מטלת מעבדה מס׳ 2

הנחיות:

- הגשה:
- יש להגיש את המטלה עד לתאריך 15.7.18.
- ההגשה תתבצע ע״י קובץ zip המכיל את הקוד שנכתב בהתאם לדרישות. פורמט קובץ ה-zip יהיה ID01_ID02.zip. שם קובץ ההרצה יהיה Lab2.py.
- ההגשה הינה **בזוגות**.
- חבר קבוצה אחד בלבד יעלה את הפתרון לאתר.
- בעיות אישיות בנוגע למועד ההגשה יש להפנות לבודק התרגילים הקורס טרם מועד ההגשה.
- **כל חריגה מנהלים אלו, ללא אישור בכתב מצוות הקורס, מהווה עילה לפסילת המטלה או להפחתת נקודות.**
- **אין להעתיק פתרונות ואין לשתף קוד בין סטודנטים. אין להעתיק קוד מוכן באינטרנט!**
- **לפתרון המטלה יש להשתמש בגרסאות פייטון 2.7 או 3 (יש לציין בהערות בתחילת הקוד את הגרסא).**
- להבהרות, הכוונות או כל עזרה אחרת, ניתן לשאול שאלות בפורום המתאים למטלה זו באתר הקורס.
- בדיקת המטלה תתייחס בין השאר לפרמטרים הבאים: נכונות הקוד, יעילות הקוד וזמני ריצה. יש לבדוק מקרי קצה.

מבוא

במטלה זו עליכם לבצע משימות חיזוי, חלקו באמצעות אלגוריתמי למידת מכונה.

בחלק א' של המטלה תתמקדו בחישוב מאפיינים לכל רשומה, קביעת class label וכן בחלוקה של הנתונים לסט אימון וסט בחינה.

בחלק ב' של המטלה תאמנו מודלים שונים לביצוע חיזויים שונים.

משימות החיזוי הינם:

משימה א': חיזוי שכר לעובד בקבועי זמן שונים (חיזוי שכר שבועי, חיזוי שכר חודשי)

זוהי משימת supervised מכיוון שניתן לחשב לכל משתמש את השכר שלו מהנתונים הקיימים.

המודלים שתאמנו יהיו אישיים (מודל לכל משתמש) וכן מודל לכל קבוע זמן (מודל חיזוי שכר שבועי ומודל חיזוי שכר חודשי)

משימה ב': האם טרנזקציה כלשהי (למשל תשלום עבור מכון כושר) היא חיוב קבוע.

זוהי משימת supervised מכיוון שבחלק א' של המטלה נוסיף label לכל אחד מהטרנזקציות המופיעות בנתונים האם היא חיוב קבוע – לפי ההסבר הבא:

חיוב קבוע מוגדר על ידי חיובים מאותה הקטגוריה (ומאותה החנות) שמתבצעים בתדירות זהה (התדירות דינמית - נניח אחת לחודש מתבצע אותו החיוב, או אחת לשבוע) בסכומים זהים שביניהם אין הפרש של למעלה מ-20$ מהחיוב הממוצע.

לשם הפשטות, החיוב הקבוע חייב להיות באותם הפרשי תאריכים – הפרש של 30 ימים או הפרש של 7 ימים.

(שימו לב – אסור להוסיף מאפיינים של הפרשי ימים למודל החיזוי!)

למשל חיוב עבור טיול, שמתבצע 3 חודשים ברציפות, ב-1 לחודש בסכומים של 60$, 68$, 75$ (כאשר ממוצע החיובים הוא 65$) הוא חיוב קבוע.

בחלק ג' תממשו webServices אשר ישתמשו בחלקים הקודמים ויחזו עבור רשומה ספציפית את המשימות הנדרשות.

ניקוד מלא יינתן לקבוצות עם הדיוקים הגבוהים ביותר בכל אחת מהמשימות.

