

מבוא ללמידת מכונה ומדע הנתונים 320101

מטלה מסכמת

שנה"ל תשפ"ו, סמסטר א'

הוראות הגשה:

1. העבודה תבוצע ותוגש ע"י סטודנט יחיד או בזוג (כפי שהוגשו שאר המטלות).
2. ניתן להיעזר בכלי בינה מלאכותית, תחת התנאים הבאים:
 - א. יש לציין במפורש במסגרת ההגשה את כל המקומות בהם נעשה שימוש בתוצרים של כלים אלה.
 - ב. בכל שימוש ב AI יש לצרף את ה prompt שהוזן אליו, ואת הפלט שהתקבל ממנו (כנספחים להגשה).
 - ג. שימוש ב-AI עשוי להיות צעד ראשון לפתרון, ומצופה מהסטודנטים לערוך ולדייק אותו לנדרש בשאלה תוך מתן ביטוי לידע, להבנה, וליצירתיות שלהם. גם לאחר עריכת הפלט לפני ששולב בפתרון שהוגש, הנחיה ב' תקפה ונדרש להבהיר מה השינוי שנעשה ומה הסיבה לביצועו על-ידי הסטודנט/ים המגישים.
 - ד. אין לשלב בהגשה תוצר של כלי בינה מלאכותית ללא הבנה מלאה של התוצר, במידה השקולה לנדרש לשם כתיבתו מראשיתו ועד סופו על ידי הסטודנט/ים המגישים.
3. **שימו לב:** שימוש בכלי בינה מלאכותית ללא עמידה מלאה בסעיפים א-ד, שקולה להעתקה. בכל מקרה בו מתעורר ספק לגבי השימוש המותר בכלי בינה מלאכותית, נדרשת הסכמת המרצה לשימוש המבוקש מראש ובכתב.
4. את הפתרונות לשאלות המבוססות על קוד יש להגיש במחברת/ות jupyter עם תיעוד מלא של הקוד והתוצרים (גרפים, הדפסות טקסט וכיו"ב), כשכל מחברת מכילה את כל הדרוש להבנתה והרצתה. את ההסברים המפורטים של כל שלבי הפתרון יש להגיש במסמך Word/PDF נפרד.
4. נא הקפידו לפרט הנחותיכם ולבסס את מסקנותיכם בתוצאות הניסויים שביצעתם, כולל גרפים/טבלאות וכיו"ב במחברת ובמסמך המלווה – לאיכותם ובהירותם של מרכיבים אלה יינתן משקל מרכזי בציון.
5. למרצה שמורה האפשרות לזמן את הסטודנטים להגנה (בחינה בעל פה) לפני מתן ציון לעבודה.
6. באופן כללי הציון של הגשה בזוג יהיה זהה לשני הסטודנטים המגישים אך, במידה ויהיה פער בבחינה בעל פה בין ביצועי הסטודנטים המגישים, הציון לכל סטודנט יהיה שונה.

בהצלחה!

חלק ו: אימון והערכת ביצועי מסווגים (80 נקודות)

בחלק זה של הפרויקט זה הינכם מתבקשים לאמן ולבחור, עבור המידע המצורף ומבין כל המסווגים שלמדנו בקורס את המסווג שלהערכתכם סיכויי לסווג נכון מידע לא מתויג הינם הגבוהים ביותר.

הערה: כפי שדנו בהרצאה, המידע שישמש אותנו במסגרת הפרויקט נלקח מקישור [זה](#). לנוחותכם צורפו הדאטה ותיאורו, ואין חובה לעשות שימוש בקישור הנ"ל.

- PassengerId - Each Id takes the form gggg_pp where gggg indicates a group the passenger is travelling with and pp is their number within the group. People in a group are often family members, but not always.
- HomePlanet - The planet the passenger departed from, typically their planet of permanent residence.
- CryoSleep - Indicates whether the passenger elected to be put into suspended animation for the duration of the voyage. Passengers in cryosleep are confined to their cabins.
- Cabin - The cabin number where the passenger is staying. Takes the form deck/num/side, where side can be either P for *Port* or S for *Starboard*.
- Destination - The planet the passenger will be debarking to.
- Age - The age of the passenger.
- VIP - Whether the passenger has paid for special VIP service during the voyage.
- RoomService, FoodCourt, ShoppingMall, Spa, VRDeck - Amount the passenger has billed at each of the *Spaceship Titanic*'s many luxury amenities.
- Name - The first and last names of the passenger.
- Transported - Whether the passenger was transported to another dimension. This is the target, the column you are trying to predict.

1. שלב א – חקר נתונים

טענו את הדאטה ובצעו Exploratory Data Analysis (EDA) מפורט תוך שימוש במדדים כמותיים ובויזואליזציות מפורטות, כולל

א. בחנו את מבנה המידע, מספר הדוגמאות והעמודות, בדיקת סוגי המשתנים, האם ישנם ערכים חסרים

ב. בחנו את התפלגויותיהם של המשתנים השונים

ג. בחנו האם יש איזון בין שכיחויות ערכי התיוגים

- ד. בחנו האם יש משתנים עם מתאם גבוה ביניהם
- ה. דונו בפירוט במסקנות מכל אחד מהשלבים הנ"ל, ופרטו בחינות נוספות במידה וביצעתם כאלה

