תרגיל בית 3 – מבוא ללמידה

הנחיות כלליות:

- תאריך ההגשה: 01/07/2021 ב23:59
- התרגיל הוא להגשה **בזוגות.**
- המתרגל האחראי על התרגיל: רפאל גד, נא לפנות בשאלות אך ורק למייל ai.technion@gmail.com לאחר שבדקתם ששאלתכם לא נמצאת כבר בFAQ.
- הקוד שלכם ייבדק אוטומטית וגם ידנית, רמאות כלשהי תוביל לועדת משמעת.
- אנה הקפידו על ההנחיות, כל הפרה תגרום להורדת ניקוד לרבות אי עמידה בציפיות פורמט השונות. בסוף התרגיל תמצאו סיכום של כל הטעויות הנפוצות בהנחיות/ פורמט. נא עיינו בו לפני הגשה.
- קיימת הגבלת שורות לפתרון שלכם לכל סעיף.
- מותר להשתמש בספריות scipy ,sklearn, pandas ,numpy, random, matplotlib, csv, math, datasets, statistics. אך כמובן שאין להשתמש באלגוריתמי הלמידה שתתבקשו לממש בעצמיכם במהלך התרגיל.
- שימו לב אסור שאחד הקבצים שתגישו ידפיס פלט או גרף בהרצתו אלא הפלטים צריכים לנבוע מקריאת הפונקציות כמבואר בכל שאלה.

מצורף לכם דטה של בדיקות רפואיות.כל שורה מתארת בדיקה של אדם בודד כאשר העמודה הראשונה מציינת האם האדם חולה (M) או בריא (B) ושאר העמודות כל מיני תכונות רפואיות של אדם (התכונות קצת מורכבות ואינכם צריכים להתייחס למשמעות שלהם כלל). עליכם לבנות כל מיני אלגוריתמים בשאלות הבאות על מנת לאפשר חיזוי מדוייק כמה שיותר להאם אדם חולה על בסיס תכונות אלה. יצויין שהדטה שמחולק לכם לא מחולק לtrain ו test כנלמד בתרגול שכן הtest יבוצע אצלנו ולכן מוגש לכם רק ה train. בהצלחה.

חלק א' – ID3:

1. **ממשו את אלגוריתם ID3**

שימו לב שכל התכונות בדאטה רציפות, אתם מתבקשים להשתמש בשיטה של חלוקה דינמית המתוארת בהרצאה. במקרה שיש כמה תכונות אופטימליות בצומת מסויים בחרו את התכונה בעלת האינדקס המקסימלי וכאשר יש כמה ערכי סף עבור אותה תכונה אופטימליים בחרו את גדול מביניהם. כמו כן כאשר בוחנים ערך סף לפיצול של תכונה רציפה, דוגמאות עם תכונה שערכה שווה לערך הסף שנבחן משתייכות לקבוצה עם הערכים הגדולים מערך הסף. ממשו אותו בקובץ בשם ID3.py.

על הקבוץ להכיל מחלקה בשם ID3 ועל המחלקה לתמוך בפונקציה:

fit_predict(self, train, test)

שבעזרתה האלגוריתם שלכם לומד על הקבוצה *train* ונבחן על הקבוצת test ומחזיר numpy_array שבמיקום *i* מכיל את סיווג של השורה *ה*i *בקבוצת test.*

ליתר דיוק:

train הינו מערך מטיפוס *numpy.ndarray עם מימדים:*

(#number of samples, number of features + 1)

כאשר העמודה הראשונה מתארת את הסיווג.

Test הינו מערך מטיפוס *numpy.ndarray עם מימדים:*

*(#number of samples, number of features)*כאשר הסיווג לא נתון במערך test.

