תרגיל בנושא סיווג רקטות

לפניך סט של כ 28,000 מקטעי מסלולי רקטות - [train.csv.gz](https://drive.google.com/file/d/0B_bjbieS2ePVWVJYeVEzcFo4MW8/view?usp=sharing)

* כל מקטע מסלול הוא בין 5 ל 15 שניות, עם תדר דגימה של 2 הרץ (דגימה כל חצי שנייה). בכל דגימה נתונים מיקום XYZ ומהירות XYZ מורעשים.
* כל מקטעי המסלולים מיושרים לכיוון המהירות ההתחלתית ומתחילים מ X=0, Y=0 ובגובה האמיתי. הכיוון החיובי של ציר Z הוא כלפי מעלה, כלומר VelZ חיובי מציין מטרה בעלייה

הכרות עם הדאטה

1. הורידו את סט המסלולים
   1. טענו אותו כ dataframe. השתמשו בעמודה הראשונה כאינדקס, ומחקו את העמודה 'targetName'.
   2. הציגו מהי כמות המסלולים הקיימים מכל סוג
   3. הציגו היסטוגרמה של אורך המסלולים מכל סוג (היסטוגרמה שונה לכל סוג)

ציור

1. כתבו פונקציה באמצעות matplotlib (או כל חבילה אחרת) שמציירת מסלולים של רקטה ( ניתן להתעלם מקואורדינטת ה Y שערכה קרוב ל 0)
   1. ציירו את המסלול הראשון.
   2. ציירו את 50 המסלולים הראשונים של רקטות מסוג 1-6. צבע שונה לכל סוג של רקטה. האם ניתן להבדיל בין הסוגים השונים?
   3. ציירו את 50 המסלולים הראשונים מסוג 1 ומסוג 6 שאורכם 15 שניות. האם ניתן להבחין בהבדל ביניהם?
   4. חזרו על הסעיף הקודם תוך התמקדות ברקטות שגם עולות וגם יורדות. האם קל יותר להבחין בין הרקטות?

הכנת דאטה - 2 קבוצות

1. צרו טבלה חדשה מהטבלה - בה קיימים רק מטרות מסוג 1 וסוג 16
   1. פצלו את הטבלה לשתי טבלאות חדשות: 80% מהשורות בטבלה חדשה שתקרא סט אימון ו 20% מהשורות בטבלה חדשה שתקרא סט בדיקה. הגרילו אילו שורות יהיו בסט האימון ואלו בסט הבדיקה
   2. פרקו את העמודה האחרונה (הסיווג) בטבלת סט הבדיקה לוקטור נפרד, כך שבטבלת סט הבדיקה לא יהיה סיווג (אך שמרו את וקטור הסיווג לשימוש עתידי)
2. סיווג – מבוסס כללים (RULE BASED)

הביטו במידע הגולמי ובמסלולים שציירתם כדי למצוא מסווג.

* 1. השתמשו בסט האימון (ובו בלבד) כדי לכתוב תוכנה ובה אוסף כללים אינטואיטיביים שעליהם חשבתם בעצמכם שיכולים להכריע איזה סוג רקטה קיבלתם.

זכרו כי המטרה היא להיות טובים ככל האפשר, אבל גם אם לא ניתן להגיע ל100 אחוזי הצלחה, עדיין אפשר לקבל תוצאה טובה משמעותית מסיווג אקראי לחלוטין.

* 1. הפעילו את הכללים על סט הבדיקה. הציגו את ה Confusion matrix (<https://en.wikipedia.org/wiki/Confusion_matrix>) וחשבו את ה F1 score (https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9).

סיווג – למידת מכונה

1. בחרו מסווג (למשל random forest שנתמך בחבילת sklearn)
   1. השתמשו בסט האימון כדי לאמן מודל בעזרת המסווג שנבחר.
   2. הפעילו את המודל על סט הבדיקה. הציגו את ה Confusion matrix (<https://en.wikipedia.org/wiki/Confusion_matrix>) וחשבו את ה F1 score (https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9).
   3. השוו למסווג ה rule\_based
   4. בחרו מסווג אחר וחזרו על שני הסעיפים האחרונים. איזה מסווג טוב יותר?
2. חזרו על שאלות 3-4-5 עבור זוגות המטרות (3,9), (5,6), (12,15)
   1. איזה זוג מטרות הכי קשה לסווג?
   2. לכל זוג מ 4 הזוגות האלו – ציירו מסלולים ונסו להבין מדוע יש זוגות שקשה יותר להבחין ביניהם.
   3. נסו לסווג את אותן זוגות מטרות שסווגו בשאלות 3-6. השתמשו הפעם בטבלה החדשה. האם התוצאות השתפרו?

הכנת דאטה וסיווג - 4 קבוצות

1. חזרו לשאלה 3 וצרו טבלה בה קיימות מטרות מסוגים 1,4,7,10.
   1. חלקו את הטבלה לסט אימון וסט בדיקה
   2. צרו מסווג rule based בעזזרת סט האימון
   3. בחרו מסווג למידת מכונה וצרו מודל בעזרת סט האימון.
   4. הפעילו את מסווג ה rule based ואת מודל למידת המכונה על סט הבדיקה.
   5. הציגו את ה Confusion matrix וחשבו את ה F1 score (איך מחשבים F1 score כשיש 4 אפשרויות?).
   6. ציירו מסלולים של מטרות מסוגים אלו ונסו להבין את התוצאות.