לימוד מכונה 364-1-1811

תרגיל מספר 2

הנחיות הגשה: דו"ח התרגיל השני וקוד ה- Python שכתבתם יוגשו לתיבת ההגשה במודל עד לתאריך ה- 14.03.2024 בשעה 23:59.

מטרת התרגיל: בתרגיל נשתמש בנתונים שהכנו בתרגיל הראשון לצורך אימון ובחינה של מערכות לומדות והשוואה ביניהן. גם בתרגיל זה נשלב אלמנטים שנלמדו בכתה. נראה את החיבור של התיאוריה לכתיבה הפרקטית של קוד בעזרת חבילות מוכנות בpython.

דגשים לדו"ח: אורך הדו"ח **לא יעלה על 5 עמודים** (לא כולל עמודים נלווים כמו שער, תוכן עניינים), בגודל כתב 12, פונט Arial, רווח שורה וחצי. הדו"ח יוגש כקובץ word. חריגה ממספר עמודים זה תגרור הורדת נק'. **יש לשמור על תמציתיות ולהתמקד בתובנות המרכזיות שלכם בכל סעיף. אין להכניס פלטים מרכזיים בנספחים, אלא בגוף הדו"ח!**

שאלות בנוגע לתרגיל, יש לפרסם **בפורום הייעודי במודל.**

דגשים

- מחקר data science הוא אמפירי במהותו, כלומר אין פתרון אנליטי מוגדר מראש שמושג על ידי מידול נכון של הבעיה ופתירת משוואות עם תשובה סופית ברורה. במהלך המחקר יש בידיו של החוקר ארגז כלים ושיטות שבעזרתם הוא מתמודד עם הבעיה בדרך הטובה ביותר שניתן. ככל שארגז הכלים והשיטות שהחוקר מכיר נרחב יותר, כך הוא יוכל להתמודד עם מגוון רחב יותר של בעיות בעולמות תוכן שונים. אנחנו מעודדים פתרונות יצירתיים לבעיות שתתקלו בהם במהלך העבודה ועליהם יינתנו עד 5 נקודות בונוס לעבודה (עד ציון 100 בחלק ב' ולשיקול דעתו של הבודק בהתאם למידת ההשקעה בדוח והפגנת הידע). לכן חשוב לנו לחשוף אותכם למגוון כלים מוכרים שניתן להשתמש בהם כאשר מחפשים פתרון לבעיה במהלך מחקר:
  1. קבוצת machine and deep learning Israel בפייסבוק שם ניתן לערוך חיפוש ולקבל מגוון פתרונות של אנשים או לחלופין לעלות פוסט ולבקש עזרה בפתרון הבעיה.
  2. מאמרים מחקריים שאליהם ניתן לגשת בעזרת Google scholar.
  3. חיפוש בגוגל ובאתרים רלוונטיים כמו Stack Overflow, Kaggle (פתרונות קוד). בלוגים כמו Machine Learning Mastery, Medium, Towards Data Science (הסברים על קונספטים של ML).
- מומלץ להסתכל בדוקומנטציה של החבילות שאתם עובדים איתם, כדי להבין איזה אפשרויות קיימות (החבילות המרכזיות מכילות את מרבית הפונקציות שתצטרכו בעבודה שלכם, תחפשו טוב ותוכלו לחסוך זמן בכתיבת קוד נוסף ומיותר)
- ניתן בכל שלב בעבודה לשנות החלטות קודמות (לדוגמא, אם החלטתם בחלק א' להסיר משתנה מסוים וכעת אתם רוצים לבדוק האם הוספה שלו משפרת את ביצועי המודל/ים – כמובן, יש לציין זאת בדו"ח.)
- ניתן להשתמש בחבילות שלא נלמדו במעבדות (גם מימושים שונים של אותם המודלים) בחלק של Improvements .
- לעיתים שני מודלים דורשים Pre-processing (עיבוד מקדים) שונה. לדוגמא: משתנים קטגוריאליים בעצי החלטה ב-Python: עבור מודל מסוג DecisionTreeClassifier של sklearn יש להמיר את המשתנים הקטגוריאליים למשתני דמה (למשל, קידוד של 1/0). מומלץ (תמיד) לעיין בדוקומנטציה של החבילה שהשתמשתם בה כדי להימנע מטעויות כאלו.

