מכון טכנולוגי חולון
Holon Institute of Technology

50058מדע נתונים- תאוריה ומעשה | סמסטר א' תשפ"א

פרויקט מסכם

בפרויקט זה עליכם להתחלק לצוותים של שלושה סטודנטים לכל היותר. כל צוות יבחר באחד הפרויקטים המופיעים בטבלה וישתבץ באמצעות רישום עצמי (ללא כפילויות- צוות לכל פרויקט), בכתובת הבאה:

https://docs.google.com/spreadsheets/d/1vwlnCG2zuCpmVRUsGqYno2CooBBxnRTVm1u_gBViokg/edit?usp=sharing

במודול תוכלו למצוא תיקייה שכוללת את כל סטי הנתונים הרלוונטיים, מחברת לדוגמא לפרויקט ה- Titanic וכן מחברת "load data", שקוראת את סטי הנתונים ל-pandas data frame.

עליכם להציג את הפרויקט במסגרת יום הצגת הפרויקטים. שעות ההצגה תקבענה בהמשך, גם בהתאם לכמות הסטודנטים שיירשמו לכל אחד מהימים. אנא שריינו את היום שבחרתם לצורך העניין.

דרישות המצגת- במסגרת הצגת הפרויקט, עליכם להתייחס לאלמנטים הבאים:

1. הגדרת הבעיה- מה מטרת הפרויקט? עם איזו בעיה הוא מתמודד? מהם הנתונים הזמינים לצורך הפרויקט (input)? מה המוטיבציה לפרויקט ואפליקציות רלוונטיות.
2. הערכת ביצועים- מהי המטריקה שבה בחרתם להשתמש לצורך הערכת ביצועי המודל ומדוע; כיצד חלקתם את סט הנתונים?
3. נתונים- תאור סט הנתונים:
   (1) כמה דוגמאות יש בסט?
   (2) כמה מאפיינים יש בסט?
   (3) מהי התפלגות ה- labels?
   (4) האם ישנם ערכים חסרים?
   (5) הצג 2-3 גרפים המתארים את האספקטים השונים של סט הנתונים.
4. הנדסת נתונים (-data engineering)
   (1) האם הסרתם מאפיינים כלשהם?
   (2) האם הוספתם מאפיינים כלשהם?
   (3) כיצד התמודדתם עם ערכים חסרים?
5. ביצועי המודל- נתחו והשוו את ביצועי של המודלים הבאים-
   (1) מודל "baseline"- בבעיית רגרסיה מודל זה יכול להתבסס, לדוגמא, על הערך הממוצע; בבעיית סיווג- מודל זה יכול להתבסס על אחד המסווגים הבסיסיים וסט המאפיינים הראשוני.

מכון טכנולוגי חולון
Holon Institute of Technology

(2) מודלים מוכרים-

a. K nearest neighbors, try 3 different k values.
b. K nearest neighbors with scaled values, try 3 different k values.
c. Decision tree with 3 different max depth values.
d. Random forest with 3 different max depth values and 100 estimators.
e. Ada Boost with 3 different max depth values and 100 estimators.
f. Lasso regression with scaled values, try 3 different alpha values (only for regression).

6. בחן את המודלים שקבלת. לדוגמא, במקרה של random forest כדאי לבחון את ה-feature importance המתקבל.
7. היפר-פרמטרים- השתמש במודל שהניב את התוצאות הטובות ביותר. חפש את ההיפר-פרמטרים של המודל אשר משפרים בצורה הגדולה ביותר את ביצועי המודל. הצג השוואת ביצועים עבור ערכי היפר-פרמטרים אלו.
8. ניתוח נוסף (בחר אחד מהבאים)-

א. ביצועים כפונקציה של גודל סט הנתונים- השתמש במודל המוצלח ביותר. הצג גרף של ביצועי המודל כאשר משתמשים ב- 10%, 30%, 50%, 70% או 100% מסט האימון. האם היית ממליץ להגדיל את סט הנתונים?
ב. מיצוע מודלים- בחר שלושה מודלים שונים שבחנת קודם לכן. צור מודל "חדש" אשר מבוסס על הממוצע של מוצאי שלושת המודלים שבחרת. השווה את הביצועים של מודל זה ביחס לכל יתר המודלים שבחנת בפרויקט.

אין צורך להגיש דו"ח ייעודי, אלא להכין מחברת ג'ופיטר מסודרת, עם תיעוד מתאים של הקוד. את הקישור למחברת יש לעדכן בכתובת שצוינה מעלה (אין להגיש את המחברת.)

בהצלחה!

צוות הקורס