מימוש של מסווג Gaussian Naïve Bayes

הקדמה

בחלק זה של הפרויקט תאמנו מסווג מסוג GNB, שהוא פונקציה $f:\mathcal{R}^{\mathrm{D}} \longmapsto \{1,2,\dots,C\}$ המשייכת לכל וקטור במרחב המאפיינים $\mathrm{x} \in \mathcal{R}^{\mathrm{D}}$ אחד מ-C ערכים $y \in \{1,2,\dots,C\}$. המידע $\mathcal{D} = \{(\mathrm{x}_n, y_n)\}_{n=1}^{N}$ שישמש אתכם לאימון המסווגים מורכב מאוסף דוגמאות, כשכל x_n הינו מטריצה ריבועית במימד 28x28 (סה"כ 784 פיקסלים, כלומר בעבודה זאת D=784) המייצגת תמונה (grayscale) של פריט לבוש, ו-y_n התיוג של כל דוגמא מייצג את סוג הפריט ע"פ הפירוט להלן (ראו איור 1)

```
Label 0: T-shirt/top
Label 1: Trouser
Label 2: Coat
Label 3: Sandal
Label 4: Ankle boot
```

ראינו כי מסווג מסוג GNB מתבסס על שערוך *maximum likelihood* של פרמטרי המודל הבא:

$$\forall \quad c = 1,2,\dots,C:$$

$$P(x|y=\mathrm{c},\mu_c,\Sigma_c) = \mathcal{N}(x|\mu_c,\Sigma_c) \quad \left(\text{where } \Sigma_c = \begin{pmatrix} \sigma_{c,1}^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_{c,\mathrm{D}}^2 \end{pmatrix} \text{ is diagonal}\right)$$

$$P(y=\mathrm{c}) = \pi_c$$

$$0 \le \pi_c \le 1 \qquad \left(\text{and} \quad \sum_{c=1}^{C} \pi_c = 1\right)$$

וכי בהינתן שיערוכי הפרמטרים של המודל $\Theta = \{\pi_{c,ML}, \mu_{c,ML}, \Sigma_{c,ML}\}_{c=1}^{C}$, ניתן להשתמש בחוק בייס באופן הבא

$$P(y=\mathrm{c}|x,\Theta) = \frac{P(x|y=c,\mu_{c,ML},\Sigma_{c,ML})P(y=c)}{\sum_{c'=1}^{C} P(x|y=c',\mu_{c',ML},\Sigma_{c',ML})P(y=c')}$$

$$= \frac{\mathcal{N}(x|\mu_{c,ML},\Sigma_{c,ML})\pi_{c,ML}}{\sum_{c'=1}^{C} \mathcal{N}(x|\mu_{c',ML},\Sigma_{c',ML})\pi_{c',ML}}$$

$$\propto \mathcal{N}(x|\mu_{c,ML},\Sigma_{c,ML})\pi_{c,ML} \qquad \forall \quad c = 1,2,\dots,C$$

ולסווג כל דוגמא x ע"פ הכלל

$$(1) \quad \hat{y}(x) = f(x) = \underset{c=1,2,\dots,C}{\operatorname{argmax}} P(y=\mathrm{c}|x,\Theta)$$

$$= \underset{c=1,2,\dots,C}{\operatorname{argmax}} \mathcal{N}(x|\mu_{c,ML}\Sigma_{c,ML})\pi_{c,ML}$$

איור 1: דוגמאות לסוגי פרטי הלבוש.

כמו כן הגדרנו את דיוק (accuracy) המסווג $f(\mathrm{x})$, עבור הדאטה $\mathcal{D}$:

$$P_c(f,\mathcal{D}) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}\big(y_n == f(x_n)\big)$$

כאשר $\mathbb{I}(A)$ היא פונקציית האינדיקטור

$$\mathbb{I}(A) = \begin{cases} 1 & \text{if } A \text{ is TRUE} \\ & \\ 0 & \text{otherwise} \end{cases}$$

<u>שימו לב</u>:

- בפתרונכם ניתן להשתמש בקוד שכתבתם במהלך הסמסטר בתרגילי הקידוד ובמטלות הבית.
- בנוסף למחברת שתגישו, יש לצרף מסמך מלווה עם התשובות לחלקים העיוניים במטלה ופירוט תוצאות ניסוייכם. בהירות ושלמות המסמך ישפיעו על הציון הסופי.