- **הגשה – יש להגיש שלושה קבצים: 1) דו"ח בוורד . 2) קובץ קוד. 3) קובץ חיזויים (pkl) – בפורמט אחיד: G1_ytest להחליף את מספר הקבוצה רק.**

Model Training

- במידה ועוד לא חילקתם את הנתונים לפי שיטת הווליד‌ציה שבחרתם בחלק א' בצעו זאת כעת.
2 הבחירות הלגיטימיות לחלוקה הן :
1. holdout - כלומר, בחרתם לחלק את סט הנתונים לסט אימון וולידציה כך שאתם תאמנו את המודלים על סט האימון בלבד ותבחנו ותבצעו את תהליך Hyperparameter Tuning על ידי מקסום המטריקה הנבחרת על גבי סט הוולידציה
2. K-fold cross-validation - כלומר, בחרתם לבצע K חלוקות holdout שונות ולבחון את המטריקה הנבחרת על גבי ממוצע של K התוצאות שקיבלתם.

שימו לב שעליכם קודם לחלק את הנתונים המקוריים מXY_train.pkl ורק לאחר מכן לבצע את חילוץ הפיצ'רים שפיתחתם בחלק א', וזאת ללא הזלגת מידע. ראו:https://scikit-learn.org/stable/common_pitfalls.html

הערה כללית

- ענו על השאלות אך ורק עם המידע החשוב ביותר לצורך כך. פירוט יתר יקשה למצוא את התשובה שלכם ויוריד נקודות! אין צורך להסביר איך אלגוריתמים עובדים אם לא נדרשתם לכך.
- ענו על השאלות לפי הסעיפים וסמנו כל סעיף ככה שיהיה ניתן לזהות.

בקובץ הנתונים שקיבלתם משתנה המטרה אינו מאוזן! לכן, לא נכון לבחון את המודלים על פי קריטריון Accuracy. **במקום זאת יש לבחון את כל המודלים על פי AUC-ROC**

בכל מקום בו רשום "אחוזי דיוק" הכוונה היא לAUC-ROC

Decision Trees (25 נק')

1. בנו עץ החלטה ובצעו תהליך של Hyperparameter Tuning למציאת הקונפיגורציה המיטבית עבור מודל זה. (13 נק')

- הסבירו בקצרה איזה פרמטרים כיוונתם – מה משמעותם ואיזה טווח ערכים ניסיתם? (לא צריך להראות גרפים)
- מהם אחוזי הדיוק המתקבלים על סט האימון וסט הוולידציה עבור המודל הטוב ביותר ? מה ניתן להסיק מתוצאות אלה?

3. הציגו גרף של העץ שהתקבל (אם גדול מידי – הציגו רק את השכבות העליונות) (12 נק')

- ע"י התבוננות במבנה העץ, אילו תובנות הוא מספק על הבעיה ועל החשיבות השונה של מאפייני הבעיה? הסבירו את מסקנותיכם בהקשר של משימת הלמידה ולא באופן כללי.
- השתמשו בפונקציית חשיבות המשתנים של המודל (במודל DecisionTreeClassifier תכונת feature_importances). איזה מסקנות בנוגע למשימת הלימוד ניתן להסיק?

הציגו את מסקנותיכם ככה שיהיה ניתן להפריד ביניהם ולהבין אותן

Artificial Neural Networks (35 נק')

1. בצעו תהליך של Hyperparameter Tuning למציאת הקונפיגורציה המיטבית עבור מודל זה. (ניתן לכוון גם את מספר השכבות החבויות למרות שלא למדנו בהרצאה מודל Deep. לא נדרש מכם להכיר את היתרונות של "העומק" של הרשת על פני מודל "שטוח", אך עדיין תוכלו לכוון את מספר השכבות החבויות ולראות מה התוצאה שתתקבל) (23 נק')

- עבור כל היפר-פרמטר: מה הייתה המוטיבציה בבחירה לכוונן אותו? מה המשמעות על הרשת הנלמדת בהגדלתו/הקטנתו? איזה טווח ערכים בחרתם לכיוון כל פרמטר, מדוע?
- הציגו (גרפים ו/או טבלה) את ערכי ההיפר-פרמטרים שנבחנו כפונקציה של אחוז הדיוק על סט האימון והוולידציה. מה ניתן להסיק ממצאים אלו?
- מהם אחוזי הדיוק המתקבלים על סט האימון וסט הוולידציה באמצעות המודל המדויק ביותר ?
- מה ניתן להסיק מתוצאות אלה?
- איך ניתן להסביר את ההבדל מתוצאות סעיף 1 (אם קיימים) – באופן ספציפי עבור הפרמטרים שבחנתם?

