

למידה חישובית 1 (096411)

אביב 2024

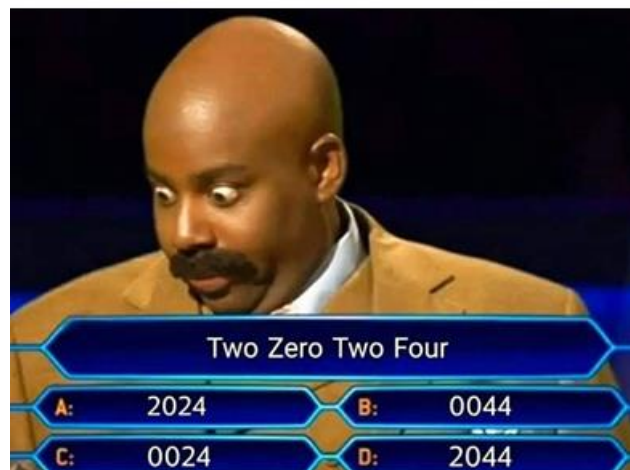
תרגיל בית 5

תאריך אחרון להגשה: 25.08.2024 בשעה 23:59

הוראות הגשה

- ההגשה בזוגות בלבד, דרך "קבוצה" ייעודית שיצרתם במודל.
- עליכם להגיש קובץ pdf בודד:
 - HW5_ID1_ID2.pdf – קובץ המכיל תשובות לכל השאלות. עבור שאלה 2, יש להוסיף צילומי מסך של הקוד והפלטים שהוא מפיק. ניתן גם לייצא מחברת בפורמט PDF ולשרשר אותה לתשובות לחלק היבש (במקום המתאים).
- קוד חייב להיות קריא, תמציתי ומתועד היטב. יש להקפיד על שימוש בשמות משמעותיים למשתנים.
- כל גרף חייב להכיל לפחות את האלמנטים הבאים: כותרת, מקרא (legend), כותרות לצירים ויחידות (ticks).
- יש להשתמש בפורום במודל לטובת שאלות על התרגיל. השאלות שלכם עוזרות לסטודנטים אחרים בקורס.

Teacher: The HW isn't hard.
HW:



בהרצאה למדנו על האלגוריתם **AdaBoost** שמשלב קבוצת לומדים חלשים לכדי לומד חזק אחד.

כפי שאתם זוכרים, בהינתן משימת סיווג בינארית עם מדגם אימון $S = \{(x_i, y_i)\}_{i=1}^m$,

כאשר $y_i \in \{+1, -1\}$ **AdaBoost** מבצע את הצעדים הבאים:

1. $D^{(1)} \leftarrow \left(\frac{1}{m}, \dots, \frac{1}{m}\right)$

2. for $t=1, \dots, T$:

i. $h_t \leftarrow WL(D^{(t)}, S)$

ii. $\epsilon_t = \sum_{i=1}^m D_i^{(t)} 1_{[y_i \neq h_t(x_i)]}$ and $w_t = \frac{1}{2} \log \log \left(\frac{1}{\epsilon_t} - 1\right)$

iii. $D_i^{(t+1)} \propto D_i^{(t)} e^{-w_t y_i h_t(x_i)}$

3. Output \hat{h} where $\hat{h}(x) = \text{sign}(\sum_{t=1}^T w_t h_t(x))$

1. הסבירו את ההנחה על $WL(\cdot, \cdot)$ וכל צעד באלגוריתם.

2. הוכיחו שהשוויון הבא מתקיים:

$$\sum_{i=1}^m D_i^{(t)} e^{-w_t y_i h_t(x_i)} = \epsilon_t \cdot e_t^w + (1 - \epsilon_t) \cdot e_t^{-w_t} = 2\sqrt{\epsilon_t(1 - \epsilon_t)}$$

בסעיף הבא תוכיחו כי אם $\forall t \in [T], \epsilon_t \leq \frac{1}{2} - \gamma$ אז:

$$L_S(\hat{h}) \leq e^{-2\gamma^2 T}$$

מספר הגדרות:

נגדיר את $f_0(x) \equiv 0$ ואת $f_t(x) := \sum_{p=1}^t w_p \cdot h_p(x)$ $\forall t \in N$,

מבאן ש- $\hat{h}(x) := \text{sign}(f_T(x))$

בנוסף, $z_t := \frac{1}{m} \sum_{i=1}^m e^{-\gamma_i f_t(x_i)}$ נגדיר $\forall t \in N \cup \{0\}$

3. ענו על הסעיפים הבאים:

1. הוכיחו כי $\forall i \in [m]: 1_{[y_i \neq \hat{h}(x_i)]} \leq e^{-\gamma_i f_T(x_i)}$ ולכן $L_S(\hat{h}) \leq z_T$.

