מבוא ללמידת מכונה
מדעים להייטק
אוניברסיטת תל אביב
אביב 2019-2020
מרצה: דור בנק
מתרגל: יעד טובלי

פרויקט מסכם

מסמכים להגשה:	1. קובץ קוד מוכן בפורמט py. או בפורמט Jupyter notebook, להחלטתכם. 2. מסמך PDF לדו"ח הפרויקט 3. קובץ csv בשם Submission_group_number.csv אשר כולל את תחזיות הקלסיפיקציה (Prediction Probabilities - מצ"ב פורמט)
תוכנות דרושות:	Anaconda for python 3
אמצעי הגשה:	קובץ zip שיכלול את שלושת הקבצים הנ"ל
יצירת קשר:	yaadtovli@mail.tau.ac.il
מועד אחרון להגשה:	25.6.2020 בשעה 23:59

כללי

בפרויקט זה יינתנו לכם נתונים (מספר פיצ'רים) אודות יום מסוים, תפקידכם לבנות מערכת המנאבת האם עתיד לרדת גשם ביום שלמחרת. כלומר בפרויקט נעסוק בבעיית Binary Classification (כלומר – משני קלאסים) בה עליכם לסווג רשומות ל**שתי קטגוריות** – האם עתיד לרדת גשם למחרת (סיווג 1), או לא (סיווג 0), על סמך מספר פיצ'רים בדאטה סט. חלק מן הפיצ'רים ידועים (כדוגמת לחות, מהירות רוח, טמפ' מקסימלית וכדומה) וחלקם אנונימיים.

המטרה בפרויקט המסכם הינה לחשוף אתכם לעבודה מעשית בה תוכלו להתנסות בחומר הנלמד בקורס, תוך כדי יישום הכלים בסביבת נתונים אמתית לחלוטין. אין בכוונת הצוות להגביל אתכם בצורת החשיבה והעבודה, אולם קיימות מספר הנחיות בסיסיות אשר עליכם לעמוד בהן.

- אין להכין ולעצב את סט הנתונים הגולמי בעזרת אקסל (אלא בחבילות python בלבד).
- מימוש הקוד ייעשה במחברת/בקובץ פייתון בסביבת Anaconda, ויכלול **הסברים מלאים** על אופי המימוש בקוד עצמו (בעזרת markdowns למשתמשים במחברת או notes למשתמשים בקובץ py.)
- משך הרצת הקוד מתחילתו ועד סופו לא תעלה על שעה.
- תצטרכו להשתמש אך ורק בקבצים train.csv וב-test_without_target.csv
- כפי שתלמדו במהלך הסמסטר אין חובה להשתמש בכל הפיצ'רים של סט הנתונים, תוכלו להנדס ולעצב את הפיצ'רים כרצונכם.
- טיב הביצועים של המודל שלכם ייעשה **על ידי מטריקת AUC**.
- שימושים בפונקציות וטכניקות שלא נלמדו בקורס הם מבורכים, אולם הם אינם מהווים תחליף לשיטות המסורתיות.
- **בונוס** יינתן למי שיעשה שימוש עשיר בוויזואליזציה של הנתונים.
- **בונוס** יינתן גם למי שיבנה קוד קריא, יעיל וקונפיגורבילי אשר מקל על האפשרות לחקור ולנסות כיוונים חדשים. שימושים בפונקציות ואובייקטים משלכם מבורכים.

- **בונוס** נוסף יחולק לסטודנטים לפי ביצועי המודל שלהם – 3 נקודות למקום הראשון, 2 נקודות למקום השני, נקודה למקום השלישי.

- **קנס** גדול יינתן על קוד שלא רץ ועל קובץ csv שלא מוגש בפורמט המבוקש.
- יש לפרט את ההנחות שנלקחו בכל שלב של הפרויקט. הנחות שלא יפורטו יחשב כאילו לא נלקחו בחשבון ומצב זה עלול להוביל להורדה בציון.
- גם אם ניסתם "לפצח" את הבעיה בדרך מסוימת ואין שיפור בתוצאות: אל תסירו את הניסיון מהקוד. רק חשוב שתדגישו שמדובר בניסיון לא מוצלח ואין לו מקום ב-work flow הסופי.

משימת התכנות (הניקוד עבור הסעיפים השונים בסוגריים)

חלק ראשון - אקספלורציה:

- עליכם לחקור את הנתונים בכל אופן שבו עולה על רוחכם: האופי שבו כל פיצ'ר מתפלג, התנהגות קורלטיבית בין הפיצ'רים, נתונים סטטיסטיים על הפיצ'רים. בשלב זה של הפרויקט יש המון מקום לוויזואליזציה! נצלו זאת. **(5)**

חלק שני – עיבוד מקדים:

עבור השאלות אשר מופיעות בסעיפים יש לענות בגוף המחברת (Markdown) בצמוד לחלקי הקוד הרלוונטיים

