למידת מכונה 2020 – מטלה 3

תחרות

התחרות עבור מטלה 3 נקראת **Titanic: Machine Learning from Disaster**.

הקישור לתחרות נמצא כאן: https://www.kaggle.com/c/titanic

בתחרות זו עליכם לבצע משימת סיווג.

עליכם לסווג את נוסעי הטיטניק בין 2 תוויות, אלה ששרדו את ההתנגשות ואלה שלא שרדו אותה.

לרשותכם מספר מועט של פיצ'רים, שאיתם אתם יכולים לשחק. אתם יכולים להוריד פיצ'רים או לחשב פיצ'רים חדשים שמתבססים עליהם.

תיאור הדאטה נמצא כאן: https://www.kaggle.com/c/titanic/data

קראו אותו היטב, הבינו מה המשמעות של כל עמודה וחישבו כיצד כל פיצ'ר עשוי להשפיע על התוצאה (מידת ההישרדות של הנוסע).

במקרה שיש לכם שאלות בנוגע לתחרות ולדאטה, אתם תמיד יכולים לכתוב ולקרוא שאלות של אחרים בפורום של התחרות: https://www.kaggle.com/c/titanic/discussion

אם אתם רוצים לראות מחברות של אחרים, לקבל השראה מהדרך בה הם פתרו את הבעיה, אתם יכולים להיכנס ל- https://www.kaggle.com/c/titanic/notebooks

שימו לב שהניקוד בתחרות מתבצע על ידי דיוק (Accuracy):

> **Metric**
>
> Your score is the percentage of passengers you correctly predict. This is known as accuracy.

דרישות הפתרון

על הפתרון שלכם להופיע במחברת אחת שתוגש בשני קבצים: קובץ ipynb שניתן להורדה מתפריט ה- File במחברת, וקובץ ה- html שניתן ליצירה מקובץ ה- ipynb על ידי ההדרכה פה.

יש לשים את 2 הקבצים בקובץ ZIP (שימו לב שזה צריך להיות ZIP ולא RAR) יחיד ולהעלות את קובץ ה- ZIP למטלה במודל.

יש לרשום בתחילת המחברת את השם ותעודת הזהות ולצרף קישור לדף המשתמש שלכם ב- Kaggle.

אתם לא צריכים לבצע חקירה דאטה מכיוון שעשיתם זאת כבר בתרגיל הראשון.

אתם צריכים לקחת את המחברת שלכם מהמטלה הראשונה ולהמשיך אותה עבור המטלה הנוכחית (להוסיף כותרת "Exercise 3" בסוף המחברת הקודמת ולהמשיך תחתיה). בהזדמנות זו אתם יכולים לשפר ולסדר גם את החלק ששייך למטלה הראשונה.

עליכם לסווג את הדאטה על ידי שימוש ב- KNN, או ב- NBC או ב- LDA (בחרו אחד או יותר). הסבירו את בחירתכם ונמקו אותה (אם בחרתם ביותר מאחד, הסבירו מה ההבדלים ונתחו את התוצאות בהתאם).

השתמשו ב- CV (LPO, KFold). נסו אלגוריתמי Feature Selection, השתמשו ב- Ensembles מסוגים שונים (Bagging ו- Boosting), בצעו חיפוש היפר פרמטרים (Grid Search, Random Search), הסבירו את בחירותיכם ונתחו את התוצאות.

שימו לב שאתם יכולים להגיש רשמית עד 10 הגשות ביום:

Submission Limits

You may submit a maximum of 10 entries per day.

נסו ערכים שונים של היפר-פרמטרים, נסו אימונים על תתי-קבוצות שונות של פיצ'רים ובדקו מה הם הפיצ'רים הטובים ביותר עבור הסיווג (יש לבדוק על ה- CV לפני שבודקים על ה- test).

הציגו מטריצת עמימות (Confusion Matrix) על ה- CV, הציגו את הסטטיסטיקות (KPI) שמחושבות מהמטריצה והסבירו מה המשמעות של הערכים.

הציגו גרפים של שגיאת (Loss) האימון מול שגיאת הולידציה. ודיוק (Accuracy) האימון מול דיוק הולידציה (אתם יכולים גם להשוות סטטיסטיקות נוספות בין האימון לולידציה, לנתח את ההבדלים ולהסיק מסקנות).

לבסוף, צרו את קובץ ההגשה והגישו אותו לתחרות. צלמו את דף ההגשות שלכם (אם יש לכם יותר מ- 10 הגשות, זה יספיק אם תצלמו רק את 10 ההגשות האחרונות) והדגישו את ההגשה הטובה ביותר שלכם. העלו את תמונת ההגשות למחברת. צלמו את המיקום שלכם ב- Leaderboard והעלו את גם התמונה הזאת למחברת.

סכמו את העבודה שלכם (הסיכום לא חייב להיות ארוך) והסבירו מה עשיתם, מה עבד טוב ומה עבד פחות טוב. הוסיפו בסוף המחברת רשימת מקורות למחברות אחרות שלקחתם השראה מהן ולאתרים או מדריכים שלמדתם מהם.