2. שלב ב – הכנת המידע לאימון מסווג

- א. טענו את הדאטה (8693 דוגמאות מתויגות) שקיבלתם וחלקו אותו לשתי סדרות ללא דוגמאות משותפות, סדרת אימון $\mathcal{D}_{\text{train}} = \{x_{\text{train},n}, y_{\text{train},n}\}_{n=1}^{N_{\text{train}}}$ וסדרת מבחן $\mathcal{D}_{\text{test}} = \{x_{\text{test},n}, y_{\text{test},n}\}_{n=1}^{N_{\text{test}}}$ כאשר $N_{\text{train}} + N_{\text{test}} = 8693$. הסבירו את שיקוליכם בחלוקת הדאטה לשתי הסדרות כולל התייחסות לבחירת הערכים עבור $N_{\text{train}}, N_{\text{test}}$.
- ב. טפלו בערכים חסרים בדאטה. אם בחרתם להשלים הערכים, תארו בפירוט איך עשיתם זאת ומדוע באופן זה. אם בחרתם להסיר דוגמאות עם ערכים חסרים מהמידע, נמקו היטב מדוע העדפתם זאת (תוך התייחסות למטרת המטלה – אימון המסווג המיטבי לסיווג דאטה לא מתויג).
- ג. אם ויתרתם על מאפיין (עמודה במידע) שניתן לכם, נמקו את בחירתכם.
- ד. אם הגדרתם מאפיינים חדשים (feature engineering) על-סמך המידע שקיבלתם, הסבירו בפירוט כיצד חישבתם אותם ונמקו בקצרה מדוע לדעתכם הם עשויים לשפר את ביצועי המסווג.
- ה. במידת הצורך, יצגו משתנים קטגוריאליים באמצעות קידודים כמותיים.
- ו. פרטו ונמקו כל פעולת preprocessing אחרת שבחרתם לבצע.

3. שלב ג – אימון ובחירת המסווג

מבין המסווגים שלמדנו

- KNN
- QDA
- LDA
- GNB
- Decision Tree
- Random Forest
- SVM
- Logistic Regression
- Multilayer Perceptron
- xgboost

מיצאו את המסווג אשר, ע"פ תוצאות ניסוייכם, דיוקו הצפוי על דוגמאות עתידיות לא מתויגות הינו הגבוה ביותר הניתן להשגה.

הינכם מתבקשים **לנמק בפירוט ובבהירות** כל אחד מהשלבים באימון המסווגים השונים, את שיקוליכם בנוגע למסווג שבחרתם מבין כולם כולל בחירת ערכי ה hyperparameters שלו במידה ויש כאלה, ומדוע לדעתכם סיכויי הדיוק שלו הם הגבוהים ביותר. בתשובתכם ציינו את המסווגים שבחנתם, תארו מילולית והציגו גרפית את התוצאות שקיבלתם, ונמקו את המסקנות שהגעתם אליהם בהתבסס על תוצאות הניסויים שביצעתם. יש להשתמש בכל השיטות שדנו בהן בהקשר של אימון מסווג כולל

- cross-validation
- חישובי accuracy, precision, recall
- שימוש בדיאגרמות ROC
- וכיו"ב, ע"פ הצורך ובהתאם לשיקול דעתכם.

חלק II - Clustering (20 נקודות)

בחלק זה של התרגיל תשתמשו באלגוריתמי clustering מסוג Gaussian | K-Means ו Mixture Models (GMM) בכדי למצוא קבוצות בתמונות של ספרות בכתב יד (MNIST), ו תמדדו עד כמה הקבוצות המתקבלות מיושרות עם התוויות האמיתיות של הספרות.

4. שלב א – חקר נתונים

טענו את הדאטה Digits MNIST ובצעו Exploratory Data Analysis (EDA) מתאים תוך שימוש במדדים כמותיים ובויזואליזציות.

5. שלב ב – אישכול למספר קבוצות השווה למספר התיוגים

א. השתמשו ב-K-Means עם $K=10$ ובצעו אישכול של הדוגמאות בדאטה (ללא שימוש בתיוגים הנתונים)

ב. שיוך קלאסטרים לתוויות: מאחר ולמספרי הקלאסטרים אין משמעות סמנטית, שייכו כל קבוצה לתווית הנפוצה ביותר בה, והגדירו תווית זו כתווית המייצגת של הקלאסטר.

ג. הגדירו והציגו גרסת confusion matrix המתאימה לבעיה הנתונה ומידדו, עבור כל תיוג אפשרי בדאטה, את מספר הדוגמאות שהקלאסטר אליו הן שייכות מיוצג עם התווית האמיתי שלהן.

ד. הציגו מספר דוגמאות מכל קלאסטר, ולצידן את הממוצע של הקלאסטר (כתמונה), וציינו את התיוג המייצג של הקלאסטר.

האם הקלאסטר קוהרנטי ויזואלית?

עבור כל תיוג, עם איזה קבוצות נוטות הדוגמאות שלו להתקבץ?

6. חיזרו על שאלה 5 תוך שימוש ב Gaussian Mixture Model. דונו במשותף ובהבדלים בין השימוש ב-Kmeans וב-GMM מבחינת זמני ריצה, תוצאות האישכול, וכל היבט אחר שהינו רלוונטי לדעתכם.

7. שלב ג – אישכול למספר קבוצות הגדול ממספר התיוגים

חיזרו על שלב ב (שאלה 5) למעלה עם K גדול יותר ממספר התיוגים האפשריים (למשל $K=20,30,50$). לכל ערך של K , השוות את התוצאות לאלו שהתקבלו בשאלה 5. השתמשו בויזואליזציה מפורטת ובמדדים כמותיים היכן שניתן בכדי להציג את תוצאותיכם ולתמוך במסקנותיכם.

בהצלחה!