1. (1 נקודות) טענו את הדאטה שקיבלתם, והכינו אותו לאימון המסווג ע"י ייצוג כל תמונה כוקטור במרחב $\mathcal{R}^{784}$.

סדרת האימון מכילה 24,000 דוגמאות, כל אחת כוללת מטריצה במימד 28x28 (המאפיינים x) ותיוג (y) לכל אחת מהדוגמאות. אוסף כל התמונות נתון לכם כמטריצה תלת-ממדית במימד 28x28x24,000, והתיוגים של כל הדוגמאות כוקטור באורך 24,000.

א. טענו את הדאטה הנתון בקובץ TrainData.pkl, הכולל את מטריצת התמונות ואת וקטור התיוגים של סדרת המבחן.

ב. הציגו בתמונות נפרדות דוגמה אחת מכל אחד מהתיוגים הקיימים בדאטה, והוסיפו לתמונה את תיוגה ע"פ הדאטה.

ג. הגדירו מטריצה חדשה במימד 24,000x784, כאשר השורה ה-n-ית בה מכילה וקטור באורך 28x28=784 עם ערכי כל הפיקסלים של הדוגמא ה-n-ית x_n (המתקבל לאחר שכל השורות בה שורשרו זו לזו).

הערה - ניתן לבצע הנ"ל על-ידי np.reshape(...)

2. (34 נקודות) ממשו את אלגוריתם האימון של מסווג GNB עבור מידע רב-ממדי כפי שלמדנו בקורס

א. כתבו פונקציה המקבלת את סדרת האימון, ומחשבת את שערוך maximum likelihood של כל אחד מפרמטרי המודל $\Theta = \{\pi_{c,ML}, \mu_{c,ML}, \Sigma_{c,ML}\}_{c=1}^{C}$.

הערה I: ערכי הפיקסלים בתמונות שקיבלתם מוגדרים כמשתנה מסוג int8. בכדי להימנע מבעיות נומריות בשערוכי הפרמטרים מומלץ ראשית לשמור אותם בפורמט אחר דוגמת float64.

הערה II: בתשובתכם יש לבצע חישוב מלא של השערוכים (אין להשתמש בפונקציות כגון np.cov)

ב. כתבו פונקציה נוספת המקבלת את כל פרמטרי המודל ששערכתם ודוגמאות לא מתויגות, ומחזירה את סיווג כל אחת מהתמונות ע"פ

$$\hat{y}(x) = \underset{c=1,2,\dots,C}{\operatorname{argmax}}\, P(y = c|x, \Theta)$$

ג. יצגו כל אחד מהוקטורים הממוצעים ששערכתם כמטריצה ריבועית (כלומר יצגו כל וקטור באורך 784 כמטריצה במימד 28x28) והציגו אותו כתמונה. הוסיפו לתמונה את התיוג הרלוונטי ודונו בקצרה בתוצאות שקיבלתם.

3. (5 נקודות) חשבו והציגו עבור סדרת האימון את

א. דיוק המסווג $P_c(f, \mathcal{D})$

ב. מטריצת הערבול (confusion matrix)

דונו בתוצאות שקיבלתם, תוך התייחסות מפורשת לכל אחד מהמדדים הנ"ל.

4. (10 נקודות) חיזרו על שאלה 3 עבור סדרת המבחן הנתונה בקובץ TestData.pkl. דונו <u>בהרחבה</u> בתוצאות שקיבלתם, תוך התייחסות <u>מפורטת</u> לדומה ולשונה בין הביצועים שמדדתם עבור כל אחת מהסדרות.

בהצלחה!