מבנה המחברת

1. שם, ת.ז., קישור למשתמש ה- Kaggle שלכם.
2. **כל מה ששייך למטלה 1 (חקירת הדאטה וכל האימונים שעשיתה במטלה הראשונה).**
3. ניסויים עם Feature Selection, מודלים שונים, Ensembles ובחירת היפר-פרמטרים (בדיקות יש לעשות על ה- CV).
4. הצגת מטריצת עמימות, סטטיסטיקות, גרפים וניתוח התוצאות.
5. תמונות ההגשות והמיקום ב- Leaderboard.
6. סיכום העבודה, מה עבד ומה לא עבד (אם יש דברים כאלה).
7. רשימת מקורות.

מבנה הציון

לכל אחד מהשלבים במבנה המחברת יש ערך בציון. גם למראה של המחברת יש ערך בציון. חשוב לשמור על מחברת יפה מסודרת וברורה. חשוב לשמור על קוד נקי ברור ופשוט. חשוב מאוד להבין שבסופו של דבר אתם תמיד תעבדו עם עוד אנשים, והם יצטרכו להבין את מה שכתבתם בצורה הקלה והנוחה ביותר. תחשבו שאתם מסבירים את מה שאתם עושים לתלמיד תיכון שמכיר פה ושם את המינוחים אך הוא לא מומחה גדול בנושא. הסבירו במילים פשוטות ושלבו קישורים במידה ואתם

חושבים שההסבר במחברת לא מספיק. זכרו שתמונות וגרפים מובנים הרבה יותר ממילים, וביחד, תמונות גרפים ומילים, אפשר להסביר כמעט כל דבר לכל אחד.

קוד פשוט, מסודר, מוסבר ונקי – 10%

מחברת מסודרת, מוסברת, נקייה ונוחה לקריאה – 10%

השקעה, לימוד עצמי, עשייה מעבר למינימום הנדרש – 10%

מימוש נכון של דרישות הפתרון ומבנה המחברת – 70%

יתכנו בונוסים על התרגיל הנוכחי, שיתקבלו עבור ביצועים שיפתיעו אותנו לטובה (לדוגמה, ניתוח תוצאות מושקע מאוד), כמובן עד ניקוד מקסימלי של 100 בתרגיל הנוכחי.

הערות

בחרו מודל, בצעו Feature Selection, השתמשו בשיטות שונות ליצירת Ensembles (Bagging, Boosting), ובשיטות שונות לחיפוש היפר-פרמטרים (Grid Search, Random Search), הסבירו את בחירותכם (מומלץ לקרוא על זה באינטרנט ולהתנסות בעצמכם).

מומלץ להשתמש בפונקציות קצרות וייעודיות ולא ליצור כפל קוד. מומלץ לתת שמות משמעותיים לפונקציות ולמשתנים ולהסביר לפני כל פונקציה מה היא עושה (אפשר גם בהערה).

חשוב להבין שבהרבה מקרים, אין נכון או לא נכון. ניסוי וטעייה זה חלק בלתי נפרד מלמידת מכונה. תבינו מה אתם עושים ותסבירו את המהלכים שלכם. יש מהלכים שיהיו מוטעים מבחינה מתודולוגית ומהלכים שיהיו נכונים מבחינה מתודולוגית.

חשוב להראות הבנה, לכתוב מה ניסיתם ועבד ומה ניסיתם ולא עבד, ולנסות להסביר את זה.

קישורים נחוצים

אתר קאגל – https://www.kaggle.com/

תחרות הטיטניק – https://www.kaggle.com/c/titanic

הדאטה של תחרות הטיטניק – https://www.kaggle.com/c/titanic/data

הפורום של תחרות הטיטניק – https://www.kaggle.com/c/titanic/discussion

המחברות של תחרות הטיטניק – https://www.kaggle.com/c/titanic/notebooks

אתר להמרת קבצי ipynb לקבצי html – https://htmtopdf.herokuapp.com/ipynbviewer/

ויקיפידה על מטריצת עמימות וסטטיסטיקות שנובעות ממנה – (https://en.wikipedia.org/wiki/Precision_and_recall#Definition_(classification_context

אתר ללימוד Markdown – https://guides.github.com/features/mastering-markdown/

סיכום הדרישות הטכניות

מודל: לבחור אחד או יותר מ- KNN, NBC, LDA

וולידציה: להשתמש ב- CV (לבחור אחד או יותר מ- LPO, KFold).

בחירת פיצ'רים: לבצע Feature Selection.

אנסמבל: לבחור אחד או יותר מ- NFold Bagging, Bootstrap Bagging, Gradient Boosting, AdaBoost)

תוצאות: לבצע ניתוח תוצאות למשימת קלסיפיקציה עם Confusion Matrix ו- KPI

נושאים שנלמדים במטלה זו

KNN

NBC

LDA

Gradient Boosting

AdaBoost

NFold Bagging

Bootstrap Bagging

Grid Search

Random Search

Confusion Matrix

KPI

בהצלחה