פרויקט - חלק א' (תשפ"ד סמסטר א')

זהו החלק הראשון מבין שני חלקים.

הנחיות הגשה: דו"ח התרגיל הראשון וקוד ה- Python שכתבתם יוגשו לתיבת ההגשה במודל עד לתאריך ה-8.2.2024 בשעה 23:59. מספיקה הגשה של אחד מבני הזוג.

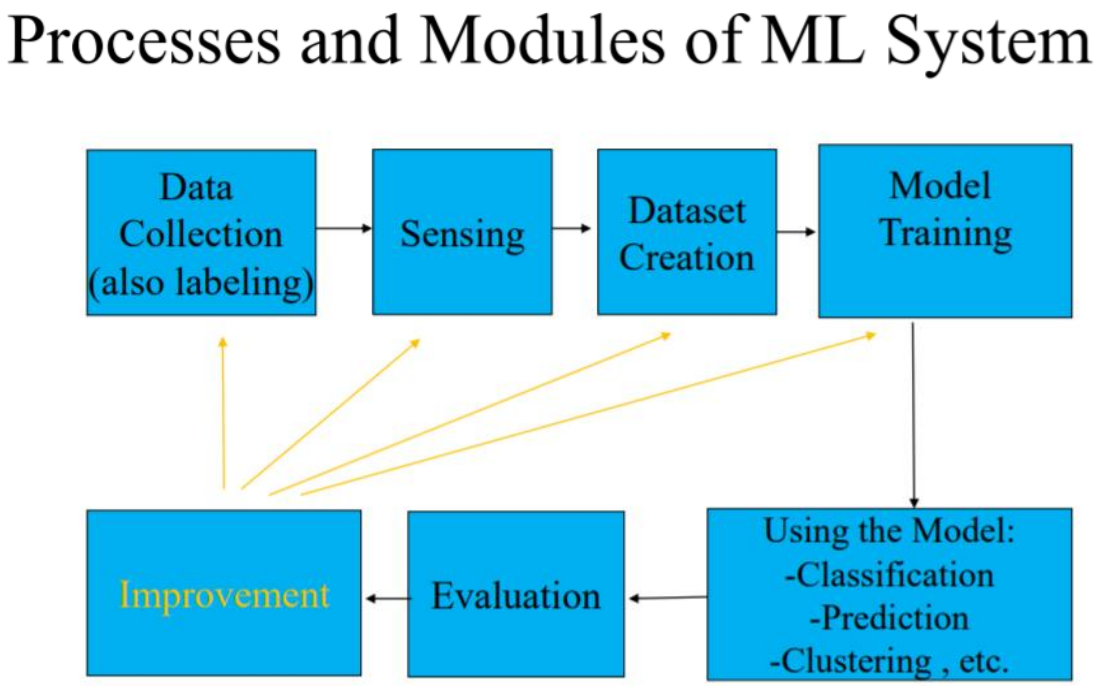
מטרת התרגיל: בתרגיל זה נתרגל את שלושת השלבים הראשונים בתהליך יצירת מכונה לומדת. למעשה השלב הראשון והשני (Data Collection, Sensing) כבר ניתן לכם, ולכן מרבית העבודה תהיה בשלב השלישי (Dataset Creation). שלבים אלה ישמשו אותנו בהבנת והכנת הנתונים לקראת התרגיל השני בו נשתמש בנתונים לאימון ובחינה של מערכות לומדות והשוואה ביניהן.

צוותי הגשה: הגשת התרגיל הינה בזוגות, בהתאם לקבוצות המוגדרות במודל.

דגשים לדו"ח: **אורך הדו"ח לא יעלה על 10 עמודים** (לא כלל עמודים נלווים כמו שער ותוכן עניינים), בגודל כתב 12, פונט Arial ורווח של שורה וחצי. **חריגה מהגדרות אלו תגרור הורדת נק'**. יש לשמור על תמציתיות ולהתמקד בתובנות המרכזיות שלכם בכל סעיף. יש להגיש קובץ Word המאפשר השארת הערות על גביו, **אין להגיש קובץ PDF!**

שפת תכנות: Python (סביבת הפיתוח היא לבחירתכם- Spyder, Jupyter, etc.).