2. הוכיחו באינדוקציה כי $\forall t \in N, i \in [m]$ מתקיים:

$$D_i^{(t)} \propto e^{-\gamma_i f_{t-1}(x_i)}$$

3. הוכיחו כי $\forall t \in N$

$$\frac{z_t}{z_{t-1}} = 2\sqrt{\epsilon_t(1 - \epsilon_t)}$$

4. הוכיחו כי אם $\forall t \in [T], \epsilon_t \leq \frac{1}{2} - \gamma$ אז:

$$2\sqrt{\epsilon_t(1 - \epsilon_t)} \leq e^{-2\gamma^2}$$

רמז: הפונקציה $g(a) = a \cdot (1 - a)$ היא מונוטונית עולה בתחום $\left[0, \frac{1}{2}\right]$. בנוסף, $\forall a: 1 - a \leq e^{-a}$.

הראו כי $z_T = \prod_{t=1}^T \frac{z_t}{z_{t-1}}$ והוכיחו כי אם $\forall t \in [T], \varepsilon_t \leq \frac{1}{2} - \gamma$ אז:

$$L_S(\hat{h}) \leq e^{-2\gamma^2 T}$$

4. הניחו כי $\forall t \in [T], \varepsilon_t \leq \frac{1}{2} - \gamma$. מהו מספר האיטרציות המינימלי T שאלגוריתם **AdaBoost** צריך לבצע בכדי להבטיח שהמסווג שיתקבל \hat{h} ישיג בהכרח שגיאת אימון השווה ל-0?
רמז: איזה ערכים $L_S(\hat{h})$ יכול לקבל?

הוכיחו שהשגיאה של h_t ביחס להתפלגות D^{t+1} היא בדיוק $\frac{1}{2}$. כלומר הוכיחו ש $\forall t \in [T]$ - מתקיים:

$$\sum_{i=1}^m D_i^{t+1} \cdot 1_{[y_i \neq h_t(x_i)]} = \frac{1}{2}$$



בשאלה הזאת נעבוד עם סט הנתונים wine-quality שמצורף לתרגיל במודל. סט נתונים זה מכיל 1599 תצפיות עם 11 פיצ'רים. העמודה quality מתארת את התויות של כל יין (ציון הנע בין 3 ל-8).

בצעו את השלבים הבאים:

- טענו את קובץ הדאטה (winequality-red.csv) והמירו אותו ל-DataFrame של pandas.
- המירו את העמודה של quality לעמודה בינארית (כלומר הפכו את התויות להיות בינארית). כאשר כל תצפית עם ערך $quality > 5$ תקבל תווית 0 ואחרת 1. המטרה תהיה לחזות את התויות הבינארי החדשה.
- חלקו את הנתונים למטריצת פיצ'רים (X) ווקטור לייבלים (y).
- חלקו את סט הנתונים למדגם אימון ומדגם מבחן באמצעות הפקודות הבאות:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.40, random_state=42)
```

בסעיפים הבאים נשתמש במודלים הבאים מתוך הספרייה של sklearn:

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
```

