למידת מכונה 2020 – מטלה 4

תחרות

התחרות עבור מטלה 2 נקראת **House Prices: Advanced Regression Techniques**.

הקישור לתחרות נמצא כאן: https://www.kaggle.com/c/house-prices-advanced-regression-techniques

בתחרות זו עליכם לבצע משימת חיזוי.

עליכם לחזות מחירי בתים בעיר איימס שבמדינת אייווה (Ames, Iowa) בארצות הברית.

לרשותכם מספר גדול מאוד 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/house-prices-advanced-regression-techniques/data

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

במקרה שיש לכם שאלות בנוגע לתחרות ולדאטה, אתם תמיד יכולים לכתוב ולקרוא שאלות של אחרים בפורום של התחרות: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/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/house-prices-advanced-regression-techniques/notebooks

שימו לב שהניקוד בתחרות מתבצע על ידי RMSE (Root Mean Squared Error):

> **Metric**
>
> Submissions are evaluated on Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price. (Taking logs means that errors in predicting expensive houses and cheap houses will affect the result equally.)

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 חקירת דאטה מכיוון שעשיתם זאת כבר במטלה השנייה.

אתם צריכים לקחת את המחברת שלכם מהמטלה השנייה ולהמשיך אותה עבור המטלה הנוכחית (להוסיף כותרת "Exercise 4" בסוף המחברת הקודמת ולהמשיך תחתיה). בהזדמנות זו אתם יכולים לשפר ולסדר גם את החלק ששייך למטלה השנייה.

עליכם לבצע חיזוי לדאטה על ידי שימוש ב- LWLR (Locally Weighted Linear Regression) או ב- KNN (K Nearest Neighbors) או בעצי החלטה (אחד או יותר) או ב- SVM (Support Vector Machine) (עם קרנל אחד או יותר). הסבירו את בחירתכם ונמקו אותה (אם בחרתם ביותר מאחד, הסבירו מה ההבדלים ונתחו את התוצאות בהתאם).

השתמשו בשיטות אנסמבל, ב- Random Forest, ב- VotingRegressor או ב- StackingRegressor (אחד או יותר). הסבירו את בחירתכם ונמקו אותה (אם בחרתם ביותר מאחד, הסבירו מה ההבדלים ונתחו את התוצאות בהתאם).

השתמשו ב- PCA להורדת מימדים.

שימו לב שאתם יכולים להגיש רשמית עד 5 הגשות ביום:

Submission Limits

You may submit a maximum of 5 entries per day.

נסו ערכים שונים של היפר-פרמטרים, מומלץ לנסות שיטות אנסמבל שונות, מומלץ לנסות מודלים שונים, השתמשו ב- PCA להורדת מימדים.

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 2 (חקירת הדאטה וכל האימונים שעשיתה במטלה השנייה).**
3. ניסויים עם PCA, מודלים שונים, אנסמבלים ובחירת היפר-פרמטרים.
4. הצגת 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

הציגו ונתחו את ההשפעה של PCA על המודלים שבחרתם.

בחרו מודל, בחרו Ensembles שאתם חושבים שיתאימו לדאטה ולמודל והסבירו את בחירתכם (מומלץ לקרוא על זה באינטרנט ולהתנסות בעצמכם).

שימו לב שלמדנו במחברת תרגול 9 להשתמש ב- SVC (Support Vector Classification) ולכן אתם יודעים להשתמש ב- SVR (Support Vector Regression). שימו לב שב- Scikit-learn יש שיטות שונות של SVR וניתן להשוות ביניהן.

שימו לב שלמדנו במחברת תרגול 9 להשתמש VotingClassifier ולכן אתם יודעים להשתמש ב- VotingRegressor. לא למדנו להשתמש ב- StackingRegressor, לכן תצטרכו ללמוד מהדוקומנטציה איך להשתמש בזה.

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 מחירי הבתים – https://www.kaggle.com/c/house-prices-advanced-regression-techniques

הדאטה של תחרות מחירי הבתים – https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

הפורום של תחרות מחירי הבתים – https://www.kaggle.com/c/house-prices-advanced-regression-techniques/discussion

המחברות של תחרות מחירי הבתים – https://www.kaggle.com/c/house-prices-advanced-regression-techniques/notebooks

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

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

סיכום הדרישות הטכניות

מודל: לבחור אחד או יותר מ- Decision Tree, KNN, LWLR, SVM

אנסמבל: לבחור אחד או יותר מ- Random Forest, VotingRegressor, StackingRegressor

פיצ'רים: להשתמש ב- PCA

תוצאות: לבצע ניתוח תוצאות למשימת רגרסיה

נושאים שנלמדים במטלה זו

LWLR

SVM

KNN for Regression

Decision Tree

Random Forest

PCA

Voting Ensemble

Stacking Ensemble

בהצלחה