מעבדה בניתוח והצגת נתונים (094295)

אביב תשפ"א 2021

תרגיל בית 1

Box Office Revenue Prediction

תאריך הגשה: 13/05/2020

מבוא

בתרגיל זה נשתמש בנתונים אודות סרטי קולנוע על מנת לנסות ולחזות את רמת ההכנסה הגלובלית שלהם בקופות.

לרשותכם רשומות של ***6953*** סרטים עם **27** מאפיינים מגוונים: שם, שנה, ז'אנר, שפה, שחקנים, מדד פופולריות ועוד. רשומות הסרטים מחולקות על פני 2 קבצים: ***train.tsv*** ו- ***test.tsv*** עם 5215 ו-1738 רשומות בהתאמה.

משימה

מטרתכם בתרגיל היא לחזות את ערך ה ***revenue*** לכל סרט בקובץ ה ***test.tsv***. ערך ה ***revenue*** הוא ערך מספרי רציף, לכן זוהי משימת רגרסיה.

מה מגישים

1. דו"ח הגשה ב-pdf (הסבר בהמשך)
2. קישור ל GitHub Repository שלכם המכיל לכל הפחות את הקבצים הבאים:
   a. **environment.yml** (פרטים במדריך ה DevOps)
   b. **predict.py**

קובץ הרצה עבור האלגוריתם שלכם (הסבר מטה + קובץ לדוגמה מצורף)

בנוסף, ה Repository שלכם צריך להכיל את כל קבצי הקוד שכתבתם וקבצים נוספים שהשתמשתם בהם בתרגיל. אין צורך שיכיל את קובץ ה ***train.tsv*** ו/או את קובץ ה ***test.tsv***.

סקריפט **predict.py**

סקריפט זה אמור לקבל ב command line ארגומנט בודד שהוא נתיב לקובץ tsv. מובטח כי קובץ זה יהיה זהה במבנה העמודות שלו לקובץ ה ***test.tsv*** (וכולל שורת header). דוגמא לקריאה ל predict.py:

>>> python predict.py test.tsv

בסיום הריצה, הסקריפט צריך לשמור בתיקייה הנוכחית קובץ בשם ***prediction.csv***. קובץ זה מכיל את כל ה ***ids*** מקובץ ה ***test.tsv*** ולצידם ערך ***revenue*** חזוי (ללא שורת header). מבנה קובץ ***prediction.csv*** לדוגמא:

1234	100000
5678	250000

העמודה השמאלית היא עמוד ה ***id*** והעמודה הימנית היא עמודת ה ***revenue***. על הסקריפט לדעת לקבל **כל** קובץ tsv במבנה העמודות של ***test.tsv*** ולחזות את ערכי ה ***revenue*** של הסרטים שבו. שימו לב כי ייתכן ויהיו ערכים חסרים בקובץ שיסופק לסקריפט. מובטח כי עמודת ה ***id*** תהיה ללא ערכים חסרים. אתם יכולים לשנות את הסקריפט כרצונכם כל עוד הוא קורא וכותב קבצים בהתאם להוראות הנ"ל. קובץ ***predict.py*** לדוגמא מצורף לקבצי התרגיל.

דו"ח הגשה

עליכם להגיש דו"ח בן 7-12 עמודים לפי המבנה הבא:

1. Exploratory Data Analysis
    a. Which features are available in the dataset
    b. Feature distribution, comparative analysis between features (with graphs)
    c. Missing data
2. Feature Engineering
    a. Which features you will be using (and why)
    b. Feature transformation (if any)
    c. Handling missing data (if any)
    d. Data enrichment (if any)
3. Prediction
    a. Example of at least 3 different algorithms used, for each:
        i. Hyperparameter selection, regularization
        ii. Training and Validation

הדו"ח חייב להכיל לפחות 4 גרפים או טבלאות. הדו"ח יכול להיות מחברת jupyter (שמורה כ- pdf או html).