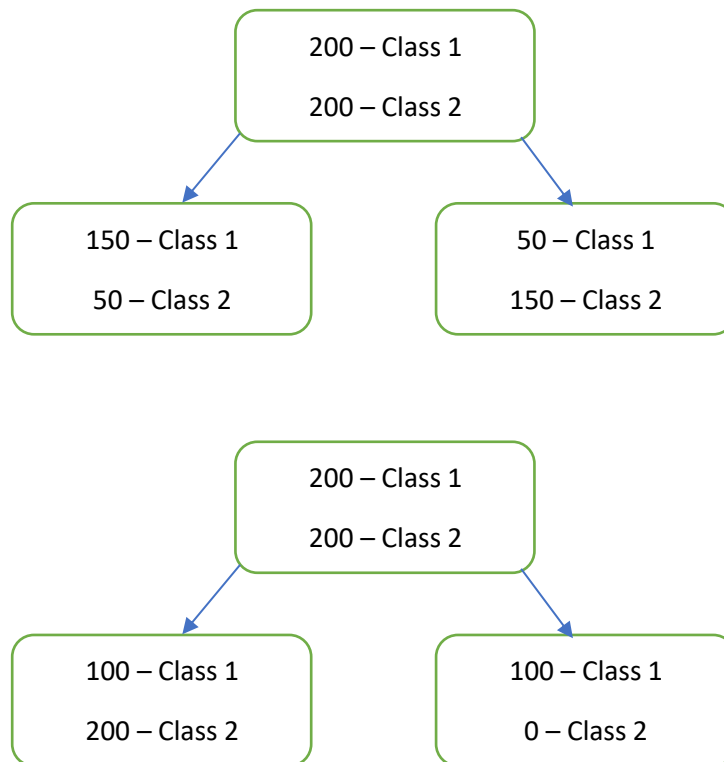
1. הגדירו מודל של עץ החלטה DecisionTreeClassifier, עם הפרמטרים הבאים:
`max_depth=12, random_state=0`
הפעילו את המודל על מדגם האימון. דווחו אחוז דיוק על מדגם האימון ומדגם המבחן.

2. הגדירו מודל של RandomForestClassifier, עם הפרמטרים הבאים:
`n_estimators=100, max_depth=12, random_state=0`
הפעילו את המודל על מדגם האימון. דווחו אחוז דיוק על מדגם האימון ומדגם המבחן.
הציגו גרף המתאר את אחוז הדיוק על מדגם המבחן בלבד כפונקציה של מספר העצים (n_estimators) כאשר טווח הערכים של מספר העצים הוא מ 1 עד 100 עצים.

3. הגדירו שוב מודל של RandomForestClassifier עם אותם פרמטרים כמו בסעיף (ב) רק שעכשיו נרצה להוריד את מנגנון האקראיות בבחירת הפיצ'רים כך שבכל פיצול לא תהיה דגימה אקראית של תת קבוצת פיצ'רים אלא שכל פיצול יסתכל על אוסף כל הפיצ'רים. (רמז: השתמשו בפרמטר max_features).
הפעילו את המודל על מדגם האימון. דווחו אחוז דיוק על מדגם האימון ומדגם המבחן.

הציגו גרף המתאר את אחוז הדיוק על מדגם המבחן בלבד כפונקציה של מספר העצים ($n_estimators$) כאשר טווח הערכים של מספר העצים הוא מ 1 עד 100 עצים.

4. על סמך התוצאות והגרפים שהצגתם:
- האם הביצועים של המודל בסעיף (ב) הוא יותר טוב או יותר גרוע מהביצועים של המודל בסעיף (א). למה הוא יותר טוב/גרוע? הסבירו את תשובתכם.
 - האם השינוי הנעשה בסעיף (ג) נתן ביצועים טובים יותר או גרועים יותר בהשוואה ל Random Forest הרגיל שהגדרתם בסעיף (ב). למה הוא נתן ביצועים טובים/גרועים יותר? הסבירו את תשובתכם.
- סעיף ה הוא סעיף נפרד ובלתי תלוי בסעיפים הקודמים.
5. נניח שיש מדגם של בעיית סיווג בינארית עם 200 תצפיות לכל מחלקה. נסתכל על שני פיצולים אפשריים של התצפיות (בכל צומת כתוב מספר התצפיות שיש בכל מחלקה):



- איזה פיצול לדעתכם הוא פיצול טוב יותר?
 - לכל אחד משני הפיצולים חשבו את מדדי ה- $impurity$ הבאים: $gini$, $entropy$ ו- $misclassification\ rate$.
- הסיקו איזה מדד מתוך השלושה לא כדאי להשתמש בשביל בניית עץ החלטה.