פרק 4: מבוא לסיווג – למידה מדוגמאות

מקור: DHS(2001):2.1-2.7.

4.1 בעיית הסיווג – למידה מדוגמאות

בעיית הסיווג הבסיסית בהקשר של למידה ממוחשבת הינה כלהלן:

- נתונות n דוגמאות מסווגות (או מתויגות, labeled), דהיינו $\{x_k, y_k\}_{k=1}^n$, כאשר (אידיאלית) y_k הוא הסיווג הנכון של תבנית הקלט x_k.
- על סמך דוגמאות אלו, נדרש לתכנן מסווג ($f: X \to \Omega$), אשר יסווג כל קלט חדש x_{new} למחלקה המתאימה עם "שגיאה קטנה ככל האפשר".

הערות:

- סדרת הדוגמאות המתויגות $\{x_k, y_k\}_{k=1}^n$ נקראת גם סדרת הלימוד.
- אופן הלמידה הנדון הוא כמובן למידה אינדוקטיבית: הכללה מהפרט (סדרת הלימוד הנתונה) אל הכלל (קלט שאינו כלול בסדרת הלימוד).
- בעייה יסודית אליה נידרש הינה: כיצד נגדיר מתמטית את הדרישה ל"שגיאה קטנה ככל האפשר", כאשר כל הנתון הוא סידרת הלימוד.

הסימונים בהם נשתמש דומים לאלו שהגדרנו בהקשר לסיווג בייסיאני. בפרט:

- $X = \{x\}$ הוא מרחב הקלט. באופן טיפוסי $x = (x_1, \ldots x_d) \in \mathbb{R}^d$, כאשר $d > 1$. במקרה זה x נקרא גם "וקטור המאפיינים".
- $\Omega = \{1, 2, \ldots, C\}$ - אוסף המחלקות שונות, אליהן הקלט עשוי להשתייך.
- המסווג, או פונקציית הסיווג, הינו העתקה $f: X \to \Omega$.

הערה לגבי אינדקסים: עבור קלט וקטורי $x = (x_1, \ldots x_d) \in \mathbb{R}^d$, הרכיב ה-$i$ של וקטור זה יסומן x_i. סימון דומה משמש כדי לציין את איברי סדרת הלימוד $\{x_k\}_{k=1}^n$, כאשר $x_k \in \mathbb{R}^d$. על מנת למנוע בלבול נקפיד על שימוש באינדקס i או k, בהתאמה. במידת הצורך נשתמש גם בסימון החליפי $x_i \equiv x^i$ ו- $x_k \equiv x(k)$.

גישה לומדת לעומת גישה אנליטית

ניתן להבחין בין שתי הגישות הבאות לתכנון המסווג:

א. **פתרון אנליטי**: המסווג $f(x)$ מתוכנן מתוך ניתוח הבעייה וידע מוקדם שקיים לגביה.

ב. **הגישה הלומדת**: תכנון המסווג מתבסס על דוגמאות מסווגות (סדרת לימוד).

למרות ההבדל בהדגש, גישות אילו אינן בהכרח סותרות ואף עשויות להיות משלימות. הפתרון האנליטי עשוי לספק בסיס לתכנון המסווג הלומד, ואילו סט הדוגמאות עשוי לספק חלק מה״ידע המוקדם״ הדרוש לפתרון האנליטי. בסעיף הבא, למשל, נראה כיצד ניתן לשלב את הפתרון הבייסיאני האופטימלי (שהוא פתרון אנליטי) עם למידה מדוגמאות.

4.2 סיווג בייסיאני אמפירי

כזכור, בבעיית הסיווג הבייסיאני אנו מניחים ידיעה מוקדמת של הפילוגים $P(\omega)$ ו- $p(x|\omega)$.

המסווג האופטימאלי (במובן הסתברות שגיאה מינימאלית) הוא

$$f_{MAP}(x) = \arg\max_{\omega\in\Omega}\{p(x|\omega)P(\omega)\}.$$

מסווג זה אינו מסתמך על למידה מדוגמאות. הבעיה היא כמובן בהנחה לגבי ידיעה מוקדמת של הפילוגים הדרושים.