שאלות בנוגע לתרגיל, יש לפרסם בפורום הייעודי שייפתח במודל בלבד, פניות למייל לא יענו!



איור 1: תמכורת מההרצאה - שלבי פרויקט machine learning

הסבר אודות בסיס הנתונים לפרויקט

את קבצי הנתונים של הפרויקט ניתן למצוא בMoodle .

רשתות חברתיות רבות משמשות מקור עשיר לסנטימנט ודעות ציבוריות. הפרויקט שלנו עוסק בניתוח הסנטימנט של ההודעות. אנו שואפים להבין את הסנטימנט (חיובי\ שלילי) העומד מאחורי ההודעות ולספק תובנות לגביו. ניתוח זה יכול לשמש כפילטר להתבטאויות בוטות.

XY_train.pkl -קובץ המשתנים המסבירים והמשתנה המוסבר עבור האימון (עליו עובדים בחלק זה)

הסבר על המשתנים בבסיס הנתונים

משתנה	הסבר
textID	מזהה ייחודי של ההודעה
text	ההודעה שנשלחה על ידי משתמש כשלהו
sentiment	**משתנה המטרה**, הסנטימנט המשויך להודעה
message_date	תאריך שליחת ההודעה
account_creation_date	התאריך שבו המשתמש נוצר
previous_messages_dates	תאריכים קודמים בהם נשלחו הודעות על ידי המשתמש
date_of_new_follower	תאריכים קודמים בהם המשתמש קיבל עוקבים
date_of_new_follow	תאריכים קודמים בהם המשתמש עקב אחר משתמש חדש
email	האימייל המקושר למשתמש
gender	מגדר השולח
email_verified	האם האימייל מאומת
blue_tick	האם המשתמש מאומת
embedded_content	תוכן נוסף בגוף ההודעה
platform	הרשת החברתית שבה נשלחה ההודעה

הקדמה:

- אחד האתגרים בכניסה לעולם של machine learning הוא הכרה והתמצאות בכל המונחים ושיטות העבודה. אנחנו מאמינים שכדי לצלוח זאת עליכם לתרגל ולהשתמש במונחים מקצועיים כמה שיותר כדי לחבר למונחים את ההבנה של הרעיונות שאותם הם מייצגים. כל המונחים בעולם זה הם באנגלית ולכן גם המצגות כתובות באנגלית, בנוסף בדו"ח הפרויקט נרצה לראות אתכם משתמשים במונחים מקצועיים ומכניסים אותם בהקשרים הנכונים בעבודה.
- בתרגיל יושם דגש על שימוש בתוכנה לצורך מענה על שאלות הקשורות בנתונים ובניתוחם. לא פחות חשובות הן התובנות משימוש זה לגבי עולם התוכן של הבעיה הנחקרת, כשהשאלה המרכזית הינה: מה בעצם למדנו מתרגיל זה? יש לשלב טבלאות, גרפים וכו', לנתחם ולהשליך מהם על עולם התוכן הנחקר.

- יש להגיש את כל קבצי Python עליהם עבדתם.

מבנה העבודה

Data collection and Sensing (10 נק')

1. **הנתונים אותם קיבלתם כבר עברו את 2 השלבים הראשונים.**

- ענו בקצרה: מהו הData collection ? איזה סוג Sensing בוצע על הדאטה (סטטי \ דינמי), הסבירו ?
- הציעו סוג Sensing שלא בוצע על הדאטה, והסבירו איך הוא יכול לעזור למשימת הלימוד (אין צורך להוסיף אותו פיזית, אלא רק ברמה הקונספטואלית)
- מהי קטגוריית וסוג משימת הלמידה (נלמד במצגת 3)? הסבירו את תשובתם. האם ניתן להשתמש בנתונים כדי לבצע עוד סוג של משימת למידה ?