- האם קיימים נתונים חריגים (Outliers) בדאטה? אם כן, עליכם להסירם או לפחות לתת עליהם את הדעת **(3)**
- האם הנתונים מנורמלים? אם לא- האם צריך לנרמל אותם? מה החשיבות של נרמול הנתונים בבעיה? **(5)**
- האם ישנם נתונים חסרים? כיצד בחרתם לטפל בהם ומדוע באופן זה? **(4)**
- האם המימדיות של הבעיה גדולה מדי? למה מימדיות גדולה עלולה ליצור בעיה? איך נזהה כי מימדיות הבעיה גדולה מדי? **(5)**
- הקטנת המימדיות על ידי טכניקה אחת שנלמדה בכיתה – PCA, ו/או על ידי בחירת תת קבוצה של פיצ'רים קיימים (Feature selection). **(12)**
- בניית פיצ'רים חדשים מניפולציה מתמטית על פיצ'רים קיימים **(2)**
- החלת העיבוד המקדים על סט ה-Test **(10)**
- ניתן לבצע ניסיונות נוספים אשר לא נלמדו בקורס על מנת לעבד את הפיצ'רים הניתנים לכם **(בונוס)**

חלק שלישי – הרצת המודלים:

- בניית שני מודלים ראשוניים מבין השלושה הבאים והחלתם על סט הנתונים: **(12)**
  - Naïve Bayes Classifier
  - KNN
  - Logistic Regression

- בחירת שני מודלים מתקדמים מבין הארבעה הבאים והחלתם על סט הנתונים: **(12)**

  - Multi-Layer Perceptron (ANN)
  - Decision Tree
  - Random Forest or Adaptive Boosting
  - Support Vectors Machine

<u>חלק רביעי – הערכת המודלים:</u>

- בניית Confusion Matrix (מדגמית) על אחד המודלים, עליכם להסביר מה אומרים התאים בתוך המטריצה בהקשר למודל שבחרתם **(5)**
- הערכת המודל באמצעות K-Fold Cross Validation, ובניית פלט ROC על כל K-Fold עבור כל אחד מהמודלים שהורצו (רצוי באותו התרשים) **(15)**
- פערי ביצועים בין הרצת המודל על ה-Train או על ה-Validation, האם המודל שלכם הוא Overfitted? מה עשיתם / עליכם לעשות על מנת להגדיל את יכולת ההכללה שלו? **(5)**

<u>חלק חמישי – ביצוע פרדיקציה</u>

- לאחר בחירת המודל, עליכם לבצע פרדיקציה על נתוני קובץ "test_without_target.csv", ולהגיש קובץ בפורמט csv בשם Submission_group_number.csv אשר כולל את תחזיות הסתברות הקלסיפיקציה (**Prediction <u>Probabilities</u>** - מצ"ב דוגמא). **(5)**

<u>הערות:</u>

1. **מיותר לציין שאת המודלים תצטרכו לבחון על סט Validation ולא על ה-Train עצמו (בחירת מודל לפי ביצועיו על ה-Train עלול להביא לתוצאות מאוד נמוכות ולהורדת ציון משמעותית!). הרצת הפרידקציות על ה-Train יכולה לסייע במציאת Overfitting אבל לא מהווה אינדיקטור לטיב המודל.**
2. עליכם לכתוב מפורשות את ההיפר פרמטרים של המודלים הנבחרים כפי שנלמדו בכיתה, גם אם הוחלט להשתמש בערכי ברירת המחדל שלהם.
3. סדר מימוש השלבים אינו מחייב, במסגרת הפרויקט סביר מאוד להניח שתצטרכו לחזור אחורה אל שלבים מוקדמים יותר (בדומה לכל פרויקט Data Science).

<u>- (הנחיות לדו"ח המסכם בעמוד הבא) -</u>

הדוח המסכם

הדו"ח יכלול **לכל היותר** 5 עמודים (לא כולל שער) אשר בו יוסברו כלל השלבים שננקטו במהלך ניתוח הנתונים. נדרש הסבר מפורט על הרציונל מאחורי בחירת כל אחת מהשיטות שצוינו בשלבים לעיל, של ההיפר-פרמטרים שנבחרו, ותוצאות המודלים השונים. אין צורך להרחיב במילים ואין צורך לצטט את חומר הקורס בפרויקט.

הדו"ח ייכתב בעברית בגופן Tahoma, עם רווח שורות של 1.15.

- על הדו"ח לכלול פרק "תקציר מנהלים" פסקה קצרה המסכמת את הפרויקט, וכן פרק "סיכום" אשר מתאר את כלל המודלים והמסקנות שנסקרו במהלך ניתוח הנתונים.
- יש לפרט את ההנחות שנלקחו בכל שלב של הפרויקט. הנחות שלא יפורטו יחשבו כאילו לא נלקחו בחשבון ומצב זה עלול להוביל להורדה בציון.
- ניתן להוסיף נספחים ככל העולה על רוחכם, ויזואליזציה תופיע בנספחים (וגם בגוף הקוד).
- יש לציין בראש הדו"ח את שם המגיש + ת"ז. את הדו"ח יש להגיש במערכת ה-Moodle עד התאריך 25.6.2020 .
- הדו"ח מהווה חלק בלתי נפרד מהפרויקט המסכם. **הניקוד אשר הוגדר בחלקי הפרויקט השונים נשען גם על טיב תיאורם בדו"ח.**

שימו לב: מועד ההגשה הינו סוף הסמסטר. לא יינתנו הארכות אז אנא תכננו את זמנכם בהתאם. בנוסף, יש להקפיד הקפדה יתרה על פורמט ההגשה.

בהצלחה!!!