ניתן בקלות להסתמך על המסווג הבייסיאני כבסיס לסכמה לומדת, ע״י שימוש בסדרת הלימוד להערכת ההסתברויות הנדרשות. הסכמה המוצעת הינה כלהלן:

1. הערך את הפילוגים הנדרשים מתוך סדרת הלימוד $\{x_k, y_k\}_{k=1}^n$.
2. חשב את המסווג הבייסיאני האופטימלי בהתבסס על הפילוגים שהתקבלו.

- הערכת $P(\omega)$: הערכה זו ניתנת לביצוע בקלות ע״י חישוב התדירות היחסית של הופעת המחלקה ω בסדרת הלימוד:

$$\hat{P}(\omega) = \frac{n(\omega)}{n} \equiv \frac{1}{n}\sum_{k=1}^{n} I\{y_k = \omega\}$$

כמובן קיימת פה הנחה שסדרת הלימוד אכן מייצגת את השכיחות היחסית של הופעת ω באוכלוסיה כולה, אולם זוהי הנחה הכרחית באין מידע נוסף על הבעיה. בבעיות רבות ניתן לקבל הערכה של $P(\omega)$ מתוך מידע מוקדם על הבעיה, ואין צורך להסתמך על סדרת הלימוד בלבד.

• הערכת $p(x\,|\,\omega)$: הבעיה פה מסובכת בהרבה, כיוון שנדרשת הערכת של מספר <u>פונקציות</u> פילוג (אחת לכל ω) במשתנה x . הנושא של הערכת פילוגי הסתברות נדון בפרק הקודם. נסתפק פה בהדגמה קצרה, תחת ההנחה כי הפילוג $p(x\,|\,\omega)$ הינו גאוסי.

<u>דוגמא</u>: הערכת $p(x\,|\,\omega)$ במקרה הגאוסי. לכל $\omega\in\Omega$, אנו מניחים כי הפילוג $p(x\,|\,\omega)$ ניתן לקירוב באמצעות פילוג גאוסי, דהיינו $p(x\,|\,\omega)\sim N(\mu_\omega,\Sigma_\omega)$. הנחה זו עשויה להסתמך על מידע מוקדם, וניתן לבחון את תקפותה בעזרת סדרת הלימוד. את הממוצע μ_ω והווריאנס Σ_ω ניתן להעריך בצורה הרגילה מתוך סידרת הדוגמאות. נסמן ב- $\{z_k\}_{k=1}^{n(\omega)}$ את תת-הסידרה של סדרת הדוגמאות $\{x_k\}_{k=1}^{n}$ שעבורן $y_k=\omega$. אזי

$$\hat{\mu}_\omega=\frac{1}{n(\omega)}\sum_{k=1}^{n(\omega)} z_k$$

$$\hat{\Sigma}_\omega=\frac{1}{n(\omega)-1}\sum_{k=1}^{n(\omega)}(z_k-\hat{\mu}_\omega)(z_k-\hat{\mu}_\omega)^T$$

למרות הפשטות הרעיונית של גישת הסיווג הבייסיאני האמפירי, הבעייה של הערכת פילוג רב-מימדי הינה מסובכת באופן כללי ועשויה להוביל למסווגים מסובכים ללא צורך. לפיכך הגישה איננה בשימוש נרחב.

<u>מסווג בייס נאיבי</u>: כאמור, בעייה מרכזית בגישה של המסווג הבייסיאני האמפירי הינה בהערכת הפילוג $p(x\,|\,\omega)$, כאשר הקלט x הינו רב מימדי. דרך אפקטיבית לפשט בעיה זו הינה להניח (לצורך השערוך) אי-תלות בין רכיבי x . בפרט, נניח כי $x=(x^1,x^2,\ldots,x^d)$. המסווג הבייסיאני הנאיבי מתבסס על הקרוב הבא לפילוג הדרוש:

$$p(x\,|\,\omega)\approx p(x^1\,|\,\omega)\,p(x^2\,|\,\omega)\cdots p(x^d\,|\,\omega)$$

לפיכך, לצורך הגדרת המסווג הבייסיאני נדרשת הערכת הפילוגים <u>החד-מימדיים</u> $p(x^i\,|\,\omega)$ עבור $i=1,\ldots,d$. הערכה זו קלה בהרבה מהערכת הפילוג הרב-מימדי.

