למידה חישובית - שיעור 5

נרצה להמשיך את הדיון שהתחלנו בתרגול להערכת ההסתברות שקיבלנו.

כיצד נעריך את ההסתברות?

מחדל פרמטרי

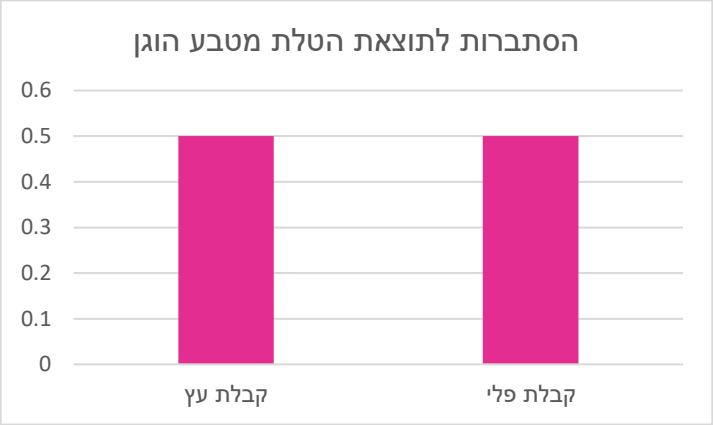
אם אנו יודעים את סוג ההתפלגות, או שאנו יכולים לנחש אותה, נוכל לעריך את הפרמטרים של ההתפלגות.

לכל Class נעריך את פרמטרי ההתפלגות לפי הtrain dataset, ובאמצעותם נחשב את הLikelihood של כל Class. לאחר שיש בידנו את ערך הLikelihood נבצע קלסיפיקציה לפי ערך ההסתברות הגדול ביותר שהתקבל מההתפלגות הנורמלית

משפט הגבול המרכזי

לכל הסתברות עם ערך μ ושונות σ^2, ההתפלגות ההמוצעת (והסכום) מתוכה היא התפלגות נורמלית עם ממוצע μ ושונות $\frac{\sigma^2}{n}$.

חשוב לזכור שלא כל התפלגות היא נורמלית, דוגמה מובהקת לכך היא הטלת מטבע הוגן. אם נבצע 10,000 הטלות הוגנות למטבע, נקבל את הגרף הבא:



נשים לב שההתפלגות שהתקבלה היא יוניפורמית, אך היא לא נראית התפלגות נורמלית. הסיבה היא כמות התוצאות האפשריות למבחן זה. מכיוון שיש לי רק 2 תוצאות אפשריות, ההתפלגות אינה נורמלית.
אם ההתפלגות אינה נורמלית- לא נשתמש בכלי זה. נלמד להתמודד עם זה בשיעור הבא.

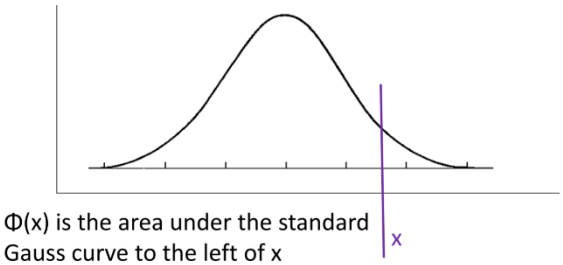
נוכל לשנות את שאלת המבחן להטלת מטבע הוגן 10 פעמים- ולבצע את המבחן הזה 10,000 פעמים, ובכך לקבל התפלגות נורמלית.

כלומר אנו מסיקים שההתפלגות הינה נורמלית (או לא) לפי המבחן עצמו ולא לפי כמות ההטלות (לפי ה-10 ולא לפי ה-10,000).

פונקציית ההתפלגות המצטברת (cumulative distribution function)

זוהי פונקציה המקבלת משתנה מקרי x ומחשבת את ההסתברות ש: $P(X \leq x)$, כאשר X הוא משתנה מקרי המקבל ערך הקטן או שווה ל-x. זהו מודל יוריסטי שמתיימר להיות קרוב לטבע. המודל הנורמלי עוזר לנו להכליל את ההתפלגות.

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{x^2}{2}\right) dx$$



<u>הערה</u>: נוכל להשתמש בהתפלגויות שאינן נורמליות, לדוגמה: פואסוניות. אנו לא נלמד על כך.

Normal Distribution Parameters

נרצה להתאים לכל attribute את ההתפלגות הנורמלית שלו. על מנת לעשות זאת, נחפש עבורו את μ ו-σ ונבצע קלסיפיקציה על ידי החישוב:

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

סטנדרטיזציה

נרצה **לנרמל** את הערכים שקיבלנו. נוריד מהמשתנה את התוחלת שלו, ונחלק בשונות.

$$U = \frac{(X-\mu)}{\sigma}$$

אינטואיציה לסטנדרטיזציה:

כשאנו מחסירים ממשתנה מקרי X, את התוחלת שלו, נקבל 0:

$$X - \mu \;\rightarrow\; E(X-\mu) \;=\; E\big(X - E(X)\big) \;=\; E(x) - E\big(E(x)\big) \;=\; E(x) - E(X) \;=\; 0$$

לאחר מכן נכפול $\frac{1}{\sigma}$ כלומר מחלקים ב$standard\ deviation$ (סטיית תקן)

$$Y = aX \;\rightarrow\; E(Y) = aE(X)$$
$$V(Y) = a^2 V(X)$$
$$std(Y) = a \cdot std(X) = \frac{1}{\sigma} \cdot \sigma(X) = 1$$

כלומר תהליך הסטנדרטיזציה עזר לנו להגיע לכך ש: $\mu = 0,\ \sigma = 1$.