חלק רטוב - מדידת ביצועים

חישוב רמת הדיוק של השערוך שלכם ברגרסיה יתבצע ע"י

Root Mean Squared Logarithmic Error (RMSLE)

המוגדרת באופן הבא:

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(ln(y_i+1) - ln(\hat{y}_i+1))^2}$$

כאשר y_i הוא ערך ה ***revenue*** האמיתי, הוא $\hat{y}_i$ הוא ערך ה ***revenue*** אותו חזיתם, $ln()$ הוא הלוגריתם הטבעי הטבעי ו- n הוא מספר הדגימות. מטריקה זו דומה למטריקת **RMSE** פרט לכך שהיא נותנת פחות משקל לשגיאה בשערוך של ערכים קיצוניים (למשל שוברי קופות).

חישוב ה **RMSLE** יתבצע על קובץ ה ***prediction.csv*** שהסקריפט שלכם מייצא, אל מול ה ***revenue*** האמיתי של הסרטים שסופקו לסקריפט בקלט. שימו לב כי ערך ה ***revenue*** הוא מספר אי-שלילי. ע"מ שחישוב ה **RMSLE** יהיה מוגדר, עליכם להבטיח כי ערכי ה ***revenue*** בקובץ ***prediction.csv*** יהיו אי-שליליים.

חלק תחרותי

בנוסף לחלק הרטוב בתרגיל בו תיבחנו על ***test.tsv***, סקריפט ה ***predict.py*** שלכם יופעל כנגד דאטה סט חבוי. כאמור, מובטח שהקלט יהיה זהה במבנה ל ***test.tsv***.

מבנה הציון

- דו"ח מסכם - 70%
- חלק רטוב - 20%
- חלק תחרותי - 10%

הערות לתרגיל

- אתם יכולים להשתמש בכל שיטת רגרסיה שעולה על רוחכם (במשאבי השרת).
- ניתן להשתמש בכל ספריה פייתונית (שניתן להתקין דרך pip).
- קובץ ה **predict.py** אמור לקרוא למודל מאומן שיודע לקבל כל דאטה סט במבנה של ***test.tsv*** - הן לחלק הרגיל והן לחלק התחרותי של התרגיל.

- אין לאמן מודלים בסקריפט **predict.py**. תוכלו לשמור מודל מאומן מראש ב repository שלכם ולקרוא אותו מתוך הסקריפט.
- **אסור להעשיר את הדאטה סט ע"י מקורות מידע חיצוניים**. עליכם לספק את כל קוד האימון למודל שלכם ב repository. במידה ונחשוד שהמודל שאימנתם קיבל דאטה ממקורות חיצוניים, נאמן אותו בעצמנו ונבדוק האם יש הבדל מובהק בביצועים. הגשה שלא תעמוד בהנחיה זו (אי צירוף קוד האימון ו/או שימוש במקור חיצוני) **תקבל ציון 0**.
- הקוד אמור להיות מסוגל לרוץ ב-Azure. יש לעקוב אחרי הנחיות מדריך ה DevOps להקמת סביבה.
- הקוד אמור לרוץ בזמן סביר.

<u>בדיקת התרגיל</u>

בדיקת החלק הרטוב והתחרותי תתבצע באופן אוטומטי, על גבי מכונה זהה למכונה שלכם ב-Azure. אנחנו נבצע:

1. clone ל repository שלכם
2. הקמת סביבה וירטואלית באמצעות קובץ ה environment.yml
3. הרצת קובץ ה predict.py שלכם כנגד קובץ ***test.tsv*** בחלק הרטוב וכנגד קובץ ***tsv*** נוסף בחלק התחרותי, מתוך הסביבה שהוקמה.

ע"מ להבטיח את תקינות השלבים, מומלץ בחום לבצע אותם בעצמכם לפני הגשת התרגיל. כישלון בבדיקת התרגיל שנבע מקונפיגורציה לא תקינה יגרור הורדת נקודות בחלק הרטוב והתחרותי.

בהצלחה!