Dataset Creation (75 נק')

1. **Exploratory data analysis – 20 נק'**

הציגו את התפלגות הנתונים של כל המשתנים בסט הנתונים כולל משתנה המטרה, מה מסקנותיכם בנוגע למשימת הלימוד? (השתמשו בגרפים ומדדים סטטיסטיים כדי להמחיש את מסקנותיכם ושמרו על תמציתיות בהסברים). בשלב זה ניתן גם להראות סטטיסטיקות הנובעות מטרנספורמציות על הנתונים (למשל חישוב אורך מערך).

אין לחזור במילים על מה שניתן לראות בגרף. עליכם להסביר את המסקנות שלכם עבור כל משתנה.

שאלות מכוונות:

- מהו טווח הערכים שכל משתנה מקבל?
- מה משמעות המשתנה בהקשר של משימת הלימוד ?
- האם הוא קשור להבנתם של עוד משתנים בDataset?
- על מה מלמדות הסתברויות אלה?
- האם סט הנתונים מאוזן? במה זה תלוי?
- קשרים מעניינים בין משתנים מסבירים

2. כעת עבדו עלפי שלבי העבודה שלמדנו בהרצאה 2 – שימו לב, ייתכן ולא כל השלבים רלוונטיים! במידה ושלב מסוים אינו רלוונטי, הסבירו מדוע הוא אינו רלוונטי (על ההסבר להיות משכנע)

- כאשר אתם בוחרים לבצע פעולה על הנתונים, הסבירו בתמציתיות מהו התהליך שאתם רוצים לבצע, ואיך בחרתם לבצע אותו ?
- עבדו באופן סיסטמתי על פי השלבים, ציינו כל שלב בכותרת מתאימה, ושאבו רעיונות מהמצגת (כאן נצפה לראות שימוש במונחים ובשיטות אותם למדנו בהרצאה).
- בשלב זה אין תשובה אחת נכונה, כל קבוצה יכולה לבחור איך לבצע שלבים מסוימים על פי שיקול דעתכם ובמידה וניתן לכך הסבר מספק.
- שימו לב לתוכן של כל שלב על פי המצגת. **הכנסת פעולה לא בשלב הנכון תוריד ניקוד!**
- הניקוד על כל שלב יהיה:

  1. dimensionality reduction – 5 נק'
  2. feature extraction; pre-processing - 10 נק' עבור כל סעיף
  3. feature selection, feature representation – 15 נק' עבור כל סעיף
- חלק מציון חלק זה יהיה על שלבי עבודה ברורים והצגתם באופן כזה שיהיה ניתן להבין מה עשיתם. כל פעולה שאתם מבצעים ציינו עם הפרדה כך שיהיה קל להבחין (לא לרשום פסקה אחת ארוכה). **דוח מבולגן יוריד ציון !**
- **בסוף כל שלב, במידה וביצעתם שינויים בנתונים, הציגו סיכום קצר בו אתם מציינים איזה שינוי ביצעתם בנתונים. הציגו פלט של כמה רשומות מתוך ה dataset שיצרתם על מנת שיהיה ניתן לראות ויזואלית מה עשיתם! ניתן להוסיף כל פלט או הסבר שתומך ומסכם כל שלב.**

Validation (5 נק')

- על סמך החומר שלמדנו בכתה בהרצאה 3, בחרו שיטת ולידציה לנתונים. הסבירו מדוע בחרתם בה ? מה היתרונות שלה על פני שאר השיטות שלמדנו ?
- הסבירו את תהליך הוולידציה שתבצעו בעזרת השיטה שבחרתם. באיזה מטריקה תבחרו להשתמש?

איכות הדו"ח ורמת שימוש בתכנת Python (10 נק')

- איכות הדוח – ויזואלית, סידור, קריא
- איכות השימוש בתוכנה – מורכבות הגרפים, הפלטים, והשיטות

בהצלחה !