

מבוא ללמידה מerb ומדע הנתונים 320101

מטלה מסכמתה

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 בכלים 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 לעורוך ולדיק אותו לנדרש בשאלת תוקן מתן ביטוי לידי, להבנה, וליצירותיות שלהם. גם לאחר עריכת הפלט לפני שלווב בפתרון שהוגש, הנחיה ב'תקפה ונדרש להבהיר מה השינוי שנעשה ומה הסיבה לביצועו על-ידי הסטודנטים המגישים.
 - ד. אין לשלב בהגשה תוצר של כלים מלאכותית ללא הבנה מלאה של התוצר, במידה הסקוללה לנדרש לשם כתיבתו מראשיתו ועד סופו על ידי הסטודנטים המגישים.
3. שימוש בכלים מלאכותית ללא עמידה מלאה בסעיפים א-ד, שකולה להעתקה. ככל מקורה בו מתעורר ספק לגבי השימוש המותר בכלים מלאכותית, נדרש הסכם המרצה לשימוש המבוקש מראש ובכתב.
4. את הפתרונות לשאלות המבוססות על קוד יש להגיש במחברת/jupyter עם תיעוד מלא של הקוד והertools (גרפים, הדפסות טקסט וכיו"ב), כשל מחברת מכילה את כל הדרוש להבנתה והרצתה. את ההסברים המפורטים של כל שלבי הפתרון יש להגיש במסמך Word/PDF נפרד.
5. נא הקפידו לפרט הנחותיכם ולבסס את מסקנותיכם בתוצאות הניסויים שביצעתם, כולל גרפים/טבלאות וכיו"ב במחברת ובמסמך המלווה – לאיכות ובהירותם של מרכיבים אלה ניתן משקל מרכזי בצינון.
6. באופן כללי הציון של הגשה בזוג יהיה זהה לשני הסטודנטים המגישים אך, במידה יהיה פער בבחינה בעל פה בין ביצועי הסטודנטים המגישים, הציון לכל סטודנט יהיה שונה.

בצלחה!

חלק א: אימון והערכת ביצועי מסותגים (80 נקודות)

ב חלק זה של הפרויקט זה הינכם מתבקשים לאמן ולבחר, עבור המידע המצורף ומבחן כל המסותגים שלמדנו בקורס את המסוג שלהערכתכם סיכיו לסוג נכון מידע לא מתויג הינם הגבויים ביותר.

הערה: כפי שדנו בהרצאה, המידע שיישמש אותנו במסגרת הפרויקט נלקח מ קישור [זה](#).
לנוחותכם צורפו הדטה וטיורו, 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 הדטה וביצעו (EDA) Exploratory Data Analysis מפורט תוך שימוש במידדים
כמותיים וביזואלייזציות מפורטוות, כולל

א. בחנו את מבנה המידע, מספר הדוגמאות והעמודות, בדיקת סוג המשתנים, האם ישם
ערכים חסרים

ב. בחנו את התפלגיותיהם של המשתנים השונים

ג. בחנו האם יש איזון בין שכיחויות ערכי התוצאות

ד. בחרנו האם יש משתנים עם מתאם גבוה ביןיהם.
ה. דונו בפתרונות מסוימים מכל אחד מהשלבים הנ"ל, ופרטו בחינות נוספות במידה וביצעתם ככלא

2. שלב ב – הכנות המידע לאימון מסויים

א. טענו את הדאטה (8693 דוגמאות מתיוגות) שקיבלתם וחלקו אותו לשתי סדרות ללא

דוגמאות משותפות, סדרת אימון $\mathcal{D}_{\text{train}} = \{\mathbf{x}_{\text{train},n}, y_{\text{train},n}\}_{n=1}^{N_{\text{train}}}$ וסדרת מבחן $\mathcal{D}_{\text{test}} = \{\mathbf{x}_{\text{test},n}, y_{\text{test},n}\}_{n=1}^{N_{\text{test}}}$ כאשר $N_{\text{train}} + N_{\text{test}} = 8693$. הסבירו את שיקוליכם בחלוקת הדאטה לשתי סדרות כולל התייחסות לבחירת הערכים עבור $N_{\text{train}}, N_{\text{test}}$.