למרות שהנחת אי-התלות אינה מבוססת, המסווג המתקבל בהסתמך על הנחה זו הוא בעל ביצועים סבירים במקרים רבים.

4.3 הגישה הפרמטרית לתכן המסווג

גישה אלטרנטיבית (ונפוצה) לתכן מסווג מדוגמאות הינה הגישה הפרמטרית. תכן המסווג מבוצע לפי השלבים הבאים:

א. בחירת מסווג פרמטרי: בחירת מבנה המסווג, עד כדי סט פרמטרים θ אותו יש לקבוע בהמשך.

ב. כיוונון (לימוד) הפרמטרים θ על פי סדרת הלימוד.

בחירת המסווג הפרמטרי כוללת, ביתר פירוט, את בחירת הצורה הפונקציונלית של המסווג, מספר הפרמטרים ("סדר המודל"), ותחום ההשתנות שלהם. שלב זה הינו חשוב ביותר להצלחת התכן, ויש להעזר בו בכל הידע המוקדם הקיים לגבי הבעייה הספציפית, בניסיון מצטבר לגבי בעיות בעלות אופי דומה, ובגישת "ניסוי וטעיה".

כיוונון הפרמטרים מתבצע, עקרונית, כך שיתקבל סיווג מיטבי של סדרת הלימוד (דהיינו, מספר שגיאות מינימלי ביחס לתוויות הנתונות).

להמחשה, נתאר את הגישה עבור סיווג בעזרת פונקציות אבחנה (דיסקרימיננציה). מסווג מסוג זה מוגדר כך: לכל מחלקה $\omega_j \in \Omega$ נגדיר פונקצית אבחנה $g_j : X \to \mathbb{R}$ על גבי מרחב הפלט. סיווג הפלט x נקבע לפי פונקציית האבחנה בעלת הערך המירבי, כלומר

$$f(x) = \arg\max_j \{g_j(x)\}$$

נשים לב כי המשערך הבייסיאני האופטימלי מוגדר באופן דומה (עבור פונקציות אבחנה מתאימות).

בגרסה הפרמטרית של מסווג זה, כל פונקצית אבחנה g_j תלויה בסט פרמטרים θ_j. נציין זאת על ידי הסימון $g_j(x) = g_j(x, \theta_j)$.

פונקצית האבחנה הפשוטה ביותר הינה הלינארית: $g_j(x) = b_j + w_j^T x$. סט הפרמטרים במקרה זה הינו $\theta_j = (b_j, w_j)$. כפי שראינו, משפחה זו של מסווגים כוללת את המסווג הבייסיני האופטימלי במקרה של פילוגים גאוסיים בעלי קוואריאנס זהה.

פונקצית אבחנה מורכבת מעט יותר הינה הריבועית: $g_j(x) = b_j + w_j^T x + x^T M_j x$. סט הפרמטרים עתה יכלול גם את איברי המטריצה M_j. כפי שראינו, משפחה זו של מסווגים כוללת את המסווג הבייסיני האופטימאלי במקרה של פילוגים גאוסיים בעלי קוואריאנס כלשהו.

בהמשך הקורס נתייחס למשפחות כלליות יותר של פונקציות הבחנה, ושל מסווגים פרמטריים בכלל, ונתאר גישות אפשריות לכיוונון הפרמטרים עבורם.

2.5 הגישה הלא-פרמטרית

מסווגים במשפחה זו מוגדרים ישירות על המידע (כלומר סדרת הלימוד), ללא שלב של כיוונון פרמטרים. מסווג נפוץ במשפחה זו הוא "מסווג K השכנים הקרובים" (K Nearest Neighbors, K-NN) אותו נתאר כאן בקצרה.

