למידה חישובית מפידנו תרגול 4 – Naïve Bayes and EM

Naïve Bayes

לכאורה האלגוריתם שלמדנו עובד כמו שצריך. אבל, למה שהוא יעבוד, אם השערוך של ה-likelihood הוא לא טוב - $\left(P(\bar{x}|A_i) \approx \prod_{j=1}^{d} P(x_j|A_i)\right)$

אז, ענינו על כך בהרצאה – הקלאס שמניב הסתברות מקסימלית בשני המקרים, הוא אותו הקלאס. כך שלמעשה אנחנו צריכים:

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

Discrete Naïve Bayes

כעת נראה דוגמה לחישוב ה-likelihood ע"י ספירה של הדאטה: $P(x_j|A_i) = \frac{n_{ij}}{n_i}$

כאשר:

- nij – הוא מספר הדגימות בדאטה אימון עם הקלאס Ai והערך xj באטריביוט הרלוונטי
- Ni – הוא מספר הדגימות בדאטה אימון עם הקלאס Ai

נשאלת השאלה מה הבעיה כאן? יכול להיות שב-training set הנתון, בהינתן קלאס לא בטוח שכל ערכי האטריביוטים הם מלאים ולכן תתכן הסתברות 0 likelihood (בגלל שאנחנו מכפילים) למרות שבחיים האמתיים הסתברות 0 לא מתרחשת. אסור לנו להניח שמה שלא ראינו (בדאטה אימון שלנו), לא מתקיים. לכן, כדי לפתור את בעיה זו ולתת איזון נשתמש ב- Laplace estimation.

תיקון לפלס – Laplace estimation

זהו תיקון מאוד פשוט אשר מוסיף אחד במונה ואת מספר הערכים של האטריביוט במכנה. כאשר:

$$P(x_j|A_i) = \frac{n_{ij} + 1}{n_i + |V_j|}$$

- nij – הוא מספר הדגימות ב-training data עם הקלאס Ai והערך xj באטריביוט הרלוונטי.
- ni – הוא מספר הדגימות ב-training data עם קלאס Ai.
- |Vj| - הוא מספר הערכים האפשריים של האטריביוט הרלוונטי.

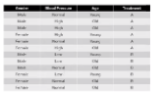
העיקרון שנשמר בתיקון זה, הוא שיישארו כאן הסתברויות בין 0-1 אשר נסכמות ל-1, לכן זו נוסחה valid-ית להסתברות.

דוגמה: נניח שנרצה לסווג בין שני טיפולים לחולים = קלאסים A ו-B. יש לנו מטופלים בעלי history data אשר מכיל 4 אטריביוטים: מין, לחץ דם, גיל והטיפול שהמטופל קיבל. נרצה לסווג מטופל חדש עם נאיב בייס. להלן הדאטה שלנו וכן החישובים הכוללים את תיקון לפלס.

Gender	Blood Pressure	Age	Treatment
Male	Normal	Young	A
Male	High	Old	A
Male	High	Old	A
Female	High	Young	A
Female	Normal	Young	A
Female	High	Old	A
Male	Low	Young	B
Male	Low	Old	B
Male	Normal	Old	B
Female	Low	Young	B
Female	Normal	Old	B
Female	Normal	Old	B

$P(A) = \frac{6}{12} = \frac{1}{2}$		
$P(male\|A) = \frac{3+1}{6+2} = \frac{4}{8} = \frac{1}{2}$	$P(female\|A) = \frac{3+1}{6+2} = \frac{4}{8} = \frac{1}{2}$	
$P(high\|A) = \frac{4+1}{6+3} = \frac{5}{9}$	$P(normal\|A) = \frac{2+1}{6+3} = \frac{3}{9}$	$P(low\|A) = \frac{0+1}{6+3} = \frac{1}{9}$
$P(young\|A) = \frac{3+1}{6+2} = \frac{4}{8} = \frac{1}{2}$	$P(old\|A) = \frac{3+1}{6+2} = \frac{4}{8} = \frac{1}{2}$	
$P(B) = \frac{6}{12} = \frac{1}{2}$		
$P(male\|B) = \frac{3+1}{6+2} = \frac{4}{8} = \frac{1}{2}$	$P(female\|B) = \frac{3+1}{6+2} = \frac{4}{8} = \frac{1}{2}$	
$P(high\|B) = \frac{0+1}{6+3} = \frac{1}{9}$	$P(normal\|B) = \frac{3+1}{6+3} = \frac{4}{9}$	$P(low\|B) = \frac{3+1}{6+3} = \frac{4}{9}$
$P(young\|B) = \frac{2+1}{6+2} = \frac{3}{8}$	$P(old\|B) = \frac{4+1}{6+2} = \frac{5}{8}$	

$$P(x_j|A_i) = \frac{n_{ij} + 1}{n_i + |V_j|}$$



לכן למשל עבור דגימה חדשה / מטופל חדש שהוא גבר, צעיר, ובעל לחץ דם גבוהה. נחשב (כנראה שבשורה הראשונה יש טעות בשוויון האחרון):