מבנה הקבצים: רשומה אופיינית בקובץ transactions מאופיינת באופן הבא:

שימו לב- הכנסה מוגדרת בשדה amount, כאשר amount<0 מדובר על הכנסה וכאשר amount>0 מדובר על הוצאה.

```
{
  "_id" : ObjectId("5b047e2a607dfa2bffc3d95b"),
  "category" : [  קטגוריה/ות של הקנייה
    "Travel",
    "Car Service",
    "Ride Share"
  ],
```

"id" : "g89NZrd198TAEyZW3A9KC94kj1zRDbFgoxXBd", מספר ייחודי של הטרנזקציה

"userId" : "GpZRBGA1ZpTlKzqgml7EiM554rA3BVi1LjoaX", מספר המשתמש במערכת

"accountId" : "ZN7pDx637NsvkpeL8vbwHxrrqJxxKatg36jpn", מספר חשבון

"amount" : 6.33, סכום הקניה/הכנסה

"categoryId" : "22006001",

"date" : "2018-05-09", תאריך הקנייה

"location" : { לרוב ריק – מיקום הקניה

"address" : null,

"city" : null,

"lat" : null,

"lon" : null,

"state" : null,

"store_number" : null, לעיתים מספר החנות קיים

"zip" : null

},

"name" : "Uber 072515 SF**POOL**", שם של הטרנזקציה

"paymentMeta" : { לרוב ריק – אופן התשלום

"by_order_of" : null,

"payee" : null,

"payer" : null,

"payment_method" : null,

"payment_processor" : null,

"ppd_id" : null,

"reason" : null,

"reference_number" : null

},

"creditCardTransaction" : false, האם בוצע באמצעות אשראי

"subscription" : None, השדה שצריך לחזות – אין נתונים על שדה זה

}

חלק א' – חישוב מאפיינים (features) עבור הנתונים

בחלק זה עליכם לחשב מאפיינים עבור כל אחד מהרשומות המופיעה בקובץ transactions.

חישוב המאפיינים יהיה רלבנטי לכל משימות החיזוי, כך שאם לרשומה מסוימת יש 10 מאפיינים קיימים, תוסיפו לה x מאפיינים חדשים, ובסה"כ לכל רשומה יהיו x+10 מאפיינים.

בחלק זה יש למלא את השדה subscription לפי החוקיות שתוארה במבוא. ביצוע ה-labeling הזה לכל טרנזקציה, יעזור לכם לאמן מודל לפי המאפיינים שתיצרו.

עליכם להחליט אילו מאפיינים חדשים שתוסיפו לכל רשומה יעזרו לחזות את סכום ההכנסה של המשתמש והאם הטרנזקציה הזו הינה חיוב קבוע.

רעיונות למאפיינים:

במשימת ההכנסה- חשבו סטטיטיקות של הכנסות אחרונות לפי זמנים, למשל:

- בוליאני – הכנסה/הוצאה (לפי משתנה amount)
- סך הכנסות בשבוע/חודש שעבר למשתמש הנ"ל
- כמות הכנסות בשבוע/חודש שעבר למשתמש הנ"ל
- סטטיסטיקות על הכנסות בשבוע/חודש שעבר (מינימום, מקסימום, ממוצע, סטית תקן) למשתמש הנ"ל

במשימת החיוב הקבוע – הוסיפו מאפיינים הקשורים לכמות הוצאות דומות לפי זמנים.