א. מסווג השכן הקרוב: תהי $\{x_k,\omega_k\}_{k=1}^n$ סדרת הלימוד, אותה אנו שומרים בזיכרון. בהינתן קלט חדש x, נמצא את תבנית הקלט x_k הקרובה ביותר ל- x, ונסווג את x בהתאם לתווית של x_k:

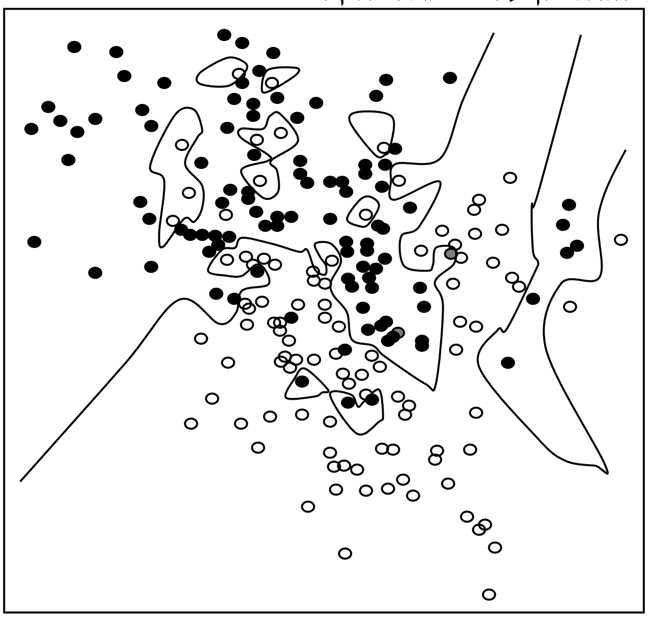
$$f_{NN}(x)=\omega_{k(x)}$$

כאשר

$$k(x)=\underset{k=1,\ldots,n}{\arg\min}\, d(x,x_k)$$

ואילו $d(x,x_k)$ הוא המרחק בין x ל- x_k.

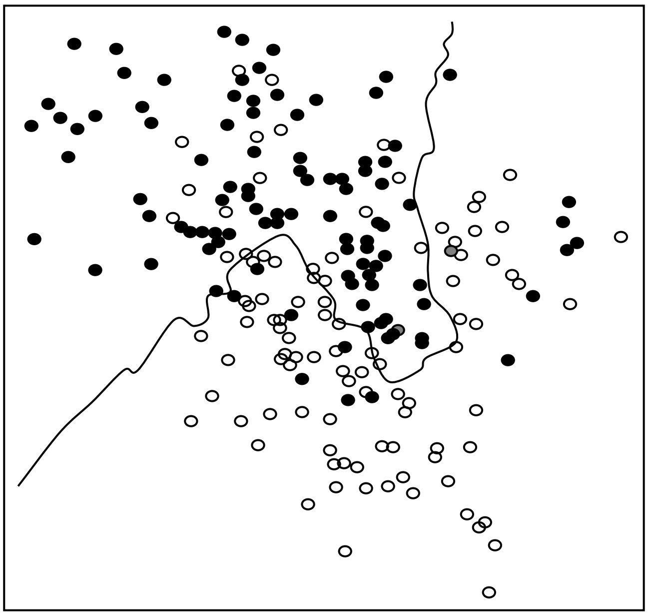
פעולת מסווג "השכן הקרוב" מודגמת בציור הבא. מובן כי מאחרי הגדרת מסווג זה עומדת הנחת רציפות כלשהי של הסיווג הנכון על פני מרחב הקלט X.



סיווג לשתי קטגוריות באמצעות מסווג K-NN עם K=15
לפי Hastie et. al. (2001), ציור 2.3

למסווג "השכן הקרוב" שני חסרונות פוטנציאליים. האחד הוא כי חלוקת מרחב הקלט X למחלוקת עשויה להיות בלתי רגולרית (לא חלקה). השני הוא רגישות גדולה לטעויות בסדרת הלימוד. לשיפור נקודות אלה ניתן להכליל את המסווג באופן הבא:

ב. מסווג "K השכנים הקרובים" (K-NN): מסווג זה מוצא את K השכנים הקרובים ל- x בסדרת הלימוד, ובוחר את התווית של x לפי תווית הרוב בין שכנים אלה. (במקרה של "תיקו" בין מספר תוויות בוחרים ביניהן אקראית, או לפי כלל בחירה אחר קבוע מראש).