הפלט צריך להיות גם הוא מטיפוס *numpy.ndarray עם מימדים: (#number of samples)*

כאשר אם השורה סווגה כחולה יש לרשום 1 ואם סווגה כבריאה 0.
וכאשר שני קבוצות אלה train/test הינם באותו פורמט בדיוק פרט למספר השורות לדטה שנתון לכם ולעובדה שבtest הסיווג לא נתון. כמובן שאתם יכולים לממש כל פונקציה/מחלקה נוספת שיעזרו לכם.

(30 נק')

2. **הוכח/הפרך**: "בהינתן דאטה כלשהו עם תכונות רציפות ותיוגים בינאריים המחולק לקבוצת אימון ומבחן, הפעלה של פונקציית נירמול MinMax הנלמד בתרגול על הדאטה אינה משפיעה על דיוק של מסווג ID3 הנלמד על קבוצת האימון והנבחן על קבוצת המבחן"(10 נק'). תשובה פורמלית לגמרי מצופה. (מוגבל ל20 שורות)

3. **הוסיפו לעץ שלכם גיזום מוקדם.**
   1. צרפו בדו"ח בלבד גרף המציג את השפעת הגיזום המוקדם על דיוק העץ, המחושב בעזרת K-fold cross validation. נסו לפחות חמישה ערכים שונים לגיזום המוקדם, דהיינו מספר מינימלי של דוגמאות בעלים על מנת להמשיך לפצל. יש לחשב כל דיוק (נקודה בגרף) באמצעות K-fold cross validation על כלל הדטה. כדי לבצע את החלוקת K סטים יש להשתמש בפונקציה sklearn.model_selection.KFold עם הפרמטרים n_split=5, shuffle=True ו- random_state שווה למספר תעודת הזהות שלכם.
   2. הסבירו מה החשיבות של הגיזום באופן כללי ואיזה תופעה הוא מנסה למנוע?
   3. תנו אינטרפרטציה לגרף שקיבלתם.
   4. לאיזה גיזום קיבלתם התוצאות הטובות ביותר ומה היא תוצאה זו?

   על המימוש שלכם להימצא גם כן בקובץ ID3.py ועל הניסוי שלכם להימצא בפונקצייה נפרדת בשם experiment בתוך אותו קובץ יחד עם שורה בהערה בקוד שמאפשר להריץ את הניסוי (אין חתימה מיוחדת לפונקציה זו אלא תוכלו לבחור בעצמיכם). (10 נק') [אורך התשובה מוגבל ל8 שורות]

4.
במקרה שלנו, בשל אופי הבעיה, שגיאת סיווג אדם חולה כבריא חמורה פי 8 מסיווג של אדם בריא כחולה. לפיכך, הדיוק שהצגתם בשאלות 1 ו- 3 אינו משקף היטב את טיב המסווג שלכם. בעזרת מישקול סוגי השגיאות, בהתאם למה שלמדתם בתרגול נגדיר פונקציה חדשה שתייצג את טיב המסווגים שלנו.

$$loss(T) \coloneqq FP + 8FN$$

כשאר FP הינו מספר הסיווגים השגויים של בריאים כחולים של T ו-FN הינו מספר הסיווגים השגויים של חולים כבריאים של T. כאשר T הינו מסווג.

1) צירפו כאן בלבד את הloss של המסווג ID3 עם הגיזום האופטימלי שמצאתם כאשר הוא נבחן בעזרת K-fold cross validation באותה צורה מאשר בשאלה 3. (דהיינו צירפו את ממוצע 5 הלוסים)
2) חישבו את הלוס שהיה מתקבל אם כל הדוגמאות היו מסווגים כחולים. השוואו את התוצאה הזאת עם התוצאה שקיבלתם בסעיף הקודם והסבירו את התוצאה.