- האם היו הוצאות במחירים זהים בשבוע/חודש שעבר למשתמש הנ"ל (להוציא 2 מאפיינים – אחד לשבוע שעבר ואחד לחודש שעבר)
- האם היו הוצאות בקטגוריות זהות בשבוע/חודש שעבר למשתמש הנ"ל
- כמה אחוז מהשם של הטרנזקציה חזר על עצמו בשבוע/חודש שעבר למשתמש הנ"ל
- סטטיסטיקות על טרנזקציות באותה הקטגוריה בשבוע/חודש שעבר (ממוצע הוצאה, סטית תקן) של המשתמש הנ"ל
- סטטיסטיקות על טרנזקציות שהשם היה דומה (ב-70% לפחות מהמילים בשם הטרנזצקיה) בשבוע/חודש שעבר (ממוצע הוצאה, סטית תקן) של המשתמש הנ"ל

חלק ב' – אימון מודלים אישיים למשימות א' + ב'

בחלק זה עליכם לאמן מודלים מבוססי למידת מכונה לכל אחת מהמשימות.

משימה א' מוגדרת כמשימת regression, לכן עליכם להפעיל מודל רגרסיה כלשהו על מנת לחזות את השכר של המשתמש. (מומלץ להתחיל עם מודל המבוסס על non-linear regression)

יש לאמן 2 מודלי רגרסיה נפרדים: אחד לשכר השבועי (עם המאפיינים של השכר השבועי) ואחד נוסף לחיזוי השכר החודשי (עם המאפיינים של השכר החודשי).

יש לחלק את נתוני האימון והבחינה במשימה א' לפי עיקרון ה-80-20 (80% הנתונים הראשונים מבחינת זמן, 20% הנתונים האחרונים) וכן החלוקה צריכה להיות לכל משתמש מסוים. המשמעות היא שיופעל מודל חיזוי נפרד (ומאותו הסוג) לכל משתמש.

משימה ב' מוגדרת כמשימה classification – האם טרנזקציה היא חיוב קבוע או לא. אם טרנזקציה כלשהי היא הכנסה, יש לסווגה באופן אוטומטי (ללא הפעלת מודל חיזוי) כ"לא". מומלץ למשימה זו להפעיל מודלים פשוטים כגון decision tree. המודל יאומן לפי ה-labeling שנתתם בסעיף א'.

יש לחלק את הנתוני האימון והבחינה במשימה ב' לפי עיקרון ה-80-20 (80% הנתונים הראשונים מבחינת זמן, 20% הנתונים האחרונים) וכן החלוקה צריכה להיות לכל הנתונים (אין צורך במודלים פרסונליים) כיוון שהמאפיינים הוצאו לכל טרנזקציה בצורה פרסונלית. המשמעות היא שיופעל מודל אחד לכל המשתמשים במערכת.

חלק ג' – מימוש webServices עבור טרנזקציה בודדת

בחלק זה עליכם לממש את השירות הבא:

```
requests.post("http://127.0.0.1:5000", data={'trans': transaction})
```

כאשר transaction היא טרנזקציה בודדת בפורמט json כפי שמופיע בקובץ transaction.

בהינתן טרזנקציה כלשהי, עליכם להחזיר את ה-json הבא:

```
{
    "subscription" : true,  זוהי תוצאות החיזוי האם זהו חיוב קבוע (משימה ב')
    "weeklyIncome": 99,  משכורת שיקבל עד לסוף השבוע (משימה א')
    "monthlyIncome": 352,  משכורת שיקבל עד לסוף החודש (משימה א')
    "yearlyIncome": 4050,  משכורת שיקבל עד לסוף השנה (משימה א')
}
```

עליכם להוציא מהטרנזקציה שקיבלתם ב-data את מספר המשתמש ולייצר עבורה את המאפיינים כפי שעשיתם בחלק א'.

לאחר שייצרתם את המאפיינים, אם אימנם עבור המשתמש הנ"ל מודל (במשימה א'), החזירו את תוצאת המודל לפי predict.

אם מספר המשתמש שקיבלתם לא הופיע בעבר בנתוני האימון, החזירו את תוצאת המודל של אחד המשתמשים (שנו את מספר המשתמש כך שזה יהיה משתמש רנדומלי שקיים אצלכם בנתוני האימון- כך תוכלו לייצר עבורו מאפיינים ולספק עבורו חיזוי מתאים).

בהצלחה! ☺