$male, young, high$

- $P(A|male, young, high) = P(A) \cdot P(male|A) \cdot P(young|A) \cdot P(high|A) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{5}{9} = \frac{5}{54}$
- $P(B|male, young, high) = P(B) \cdot P(male|B) \cdot P(young|B) \cdot P(high|B) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{3}{8} \cdot \frac{1}{9} = \frac{3}{288}$

After normalization:

- $P(A|male, young, high) = \frac{\frac{5}{54}}{\frac{5}{54} + \frac{3}{288}} = 0.9$ $\quad P(B|male, young, high) = \frac{\frac{3}{288}}{\frac{5}{54} + \frac{3}{288}} = 0.1$

ונקבל כי ההסתברות שהאדם יגיב יותר טוב לטיפול A היא גבוהה יותר ועל כן נסווג אותו לטיפול A.

אלגוריתם EM

אלגוריתם איטראטיבי שבנוי משני שלבים Expectation ו-Maximization. נסמן ב-D את קבוצת נקודות הדאטה שלנו (observed data), טטה יהיה וקטור הפרמטרים שלנו שנרצה שיניב עבורנו ML = maximum likelihood, כלומר אותו נחפש.

משתמשים ב-EM כאשר אנחנו רוצים לחשב את P(x | theta), אבל חישוב כזה באופן ישיר הוא לא פשוט. חישוב P(x,z | theta) הינו פשוט יותר, כאשר z הינו איזשהו דאטה חבוי hidden data. אנחנו מניחים שהדאטה החבוי נקבע ע"י משתנה מקרי כלשהו Z, שהוא חלק מהמודל.

הערה: המודל, תחת הוקטור טטה, שולט גם ב-X וגם ב-Z אבל ב-D אנחנו נראה רק את הערכים של X.

(כאן ניתנה הדוגמה שניתנה גם בהרצאה מספר 5 – עם ההטלה של 2 המטבעות)

<u>הערה לגבי דוגמה זו שעלתה בתרגול:</u> עדכון ה-P-ים מכפיל את הסכום ב-responsibilities, כלומר בנוסחה זו:

$$p_A = \frac{1}{(New\ w_A)N}\sum_{i=1}^{N} r(x_i, A)v(i)$$

מה שנמצא מחוץ לסיגמא (1 לחלק לWa החדש כפול N), הוא למעשה חילוק בסכום ה-responsibilities של A וניתן לראות זאת בנוסחה הבאה:

$$New\ w_A = \frac{1}{N}\sum_{i=1}^{N} r(x_i, A)$$

שבה אם מכפילים ב-N גדולה את שני האגפים נקבל מצד שמאל את המכנה שמחשב את Pa ומצד ימים את סכום ה-responsibilities של A, כאמור.

EM for GMMs

1. **שלב 1: Expectation (E-step)**

נשערך את ה"responsibilities" של כל נקודת דאטה לכל גאוסיאן באמצעות הפרמטרים הנוכחיים.

2. **שלב 2: Maximization (M-step)**

נשערך מחדש את הפרמטרים (w-ים, מיו-ים, סיגמות) בעזרת ה-"responsibilities" הקיימים. כלומר – כל נקודת דאטה, x, תורמת לכל מרכיב גאוסיאן, Gi, ביחס לאחריות שהיא קיבלה: r(x , Gi)

כאשר הנוסחאות עבור אלגוריתם זה הינן:

- Responsibilities:

$$r(x,k) = \frac{w_k N(x|\mu_k, \sigma_k)}{\sum_{j=1}^{K} w_j N(x|\mu_j, \sigma_j)}$$

- Weights:

$$New\ w_j = \frac{1}{N}\sum_{i=1}^{N} r(x_i, j)$$

- Mean:

$$New\ \mu_j = \frac{1}{(New\ w_j)N}\sum_{i=1}^{N} r(x_i, j)\, x_i$$

- Variance:

$$(New\ \sigma_j)^2 = \frac{1}{(New\ w_j)N}\sum_{i=1}^{N} r(x_i, j)\,(x_i - New\ \mu_j)^2$$

(ניתנה דוגמת ההרצה של 4 הגאוסיאנים שראינו בהרצאה)

נאמרה הערה לגבי הגרף שמתאר את <u>המיקסום של הלוג-לייקליהוד</u> לאורך האיטרציות, יש איזושהי התכנסות שאינה גלובלית בהרצות מוקדמות -10 40 ואז נעשה עוד שיפור בין האיטרציות 40-60 ובסביבות האיטרציה ה-80~ כבר נגיע להתכנסות.