תרגול 3 : מבוא לבעיית הסיווג

מסווג בייס נאיבי

סימונים:

Ω – מרחב סופי של קטגוריות (מחלקות). $\omega_i \in \Omega,\ i = 1,\ldots,N$

X – מרחב הקלט (תבניות). $x \in X$.

$f : X \mapsto \Omega$ – מסווג כל $x \in X$ ל- $\omega \in \Omega$.

$\{x_k, y_k\}_{k=1}^m$ – m דוגמאות מתוייגות (סדרת הלימוד), כאשר $x_k \in X$, $y_k \in \Omega$.

תזכורת: המסווג הבייסיאני האופטימלי נתון ע"י:

$$f(x) = \arg\max_{i=1,\ldots,N} p\left(x \mid \omega_i\right) p\left(\omega_i\right)$$

<u>מסווג בייסיאני אמפירי:</u>

1. הערך את הפילוגים $p\left(x \mid \omega\right)$ ו- $p(\omega)$ מתוך סדרת הלימוד $\{x_k, y_k\}_{k=1}^m$ ע"י אחת מהשיטות שלמדנו.
2. חשב את המסווג הבייסיאני האופטימלי ע"י שימוש בפילוגים שהתקבלו ב-1.

<u>מסווג בייסיאני נאיבי:</u>

כאשר מרחב הקלט X הינו רב-מימדי (ממימד d), הערכת הפילוג $p\left(x \mid \omega\right)$ הינה מסובכת, וברוב המקרים לא מעשית. על מנת לפתור בעיה זו נהוג להניח (לצורך השערוך) אי-תלות בין רכיבי הקלט $x = (x^1, x^2, \ldots, x^d)^T$. כלומר מניחים (באופן "נאיבי") כי:

$$p(x \mid \omega) \approx \prod_{i=1}^{d} p(x^i \mid \omega)$$

ומשערכים את הפילוגים השוליים (החד-מימדיים) $\left\{p(x^i \mid \omega)\right\}_{i=1}^d$ על סמך סדרת הלימוד.

שאלה 1

נתבונן בבעיית סיווג בינארי לשתי מחלקות ω_1,ω_2 מתוך סדרת דוגמאות $\{\mathbf{x}_k, y_k\}_{k=1}^m$, כאשר הכניסה דו-מימדית: $\mathbf{x}_k = (x_k^1, x_k^2) \in \mathbb{R}^2$, והתווית הינה $y_k \in \{\omega_1,\omega_2\}$.

הוצע להשתמש במסווג בייס נאיבי, באופן הבא:

1. הערכת הפילוג אפריורי של מחלקה ω_j לפי השכיחות היחסית בסדרת הדוגמאות.
2. הערכת הפילוג השולי של x^i (הרכיב ה-i של $\mathbf{x}$, $i=1,2$) לפי משערך סבירות מירבית, בהנחה כי הפילוג הינו גאוסי.
3. שימוש בנוסחאות המסווג הבייסיאני אופטימלי, עם הגדלים המשוערכים הנ"ל.

שאלות:

1. רשמו באופן מלא את משוואות האלגוריתם המוצע.
2. הניחו כי התקבלו ארבע הדוגמאות הבאות:
$$(-1,-1),\omega_1;\ \ (0,0),\omega_1;\ \ (0,0),\omega_2;\ \ (1,1),\omega_2$$
רשמו במפורש את חוק ההחלטה הנגזר מכך. פשטו ככל האפשר והסבירו את התוצאה המתקבלת.
3. חזרו על הסעיף הקודם עבור הדוגמאות הבאות:
$$(-1,-1),\omega_1;\ \ (1,1),\omega_1;\ \ (-1,1),\omega_2;\ \ (1,-1),\omega_2$$
4. הניחו עתה כי הדוגמאות $\mathbf{x}$ בעלות סיווג ω_1 מתקבלות לפי פילוג אחיד על הריבוע $[0,1]^2$, ואילו הדוגמאות בעלות סיווג ω_2 מתקבלות לפי פילוג אחיד על הריבוע $[-1,1]^2$. השכיחות היחסית של שתי המחלקות שווה. רשמו את חוק ההחלטה הגבולי המתקבל כאשר $n \to \infty$. הראו כי עקום ההחלטה המתקבל הינו מעגל, וחשבו את מרכזו ורדיוסו.

למידה "עצלנית" (Lazy Learning)

למידה עצלנית כשמה כן היא: עם קבלת סט האימון לא מתבצעים חישובים כלשהם ולא נעשית הכללה למקרה הכללי, ורק כאשר נדרשת קבלת החלטה מבצעת המערכת את מספר הפעולות המינימלי הנדרש לשם כך. זאת לפי הפתגם הידוע: "מדוע לדחות למחר את מה שאפשר לדחות למחרתיים?".

<u>אזהרה:</u> למידה עצלנית אינה הלמידה המומלצת בקורס זה!!

שערוך לוקלי של פילוגי הסתברות

במקום לשערך את פילוג ההסתברות המלא של משתנה מקרי כלשהו, ניתן לשערך את הפילוג בנקודה המעניינת אותנו. אם נתונה הנקודה הרצויה, ניתן לקרב את פילוג ההסתברות בנקודה זו ע"י מציאת המספר היחסי של דגימות (מסט האימון) בסביבת הנקודה. כלומר:

$$P_X(x) \approx \frac{\left|\{x_i :\ x_i \in D,\ |x_i - x| \le \varepsilon\}\right|}{\varepsilon\left|\{x_i :\ x_i \in D\}\right|}$$

מהי "סביבה" של הנקודה? בחירת ε גדול מדי תתן שערוך מאוד גס לפילוג ההסתברות בנקודה. מצד שני, בחירת ε קטן מדי עלולה ליצור מצב בו הסביבה תכלול מעט מדי (אם בכלל) דגימות.

סיווג בעזרת אלגוריתם K-NN
(k Nearest Neighbours)

1. מצא את K השכנים הקרובים ביותר לנקודה החדשה.
2. מצא לאיזו קבוצה שייכים רוב השכנים. הנקודה החדשה שייכת לקבוצה זו.

2.1. במקרה של שוויון בשלב 2, השווה סכום מרחקים. הנקודה החדשה שייכת לקבוצה בעלת הסכום המינימלי.

2.1.1. במקרה של שוויון בשלב 2.1, בחר אקראית.

שאלה 2

נתונה בעיית סיווג לשתי מחלקות עם הסתברויות אפריוריות שוות ועם הפילוגים המותנים הבאים:

$$p(x|\omega_1)=\begin{cases}2x, & 0\le x\le 1\\ 0, & otherwise\end{cases} \qquad p(x|\omega_2)=\begin{cases}2-2x, & 0\le x\le 1\\ 0, & otherwise\end{cases}$$

1. מהו כלל ההחלטה הבייסיאני למקרה זה ומהי הסתברות שגיאת הסיווג?
2. מגרילים שתי נקודות, אחת מכל מחלקה. מהי ההסתברות לשגיאה בסיווג K-NN של נקודה שלישית, המוגרלת מהמחלקה ω_1, עבור $K=1$?
3. השוו את הסעיף הקודם למקרה בו שתי הנקודות בסט האימון הן בדיוק התוחלות של שתי המחלקות, בהתאמה.