שימו לב – בהרצאה לא למדנו על כל הפרמטרים שניתן לכוון במודל MLP.
כאן ניתן להרחיב ולקרוא באינטרנט על פרמטרים נוספים אותם תרצו "לכוון".
מומלץ להיעזר בדוקומנטציה של המודל באתר של חבילת התוכנה כדי להבין איזה פרמטרים ניתנים לכוונון.

2. אמנו רשת באמצעות הקונפיגורציה הנבחרת מסעיף קודם. (12 נק')
- הציגו מטריצת מבוכה עבור 2 המחלקות – מה ניתן להסיק על התנהגות המודל הטוב ביותר שלכם?

השוואה בין המודלים - Evaluation (10 נק')

1. השוו את ביצועי שני המודלים (DT, MLP)
- על איזה מהמודלים תמליצו? מדוע?

2. ציינו באופן בולט וברור באיזה אלגוריתם ומה דיוק הוולידציה שהתקבל באמצעותו

שיפור המודל הנבחר – Improvements (25 נק')

בסעיף זה תצטרכו להרחיב מעבר לידע של הקורס כדי לשפר עוד מעט את המודל הטוב ביותר שלכם, דבר שיוכל לסייע לכם בניקוד היחסי של החיזויים הסופיים בסעיף הבא.
הניקוד של סעיף זה יהיה תלוי במידת המורכבות והייחודיות של ההצעות לשיפור אותם תבחרו ליישם.

1. הציגו כמה מסקנות עיקריות משעלו לגבי הנתונים או המודל הנבחר שלכם לאורך העבודה אותם תרצו לשפר (5 נק')
- יש להציג כל בעיה בנפרד. לתת הסבר קצר עליה, להסביר איך אתם חושבים לפתור את הבעיה, ומה אתם מצפים לשפר בפתרון הבעיה

2. חפשו באינטרנט ובכל מקור ידע שזמין לכם (כדאי להסתכל על המקורות המומלצים לחיפוש בדגשים) רעיון לשיפור המודל הנבחר שלכם (10 נק')

- תארו את ביצוע השיפור
- האם הרעיון באמת שיפר את המודל שלכם ? (לא חובה שזה יקרה)
- אם כן, מדוע שיפר לדעתכם ? אם לא, מדוע לא שיפר לדעתכם ?

חייב להיות רעיון אחד לפחות של שיפור בנתונים, ורעיון אחד לפחות של שיפור במודל.
ניתן להרחיב ולהשתמש באלגוריתמים שלא נלמדו בקורס, **ניתן להשתמש בהגשה הסופית באלגוריתמים שלא נלמדו בקורס ועליהם הרחבתם בחלק זה**.
אם שיפרת את המודל הטוב ביותר שלכם ציינו שוב את דיוק הווליידציה שהתקבל באמצעותו

הגשת חיזויים סופיים (5 נק' + 2 נק' בונוס)

בצעו חיזוי באמצעות המודל שבחרתם, על קובץ הנתונים "X_test.pkl" הנמצא במודל. את החיזויים שיתקבלו העלו למודל כקובץ אקסל, על פי הפורמט שמופיע בקובץ הדוגמא "y_test_example.pkl" הנמצא במודל (הערכים בקובץ זה אקראיים). בדקו שהקובץ שאתם מגישים הוא בעל מבנה זהה ובעל אותו מספר רשומות הקיימות בקובץ X_test, ושמו לפי הפורמט: Gx_ytest (x הוא מספר הקבוצה שלכם)
(5 נק' - בהתאם למיקומכם ביחס לשאר הצוותים בקורס + 2 נק' בונוס - לקבוצה אשר תשיג את הביצועים הגבוהים ביותר. קבוצה שתקבל דיוק נמוך מהדיוק בעבודה שלהם ב 20% ויותר תקבל 0 על סעיף זה, מהסיבה שזאת דרך עקיפה לבדוק האם עשיתם את התהליך בצורה טובה)

שימו לב שאתם לא משנים את הסדר של הרשומות בקובץ x_test כדי למנוע בעיה בהשוואה לLABELS האמיתיים בסוף.
שימו לב שאתם מעבירים את הנתונים את אותו תהליך של עיבוד מקדים כמו שהעברתם את הנתונים שעליהם בניתם את המודל.

בהצלחה !