ב. טפלו בערכים חסרים בדאטה. אם בחרתם להשלים הערכים, תארו בפתרונות איך עשיתם זאת ומדווע באופן זה. אם בחרתם להסיר דוגמאות עם ערכים חסרים מה מידע, נמקו היטב מדוע העדפתם זאת (טור התייחסות למטרת המטלה – אימון המסויים המיטבי לשיווג דאטה לא מתיוג).

ג. אם ויתרתם על מאפיין (עמודה במידע) שנייתן לכם, נמקו את בחירתכם.

ד. אם הגדרתם מאפיינים חדשים (feature engineering) על-סמן המידע שקיבלתם, הסבירו בפתרונות כיצד חישבתם אותם ונמקו בקצחה מדוע לדעתכם הם עשויים לשפר את ביצועי המסויים.

ה. במידה הצורך, יצגו משתנים קטגוריאליים באמצעות קידודים כמותיים.

ו. פרטו ונמקו כל פעולה preprocessing אחרת שבחרתם לבצע.

3. שלב ג – אימון ובחירה המסוים

מבין המסוים שלמדנו

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

מיצאו את המסוג אשר, ע"פ תוצאות ניסויים, דיווקו הצפוי על דוגמאות עתידיות לא מתואגות הינו הגבוה ביותר ביחס להשגה.

הינכם מתחקים **לنمך בפירוט ובבהירות** כל אחד מהשלבים באימון המסוגים השונים, את שיקוליםם בנוגע למסוג שבחרתם מבין כולן כולל בחירת ערכי ה **hyperparameters** השונים, שלו במידה ויש כ אלה, ומדוע לדעתכם סיכון הדיווק שלו הם הגבוהים ביותר. בתשובתכם ציינו את המסוגים שבחנתם, תארו מילולית והציגו גרפית את התוצאות שקיבלתם, וכן ציינו 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 בהתאם לשיקול דעתכם.

חלק II - Clustering (20 נקודות)

ב חלק זה של התרגיל תשתמשו באלגוריתמי clustering מסווג Gaussian ו K-Means ב כדי למצוא קבוצות בתמונות של ספרות בכתב יד (MNIST), Mixture Models (GMM) ו תמדו עד כמה הקבוצות המתקבלות מישרות עם התוויות האמיתיות של הספרות.

4. שלב א – חקר נתונים

טענו את הדטה MNIST Digits וביצעו (EDA) Exploratory Data Analysis מתאים תוך שימוש במדדים כמוותים וביוזאליזציות.

5. שלב ב – אישכל למספר קבוצות השווה למספר התוצאות

א. השתמשו ב-K-Means עם $K=10$ וביצעו אישכל של הדוגמאות בדטה (לא שימוש בתוצאות הנתונים)

ב. שיר קלאסטרים לתוויות: לאחר ולמספר הקלאסטרים אין משמעות סמנטית, שייכו כל קבוצה לתווית הנפוצה ביותר בה, והגידרו תווית זו כתווית המייצגת של הקלאסטר.

ג. הגידרו והציגו גרסה confusion matrix המתאימה לבעה הנתונה ומידדו, עברו כל תיוג אפשרי בדטה, את מספר הדוגמאות שהקלאסטר אליו הן שייכות מיצג עם התווית האמיתית שלהן.

ד. הציגו מספר דוגמאות מכל קלאסטר, ולצדן את הממציע של הקלאסטר

(כתמונה), וציינו את התיאוג המיצג של הקלאסטר.

האם הקלאסטר קוהרנטי ויזואלית?

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

6. חיורו על שאלה 5 תוך שימוש ב Gaussian Mixture Model. דנו במשותף

ובבדלים בין השימוש ב-K-means וב-GMM מבחינת זמני ריצה, תוצאות האישכול,

וכל היבט אחר שהינו רלוונטי לדעתכם.

7. **שלב ג – אישכל למספר קבוצות הגדול ממספר התיוגים**

חיורו על שלב ב (שאלה 5) למעלה עם K גדול יותר מאשר המספר התיאוגים האפשריים (למשל $K=20,30,50$). לכל ערך של K , השווות את התוצאות לאלו שהתקבלו בשאלה 5. השתמשו ביזואלייזציה מפורטת ובמדדים כמוותיהם היכן שניתן בכך להציג את תוצאותיכם וلتמוך במסקנותיכם.

בהצלחה!