אם X נורמלי, אז $U \sim (0,1)$ עם צפיפות נתונה של:

$$p(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$$

אם אנו משתמשים במודל נורמלי, ננסה למצוא את המודל הנורמלי הטוב ביותר עבור $P(x|A_i)$, לכל i.
לכל attribute שיש לי, אני אחשב:

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

<u>נחשב את השונות וסטיית התקן:</u>

$$\mu_i = \frac{1}{|A_i|}\sum_{x\in A_i} x$$

$$\sigma_i = \sqrt{\frac{1}{|A_i|}\sum_{x\in A_i}(x-\mu_i)^2}$$

את התהליך הזה נבצע עבור לכל instance כמספר הattributes שיש ברשותי. (בדוגמה בכיתה של האיריסים, נחשב את התוחלת והשונות 12 פעמים - 3 instances ו-*4* attributes, ואז נחשב את $P(x|A_i)$ עבור ערכים אלה)

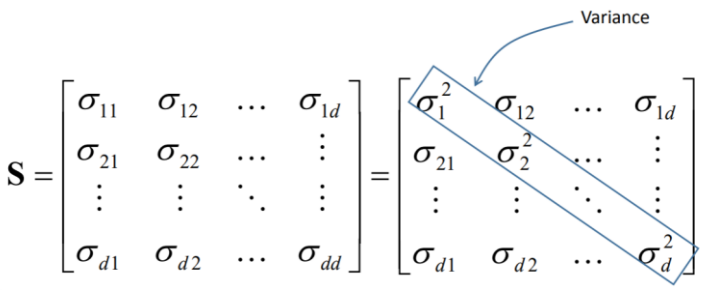
התפלגות נורמלית רב מימדית

נאמר ש$X\sim N(\mu,S)$, כאשר ל-X יש התפלגות גאוסיאנית רב מימדית:

$$p(x) = \frac{1}{(2\pi)^{d/2}|\mathbf{S}|^{1/2}}\exp[-\frac{1}{2}(x-\boldsymbol{\mu})^{t}\mathbf{S}^{-1}(x-\boldsymbol{\mu})]$$

כאשר:

- ***<u>μ</u>*** <u>-</u> וקטור הממוצעים.
- ***<u>S</u>*** <u>-</u> האלכסון הראשי והמשולש העליון של מטריצה הcovariance.



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

<u>טענה</u> (משתנים נורמליים רב מימדיים): $X_1, X_2, \dots, X_d$ משתנים מקריים ממימד d מתפלגים נורמלית <u>**אמ"מ**</u> כל צירוף לינארי $a_1X_1 + a_2X_2 + \cdots + a_dX_d$ יניב משתנה מקרי גאוסיני חד מיימדי.

Naïve Bayes Classifier

גישה נאיבית (ומכאן שמו) המבוססת על חוק בייס ועל ההנחה הנאיבית שאין תלות בין ה-attributes בהינתן ה-class, וכן כל ה-instances בלתי תלויים זה בזה. בעקבות ההנחה אנו צריכים לצפות רק ב-instances המכילים features מסוימים ולא צירופים של features עבור instances שונים.

כלומר, נקבל:

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

<u>נסווג instance עם תכונה $\vec{x}$ כ:</u>

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

הצעד הראשון בקלסיפיקציה באמצעות נאייב בייס הינו שערוך ההתפלגות המותנית לכל feature ולכל הclassים

הגישה הנאיבית פחות מדויקת, אך היא דורשת פחות data ללמוד ממנו. בנוסף השיטה הנאיבית מהירה יותר, בגלל שאנו מבצעים הנחת אי תלות שחוסכת לנו הרבה חישובים.

הגישה הנאיבית מספקת לנו שיערוך לא מספיק טוב וכן לעיתים (קרובות) ההנחה של אי התלות אינה נכונה, אך בפועל התוצאות שיתקבלו משימוש בNaive Bayes מספיק טובות בשביל להשוות בין הargmax (המקום שבו מתקבל המקסימום) של ההסתברות:

$$\arg\max_j \hat{P}(A_j) \prod_i \hat{P}(x_i \mid A_j) = \arg\max_j \hat{P}(A_j) \hat{P}(x_1, \dots, x_n \mid A_j)$$

וזה בעצם מה שאנו מחפשים.

Laplace Estimation

אם ניתקל במידע חסר שגורר מצב בו ההסתברות של המאורע היא 0 (כי יש לי binים ריקים עבור מקרים מסוימים, אבל זה לא אומר בהכרח שהם לא קיימים), נשתמש בשיערוך Laplace:

$$\hat{P}(x_i | A_j) = \frac{n_{ij} + 1}{n_j + |V_i|}$$

<u>**כאשר:**</u>

- $\boldsymbol{n_j}$ – מספר ה-instances עם class A_j
- $\boldsymbol{n_{ij}}$ – מספר ה-instances עם class A_j, עם ערך attribute x_i
- $|\boldsymbol{V_i}|$ – מספר הערכים האפשריים של ה-attribute x