Decision Trees, Ensemble Methods – 5 תרגיל בית

יש להגיש שני קבצים נפרדים: קובץ PDF ובו פתרון התרגיל כולל הפלטים של החלק המעשי וקובץ נוסף ובו הקוד שכתבתם. יש להקפיד על תשובות ברורות ומסודרות ועל קוד מסודר ומתועד היטב. רק אחד מבין חברי הזוג צריך להגיש את הפתרון.

שאלות על התרגיל יש לכתוב בפורום תרגילי הבית באתר הקורס. התרגיל מנוסח בלשון נקבה אך מתייחס לשני המינים.

שאלה 1

נתון אוסף התצפיות הבא:

ID	Family heart attacks	Gender	Cholesterol	Blood pressure	Heart attack
1	Yes	Male	160	High	No
2	Yes	Male	260	Normal	Yes
3	No	Female	245	Normal	No
4	No	Male	170	Normal	No
5	Yes	Female	230	High	Yes
6	No	Male	215	Normal	No
7	Yes	Female	240	Normal	No
8	No	Male	235	High	Yes

א. נרצה לנבא האם אדם יקבל התקף לב או לא. על פי הנלמד בהרצאה ובתרגול, ציירו את עץ ההחלטה המתאים לנתונים. על מנת לחשב השתמשו ב Entropy.

ב. כעת השתמשו בGini. האם העץ הינו אותו עץ?

ג. בהסתמך על הסעיפים הקודמים, האם השימוש בGini וEntropy יכול ליצור את אותו העץ? אם כן - האם העץ תמיד יהיה אותו עץ? במידה וכן – הסבירו. במידה ולא – הראו דוגמה נגדית.

שאלה 2 – שאלה ממבחן

אלגוריתם ה AdaBoost מבצע שלושה צעדים:

i) $h_t = WL(D^{(t)}, S)$

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

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

א. הסבירו את ההנחה על $WL()$ ואת כל אחד מהצעדים באלגוריתם.

ב. הראו ש $2\sqrt{\epsilon_t(1-\epsilon_t)} = \sum_{i=1}^{m} D_i^{(t)} e^{-w_t y_i h_t(x_i)}$.

ג. הראו ש $1/2 = \sum_{i=1}^{m} D_i^{(t+1)} 1_{[h_t(x_i) \neq y_i]}$.

ד. מהו המספר המינימלי של צעדים הנדרש באלגוריתם ה AdaBoost בכדי שפונקציית ההפסד שלו תהיה 0 על מדגם האימון?

שאלה 3 – חלק רטוב

בחלק זה של התרגיל נממש עץ החלטה ע"פ האלגוריתם CART שנלמד בהרצאה ובתרגול. נעבוד עם מדד Entropy.

1. הData עליו נעבוד הינו Data אודות אבחון של סרטן השד.
   a. על מנת לטעון את הData אפשר להשתמש בפונקציה load_breast_cancer של sklearn. קישור - https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html#sklearn.datasets.load_breast_cancer
   b. לפני מתחילים לעבוד, מומלץ לקרוא את תיאור הדאטה - https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
2. ממשו את מדד הEntropy
   a. קלט הפונקציה – רשימה של נתונים
   b. פלט הפונקציה – ערך הEntropy
3. ממשו את בניית העץ על פי האלגוריתם שראינו בתרגול ע"פ אלגוריתם CART
   a. מומלץ להיעזר בחבילות מובנות המממשות עצים (כדוגמת anytree) אך אין חובה
4. פצלו את הData למדגם אימון (80%) ומדגם מבחן (20%), את הפיצול יש לבצע באמצעות הפונקציה train_test_split עם random_state=3.
5. בנו את העץ על פי מדגם האימון ובדקו את התוצאות על מדגם המבחן
6. על מנת לבדוק את עצמכם, בדקו מול הרצה של האלגוריתם הממומש בsklearn עם אותם הפרמטרים (הציון בשאלה זו יקבע לפי דיוק הפלט אל מול הפלט מsklearn)
7. ציירו את העץ שקיבלתם (ניתן לצייר אותו בכל דרך שהיא ולצרף לפתרון)
8. הציגו את התוצאות בConfusion Matrix