סיווג לשתי קטגוריות באמצעות מסווג K-NN עם K=15
לפי Hastie et. al. (2001), ציור 2.2

הערות:

1. אופן הלמידה המיוצג על ידי מסווגים אלה מכונה Lazy Learning – הוא מאופיין שפעולות החישוב המקדים הנדרשות הינן מינימליות, מרבית פעולת החישוב של המסווג מתבצעת כאשר מגיע קלט חדש.

2. מסווגים אלה נדרשים לשמור בזיכרון את סדרת הלימוד $\{x_k,\omega_k\}_{k=1}^{n}$, ובכל פעם שמתבצע סיווג של קלט חדש יש למצוא את האיבר הקרוב ביותר (או k האיברים הקרובים) מתוך סדרה זו. כאשר מספר הדוגמאות n גדול נדרש זיכרון גדול בהתאם ועומס חישובי ניכר. קיימים מספר אלגוריתמים שמטרתם "לדלל" את סדרת הלימוד המקורית על ידי מחיקת דוגמאות שהשפעתם על המסווג קטנה.

3. מסווגים אלה הינם פשוטים יחסית לתכנון ומימוש, אולם זמן החישוב של המסווג עלול להיות גדול כאשר מספר הדגימות גדול, והביצועים תת-אופטימליים כאשר מספר הדגימות אינו גדול מספיק.

4. הגדרת מסווג זה משתמשת במידת מרחק d על מרחב הקלט X. בדוגמאות לעיל השתמשנו במרחק האאוקלידי (ה"טבעי"), אולם בשימושים מסוימים הגדרת מרחק נכונה עשויה להיות מסובכת בהרבה. לדוגמא: מה המרחק בין שתי סדרות באורך N, כאשר ידוע כי 3 איברים מכל סדרה חסרים (במיקום לא ידוע)?

<u>אנליזה:</u>

נראה כעת כי מסווג זה בעל ביצועים לא רחוקים מאילו של המסווג הבייסיאני האופטימלי, המסווג ע"פ הכלל $c_m = \arg\max_i P(\omega_i \mid x)$

נסמן את המאורע שמסווג ה-NN עבור מדגם בגודל n שגה ע"י e אזי השגיאה הכוללת שלו נתונה ע"י

$$P_n(e) = \int P_n(e \mid x)\, p(x)dx$$

נסמן את הגבול (אם קיים) של ערך זה ע"י

$$P = \lim_{lin\to\infty} P_n(e)$$

נציין שעבור המסווג הבייסיאני האופטימלי מתקיים

$$P^*(e \mid x) = 1 - P(c_m \mid x)$$

וכן $P^* = \int P^*(e \mid x) P(x)dx$

משפט: $P^* \le P \le 2P^*$

אנו מניחים כי פונקציית הפילוג של הדוגמאות $P(x)$ הינה רציפה ואינה שווה ל-0. כלומר, לכל נקודה x ולכל סביבה קטנה כרצוננו סביבה קיים סיכוי גדול ממש מ-0 שנקודה אקראית תיפול בתוך סביבה זו. ולכן בגבול של גודל מדגם שואף לאין-סוף, מתקיים כי הנקודה הקרובה ביותר ל- x, הינה x עצמה. מהיא אם כן, השגיאה של מסווג NN במקרה זה? או דוגמים פעמיים מתוך הפילוג $P(w \mid x)$ ונטעה כאשר שתי הדגימות שונות. פוטרמלית נסמן את דגימות אלו ע"י θ_1, θ_2 - כל אחד בתחום $1...C$. מכאן נקבל כי בגבול מתקיים,

$$\lim_{n\to\infty} P_n(e \mid x) = P(\theta_1 \ne \theta_2 \mid x) = 1 - \sum_{c=1}^{C} P(\theta_1 = c, \theta_2 = c \mid x)^2 = 1 - \sum_{c=1}^{C} P(c \mid x)^2$$

נסכם, קיבלנו כי בגבול מתקיים

$$P = \int \left(1 - \sum_{c=1}^{C} P(c \mid x)^2\right) P(x)dx$$

נראה כעת כי, $1 - \sum_{c=1}^{C} P(c \mid x)^2 \le 2P^*(e \mid x)$, ונסיים. אלגברה תסייע לנו לראות כי טענה זו שקולה לזהות הבאה,

$$-\sum_{c\ne c_m} P(c \mid x)^2 \le \left(1 - P(c_m \mid x)\right)^2$$

חסם הדוק יותר

ניתן לקבל חסם הדוק יותר באופן הבא:

