאין טעות עבור אינסטנס i) לפי ההסתברות הנורמלית של ei.

- **כעת נקבל:** (משמאל)
הנחנו אי תלות ולכן השורה הראשונה מתקיימת.
הנחנו שהטעות מתפלגת נורמלית ולכן ההסתברות לקבל את ה-target value בהינתן ההיפותזה, שווה להסתברות של הטעות ולכן נציב את נוסחת ההסתברות של הטעות המתפלגת נורמלית עבור כל אינסטנס (שורה שניה).
בשורה השלישית אנחנו מציבים את המרחק שלנו מהטעות, כי ההיפותזה נתונה.
לאחר כל זאת נוכל להפעיל ln. שיהפוך את הפאי, המכפלה, לסכום. ונשמיט ערכים קבועים כי הם לא משנים את המקסימום.

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}}\, p(D|h) = \underset{h \in H}{\operatorname{argmax}} \prod_i P(y_i|h)$$

$$= \underset{h \in H}{\operatorname{argmax}} \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{e_i - 0}{\sigma}\right)^2}$$

$$= \underset{h \in H}{\operatorname{argmax}} \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{h(x_i) - y_i}{\sigma}\right)^2}$$

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{h(x_i) - y_i}{\sigma}\right)^2}$$

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \ln\left(\prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{h(x_i) - y_i}{\sigma}\right)^2}\right)$$

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \sum_i \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2}\left(\frac{h(x_i) - y_i}{\sigma}\right)^2$$

$$= \underset{h \in H}{\operatorname{argmax}} \sum_i -\frac{1}{2}\left(\frac{h(x_i) - y_i}{\sigma}\right)^2$$

$$= \underset{h \in H}{\operatorname{argmax}} \sum_i -(h(x_i) - y_i)^2$$

$$= \underset{h \in H}{\operatorname{argmin}} \sum_i (h(x_i) - y_i)^2$$

- **לבסוף נקבל כי ההיפותזה הכי סבירה (maximum likelihood hypothesis) היא זו שממזערת את ה-MSE = שורה אחרונה.**

סיכום עד כה:

- Prior classifier: $P(A) > P(B)$
- ML classifier: $P(x|A) > P(x|B)$ – assuming $P(A) = P(B)$
- MAP classifier:
$$P(A|x) = P(x|A)P(A) > P(x|B)P(B) = P(B|x)$$

* Drooping $P(x)$ from the denominator

<u>איך מחשבים/משערכים את ההסתברויות? הלייקליהוד, הפוסטריור</u>

- **Parametric estimation** – אם ידוע לנו שאנחנו יכולים לנחש את סוג ההתפלגות נוכל להעריך את הפרמטרים של ההתפלגות. למשל אם נוכל לנחש שמשתנה מקרי מסויים מתפלג נורמלית נוכל לשערך עבורו את ה-miu וה-sigma, או אם הוא מתפלג פואסונית נוכל לשערך עבורו את הלמדא.
- **Non parametric estimation** – אם אנחנו לא יכולים להניח אף סוג של התפלגות על הדאטה שלנו נשתמש בהיסטוגרמה (=ספירה) או ב-Kernel Density Estimation (שזו למעשה היסטוגרמה חלקה = smooth histogram).

<u>שיערוך פרמטרי Parametric Estimation</u>

עבור כל קלאס נשערך את הפרמטרים של ההתפלגות על פי הדאטה אימון. אם אנחנו מדברים על ההתפלגות הנורמלית, עלינו לשערך את התוחלת ואת השונות של כל קלאס:

$$\mu = \frac{1}{m}\sum_{k=1}^{m} x_k$$

$$\sigma^2 = \frac{1}{m}\sum_{k=1}^{m}(x_k - \mu)^2$$

ואם נרצה לסווג לפי ההסתברות הגבוהה ביותר בהינתן ההתפלגות הנורמלית (הלייקליהוד):

$$P(x|A_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2}$$

הבעיה היא שכל זאת מתאים רק לאטריביוט/קלאס יחיד, **מה אם יש יותר אטריביוטים?**
אז במקרה כזה הסתברות הלייקליהוד תחושב לפי **התפלגות נורמלית רב-ממדית**. לשם כך נצטרך את וקטור התוחלות (כל ממד יהיה התוחלת של אטריביוט מסוים) ואת מטריצת השונות המשותפת the covariance matrix.
(במטריצה, השונויות שלא באלכסון הן ההסתברויות המשותפות)

Variance

$$\mathbf{S} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

Multivariate normal distribution:

$$P(\bar{x}|A_i) = \frac{1}{\sqrt{(2\pi)^d |S|^2}} e^{\left[-\frac{1}{2}(\bar{x}-\overline{\mu_i})^T S^{-1}(\bar{x}-\overline{\mu_i})\right]}$$

$|\mathbf{S}|$ - is the determinant of the covariance matrix

$\mathbf{S}^{-1}$ - is the inverse matrix of the covariance matrix

<u>שיערוך לא פרמטרי Non Parametric</u>

מכיוון שבשיערוך זה איננו יודעים להניח את ההתפלגות, עלינו למצוא דרך אחרת לשערך את הפריור P(Ai) ואת הלייקליהוד P(x | Ai). הסתברות הפריור P(Ai) ניתנת לשערוך מתדירות הקלאסים בדאטה אימון (מספר המופעים של הקלאס לחלק למספר הדגימות). לגבי הלייקליהוד, ראינו בהרצאה שניתן להשתמש בהיסטוגרמה וב-Interpolation, ובשבוע הבא נראה גישה נוספת.

<u>התמודדות עם multiple features</u>

כדי להעריך נכון את הלייקליהוד עבור דגימה נתונה עלינו להחזיק דאטה-סט עצום: אם בידינו d אטריביוטים בדידים, ו-k קלאסים, מספר האפשרויות של הלייקליהוד $P(x_1, x_2, \ldots, x_d|A_i)$ הבא הינו $k \cdot |V_1| \cdot |V_2| \cdots |V_d|$. עבור רק 2 אטריביוטים וקלאס בודד נוכל להגיע ל-2^d.
לכן אנחנו צריכים דרך / הנחה שתעזור לנו להתגבר על בעיה זו.
לכן נוכל להניח שהאריביוטים הם בלתי תלויים בהינתן קלאס – ברגע שנניח זאת נוכל להמיר את המשוואה למכפלת ההסתברויות של כל אטריביוט בנפרד. בדאטה מופיעים לנו כל הערכים של האטריביוטים האלה וכך הדבר ניתן לחישוב!

הנחה זו נקראית Naïve Bayes:

- אנחנו מניחים שכל האטריביוטים הם בלתי תלויים **בהינתן הקלאס**, ונקבל:

$$P(x_1, x_2, \dots, x_d|A_i) = \prod_{j=1}^{d} P(x_j|A_i)$$

- כעת נוכל למצוא את ה-MAP (Maximum A-Posterior Probability):

$$V_{NB} = \underset{i}{\operatorname{argmax}} P(A_i) \prod_{j=1}^{d} P(x_j|A_i)$$

הקלאס שיבחר הוא זה שממקסם את הביטוי הנתון.

- ברגע שאנו מניחים כך, אנחנו מורידים משמעותית את הגודל של הדאטה-סט שנצטרך הבדיד ל-

$$k \sum_{j=1}^{d} |V_j|$$

- ניתן להניח זאת גם במקרה הרציף, ניתן להשתמש בו גם בשיערוך הפרמטרי.