3) תארו דרך לגרום לID3 ללמוד מסווג אשר ממזער את פונקציית ה- loss שהוצגה כאן בצורה טובה יותר מאשר ID3 הרגיל ומהפתרון הטריוויאלי שהוצג בסעיף 2.
ממשו הצעתכם בקובץ PersonalizedID3.py. על הקובץ להכיל מחלקה בשם PersonalizedID3 ועל המחלקה לתמוך בפונקציה:

fit_predict(self, train, test)

שבעזרתה האלגוריתם שלכם לומד על הקבוצה *train* ונבחן על קבוצת test ומחזיר numpy_array שבמיקום *i* מכיל את סיווג של השורה *ה*i *בקבוצת test*.
ליתר דיוק:
train הינו מערך מטיפוס *numpy.ndarray עם מימדים:*
(#number of samples, number of features + 1)
כאשר העמודה הראשונה מתארת את הסיווג.
Test הינו מערך מטיפוס *numpy.ndarray עם מימדים:*
(#number of samples, number of features)
כאשר הסיווג לא נתון במערך test.
הפלט צריך להיות גם הוא מטיפוס *numpy.ndarray עם מימדים:*
(#number of samples)
כאשר אם השורה סווגה כחולה יש לרשום 1 ואם סווגה כבריאה 0.
וכאשר שני קבוצות אלה train/test הינם באותו פורמט בדיוק פרט למספר השורות לדטה שנתון לכם ולעובדה שבtest הסיווג לא נתון. כמובן שאתם יכולים לממש כל פונקציה/מחלקה נוספת שיעזרו לכם.

אופי שאלה זו הינו מחקרי, עליכם לתאר את השיקולים שהביאו אתכם לבחירת האלגוריתם שהצעתם לרבות כל דבר שניסיתם. בהתאם לטיב הפתרון שלכם תקבלו ציון. כאשר טיב פתרון מוגדר על תוצאות אמפיריות ועל היקף ועומק המחקר שתעשו.
כמו כן אם ביצעתם ניסויים לקביעת פרמטרים לאלגוריתם שלכם הציגו אותם כאן.
שימו לב, אינכם צריכים לדאוג מכך שהשיפורים שלכם יפגעו בדיוק ובלבד שהם ישפרו את הloss.
<u>3 הזוגות בעלי התוצאות הטובות ביותר על הטסטים שלנו יקבלו 3 נקודות בונוס על ציון הקורס.</u>
מגבלות על השיפור, הקוד שלכם צריך לא לקחת יותר מסדר גודל של 15 דקות לרוץ בהינתן דטה באותו סדר גודל מהנתון לכם ובהיותו רץ על מחשב סטנדרטי. כמו כן השיפור שלכם חייב להישאר בגדר של עץ דהיינו אסור להשתמש כאן ביער או לעבור לסוג אלגוריתם שונה לגמרי. כמו כן עיבוד מקדים אינו נחשב שיפור לאלגוריתם אך אינו אסור בתוספת לשיפור אחר.

(50 נק') [אורך התשובה מוגבל ל2 עמודות]

הוראות הגשה:

- הגישו קובץ zip בודד המכיל את כל הקבצים שהתבקשתם לממש יחד עם קבצים נוספים כרצונכם בלי קבצי הדאטה שצורפו לכם.
- ודאו שכל הקבצים שלכם מחזירים מה שביקשנו מכם ושלא עשיתם אף אחת מהטעויות המוצגות להלן.
- נוהל האיחורים מופיע בסילבוס הקורס.

טעויות נפוצות בהנחיות

לנוחיותכם רשימה חלקית של טעויות נפוצות בקיום ההנחיות אשר גם יגרמו להורדת נקודות:

- שימוש בספריות שלא אושרו במפורש.
- פונקציות שלא עומדות בחתימות המתבקשות.
- שמות לא נכונים לקבצים שהתבקשתם להגיש בדגש על אותיות גדולות/קטנות ורווחים מיותרים.
- שמות לא נכונים למחלקות ופונקציות שהתבקשתם להגיש.
- הגשת קבצים אשר מדפיסים פלטים או גרפים בפרט שכחת printים.
- פלטים לא מתאימים לפונקציות שהתבקשתם לכתוב.

הרבה בריאות והצלחה.