נרצה למצוא חסם עליון לגודל $1-\sum_{c=1}^{C}P(c|x)^2$ או חסם תחתון לגודל $\sum_{c=1}^{C}P(c|x)^2$

כאשר ידוע כי כל האיברים בסכום הם אי שליליים וכן כי,

$$1=\sum_{c=1}^{C}P(c|x)=\sum_{c\neq c_m}^{C}P(c|x)+P(c_m|x)=\sum_{c\neq c_m}^{C}P(c|x)+1-P^*(e|x)$$

דהיינו

$$\sum_{c\neq c_m}^{C}P(c|x)=P^*(e|x)$$

מינימום של סכום-ריבועים תחת אלוץ לינארי אחיד הוא כאשר כל הערכים שווים, דהיינו

$$P(c|x)=\begin{cases}\dfrac{P^*(e|x)}{c-1} & c\neq c_m\\ P^*(e|x) & c=c_m\end{cases}$$

ולכן מתקבל,

$$1-\sum_{c=1}^{C}P(c|x)^2\leq 2P^*(e|x)-\frac{c}{c-1}\left(P^*(e|x)\right)^2$$

נעשה שימוש בקשר

$$\int\left(P^*\right)^2(e|x)P(x)dx\geq\left(\int P^*(e|x)P(x)dx\right)^2=\left(P^*\right)^2$$

ונקבל

$$P^*\leq P\leq P^*\left(2-\frac{C}{C-1}P^*\right)$$

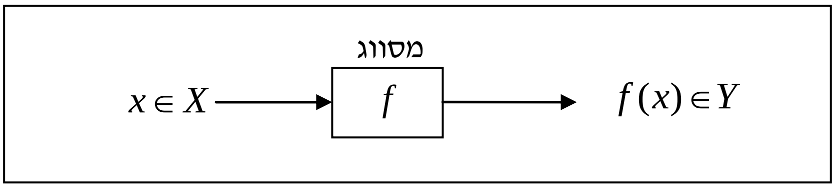
תכונות מסווגי NN:

יתרונות: מאוד פשוטים, יכולת למדל כל פילוג בגבול עם הרבה דוגמאות.

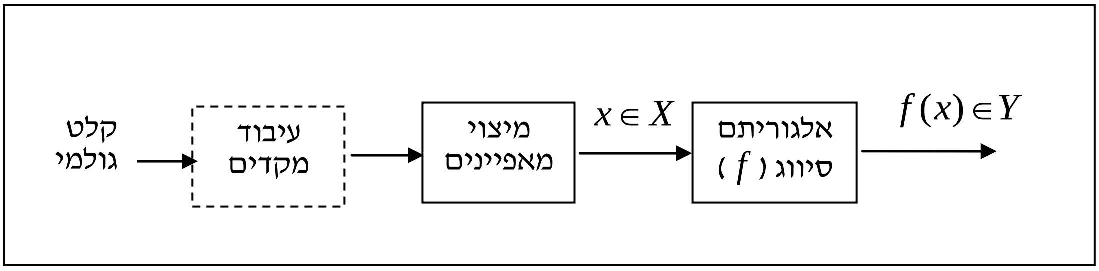
חסרונות: משאבי זכרון וחישוב, קללת המימד (במימד גבוה, כל הנקודות רחוקות זו מזו במידה דומה), רגיש מאוד ליצוגים רועשים (לדוגמא פיצ׳רים רועשים).

2.6 תהליך התכן

נתאר באופן סכמטי תהליך תכן אופייני של מסווג לומד.
נזכיר ראשית כי מסווג הוגדר בראשית הפרק באופן הבא :



בפועל, אלגוריתם הסיווג מופעל רק לעיתים רחוקות על הקלט הגולמי, והמסווג יכלול שלבים מקדימים של עיבוד הקלט הגולמי, כאשר העיקרי שבהם הינו <u>מיצוי מאפיינים</u> מתוך המידע :



תהליך התכן ניתן עתה לתיאור באופן הבא: