מעבדה בניתוח והצגת נתונים (094295) אביב תשפ"ג 2023

תרגיל בית 1

Early Prediction of Sepsis from Clinical Data

תאריך הגשה: 04/05/2023

מבוא

אלח דם (Sepsis) הוא מצב רפואי מסכן חיים, הנגרם כתוצאה מתגובה חיסונית לזיהום חמור.

בכל שנה כ1.7 מליון אנשים בארה"ב מפתחים אלח דם ומתוכם כ270,000 מתים.

חיזוי מוקדם של אלח דם והתחלת טיפול רפואי בהתאם הם קריטיים ויכולים להציל חיים רבים.

לקריאה נוספת - מאמר 1, מאמר 2, מאמר 3, אלח דם - ויקיפדיה.

הנתונים

את קבצי הנתונים ניתן להוריד דרך הרצת הפקודה הבאה בCommand Line:

```
wget https://technionmail-my.sharepoint.com/:u:/g/personal/ploznik_campus_technion_ac_il/EQc79uRBeO1FqtH6ILFDx78BuuWui3DuRaBtnzTB6Aqxqg?download=1 -O data.tar
```

ולחלץ אותם בעזרת הפקודה הבאה:

```
tar -xvf data.tar
```

תקיית הData מכילה שתי תקיות train וtest.

הנתונים בהם נשתמש נאספו ממטופלים בטיפול נמרץ משני בתי חולים בארה"ב.

עבור כל מטופל יש ברשותנו קובץ Pipe-separated values (PSV) המכיל נתונים דמוגרפיים ורפואיים אודות המטופל (מילון משתנים מצורף בסוף הגליון), כאשר כל שורה מייצגת נתונים שנאספו במשך שעת אחת. השורות ממויינות לפי השעות, כאשר ניתן להתייחס לשורה הראשונה כשעה הראשונה להגעת המטופל לטיפול הנמרץ ולשעה האחרונה כשעה בה המטופל עזב מסיבה כלשהי את הטיפול הנמרץ.

משימה

מטרתכם בתרגיל זה היא לחזות האם מטופל בטיפול נמרץ סובל מאלח דם כ6 שעות לפני שזוהה כסובל מאלח דם, על בסיס נתונים קליניים אודות מצבו הרפואי לאורך זמן.

כל טבלה, המייצגת מטופל, מכילה עמודה בשם SepsisLabel. הערך בעמודה זו הוא 1 אם השורה בה הערך מופיע היא עד כ6 שעות לפני זיהוי אלח הדם ו0 אחרת. במידה והמטופל לא סבל כלל מאלח דם העמודה תכיל אפסים בלבד. **מטופל יתוייג כסובל מאלח דם (על ידכם) אם קיימת שורה בה SepsisLabel=1.**

כמו כן, עליכם לעבד את הטבלאות כך **שהקלט למודל החיזוי לא יכיל שורות המכילות נתונים שלאחר השורה הראשונה בה SepsisLabel=1 ולא יכיל את העמודה SepsisLabel.**

דוגמא

	HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	BaseExcess	HCO3	...	WBC	Fibrinogen	Platelets	Age	Gender	Unit1	Unit2	HospAdmTime	ICULOS	SepsisLabel
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	27.92	0	1	0	-0.02	1	0
1	88.0	100.0	37.1	97.5	58.0	60.0	21.5	NaN	0.0	26.0	...	NaN	NaN	NaN	27.92	0	1	0	-0.02	2	0
2	92.0	100.0	37.1	125.0	89.0	72.0	20.0	NaN	0.0	24.0	...	7.1	NaN	297.0	27.92	0	1	0	-0.02	3	0
3	94.0	100.0	37.1	121.0	89.0	74.0	23.0	NaN	NaN	24.0	...	7.1	NaN	NaN	27.92	0	1	0	-0.02	4	0
4	93.0	100.0	37.1	124.0	93.0	77.0	21.0	NaN	NaN	NaN	...	NaN	NaN	NaN	27.92	0	1	0	-0.02	5	0
5	103.0	100.0	37.4	105.0	84.0	72.0	21.0	NaN	NaN	NaN	...	NaN	NaN	NaN	27.92	0	1	0	-0.02	6	0
6	99.0	100.0	37.7	104.0	85.0	72.0	23.0	NaN	NaN	NaN	...	NaN	NaN	NaN	27.92	0	1	0	-0.02	7	0
7	99.0	100.0	37.8	122.0	97.0	79.0	18.5	NaN	-2.0	22.0	...	6.5	NaN	276.0	27.92	0	1	0	-0.02	8	0
8	104.0	99.0	37.9	110.0	86.0	73.0	17.0	NaN	-2.0	22.0	...	6.5	NaN	NaN	27.92	0	1	0	-0.02	9	0
9	102.0	100.0	38.1	126.0	96.0	78.0	14.0	NaN	NaN	NaN	...	NaN	NaN	NaN	27.92	0	1	0	-0.02	10	0
10	106.0	100.0	38.0	103.0	75.0	58.0	17.0	NaN	-3.0	NaN	...	NaN	NaN	NaN	27.92	0	1	0	-0.02	11	1
11	118.0	100.0	37.9	105.0	80.0	65.0	19.5	NaN	-3.0	NaN	...	NaN	NaN	NaN	27.92	0	1	0	-0.02	12	1
12	122.0	100.0	NaN	99.0	74.0	63.0	22.0	NaN	NaN	20.0	...	NaN	NaN	NaN	27.92	0	1	0	-0.02	13	1
13	125.0	100.0	37.6	93.0	72.0	60.0	21.0	NaN	NaN	20.0	...	NaN	NaN	NaN	27.92	0	1	0	-0.02	14	1
14	118.0	100.0	NaN	110.0	81.0	65.0	12.0	NaN	NaN	NaN	...	NaN	NaN	NaN	27.92	0	1	0	-0.02	15	1
15	120.0	NaN	37.5	86.5	58.0	57.0	13.5	NaN	NaN	NaN	...	NaN	NaN	NaN	27.92	0	1	0	-0.02	16	1
16	112.0	NaN	NaN	105.0	72.0	54.0	13.0	NaN	NaN	NaN	...	NaN	NaN	NaN	27.92	0	1	0	-0.02	17	1
17	114.0	NaN	37.5	90.0	53.0	50.0	13.0	NaN	-5.0	20.0	...	NaN	NaN	NaN	27.92	0	1	0	-0.02	18	1
18	128.0	87.0	NaN	78.0	49.0	38.0	22.0	NaN	-5.0	20.0	...	NaN	NaN	NaN	27.92	0	1	0	-0.02	19	1
19	121.0	91.0	37.4	63.0	49.0	40.0	23.0	NaN	NaN	NaN	...	NaN	NaN	NaN	27.92	0	1	0	-0.02	20	1

הקלט למודל

השורה הראשונה המקבלת את הערך 1 בSepsisLabel מייצגת 6 שעות לפני אבחון המטופל כסובל מאלח דם.

6 שורות קדימה נקבל את השעה בה המטופל אובחן כסובל מאלח דם

בדוגמא זו המטופל זוהה כסובל מאלח דם בשורה 16 ועל כן SepsisLabel קיבל את הערך 1 בשורה 10 (ב6 שעות לפני). נרצה להפעיל את מודל החיזוי על הנתונים שנתנו עד שורה 10 (כולל. לפי עמודת האינדקס משמאל).

מצופה מכם לאמן מודל חיזוי על הקבצים שבתקיית הtrain ותימדדו על סמך ביצועי המודל על הקבצים שבתקיית הtest ועל סט נתונים נוסף שלא יהיה ברשותכם.

דרישות ומדידת הביצועים

המדד המרכזי למדידת ביצועי המודל יהיה **F1 Score** המוגדר באופן הבא:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

אמנם בדו"ח מצופה מכם לציין מדדים נוספים.

על המודל שתגישו לעמוד בF1 Score של לפחות 0.5 על תקיית הtest המצורפת.

מה מגישים

1. דו"ח הגשה ב-pdf (אפשר jupyter notebook מסודרת מאד שהומרה לPDF)
2. קישור ל GitHub Repository שלכם המכיל את כל קבצי הקוד שכתבתם ואת הקבצים הבאים:
   a. **environment.yml** (פרטים במדריך הDevOps)
   b. **predict.py**

דו"ח הגשה

עליכם להגיש דו"ח בן 7-12 עמודים. מבנה הדו"ח גמיש אך מצופה מכם לגעת בנושאים הבאים:

1. Executive summary
2. Exploratory Data Analysis
   a. Describing the features that are available in the dataset
   b. Inspecting the features distribution, comparative analysis between features (with plots and hypothesis testing)
   c. Handling missing data
3. Feature Engineering
   a. Which feature you will be using (and why)
   b. Features transformations
   c. Data enrichment (if any)
4. Prediction
   a. Use at least **3** different algorithms. For each algorithm report:
      i. The algorithm
      ii. Hyperparameter selection, regularization
      iii. Training and validation results
      iv. Post Analysis (e.g., comparison of model performance on different subgroups), model interpretability
5. Summary and Discussion

על הדו"ח להכיל בנוסף לתיאורים המילוליים גם תיאורים ויזואלים כמו גרפים וטבלאות.

סקריפט predict.py

סקריפט זה אמור לקבל ב command line ארגומנט בודד שהוא נתיב לתקיית טבלאות המטופלים.

דוגמא לקריאה ל predict.py:

>>> python predict.py blabla/path/test

בסיום הריצה, הסקריפט צריך לשמור בתיקייה הנוכחית קובץ בשם ***prediction.csv***. קובץ זה יכיל את כל ה ***Ids***, שהם שמות הטבלאות המייצגות את המטופלים שאוחסנו בתקייה, ולצידם ערך **SepsisLabel** חזוי. על השורה הראשונה בקובץ להיות שורת שמות העמודות.
מבנה קובץ ***prediction.csv*** לדוגמא:

id	prediction
patient_1	1
patient_2	0
patient_3	1

העמודה השמאלית היא עמודת ה ***id*** (לשים לב שהid הוא שם הקובץ\טבלה של כל מטופל) והעמודה הימנית היא עמודת הפרדיקציה של ה **SepsisLabel**.

על הסקריפט לדעת לקבל **כל** תקייה שמכילה קבצי PSV עם מבנה עמודות הדומה לקבצי הPSV שמופיעים בדאטא האימון שקיבלתם ולחזות עבור כל קובץ האם הוא מייצג מטופל הסובל מאלח דם (כך שהקלט למודל הוא כפי שמתואר למעלה).
שימו לב כי ייתכן ויהיו ערכים חסרים בקובץ שיסופק לסקריפט. מובטח כי התקייה תכיל אך ורק קבצי PSV וששמותיהם יהיו ייחודיים. אתם יכולים ליצור את הסקריפט כרצונכם כל עוד הוא קורא וכותב קבצים בהתאם להוראות הנ"ל.

חלק תחרותי

בנוסף לעמידה בדרישת המינימום על ביצועי המודל על תקיית הtest המודל שלכם יבחן גם על סט נתונים חבוי בעזרת הסקריפט **predict.py** שתצטרפו. ההצלחה של המודל על סט זה תימדד באופן יחסי לתוצאות שאר משתתפי הקורס.

סיכום ויצירתיות

בתחילת ההגשה יש לכלול תקציר המסכם את העבודה שביצעתם, בפורמט של בולטים.
זו גם ההזדמנות להדגיש את החלקים שנראו לכם ייחודיים ויצירתיים, שסיפקו תובנות מרתקות הקשורות למשימה.

מבנה הציון

- דו"ח מסכם (רמת פירוט, רגורסיות, מתודולוגיה נכונה) - 60%
- ביצועי המודל על הtest - 20%
- חלק תחרותי - 20%
- יצירתיות – בונוס 5 נקודות

הערות לתרגיל

- אתם יכולים להשתמש בכל שיטת קלסיפיקציה שעולה על רוחכם (במגבלת משאבי השרת).
- על הקוד להיות קריא ולהכיל הערות בעת הצורך.
- ניתן להשתמש בכל ספריה פייתונית (שניתן להתקין דרך pip/conda)
- קובץ ה **predict.py** אמור לקרוא למודל מאומן שיודע לקבל כל דאטה סט במבנה תקיות הדאטא המצורפות – 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 בשנית ונבדוק האם יש הבדל מובהק בביצועים. הגשה שלא תעמוד בהנחיה זו (אי צירוף קוד האימון ו/או שימוש במקור חיצוני) **תיפסל.**
- יש להגיש דוח מוקלד.
- יש לתת קרדיט לכל מקור חיצוני שנעשה בו שימוש.
- הגשת התרגיל תתבצע באתר הקורס במודל על ידי אחד השותפים בלבד.
- איחור בהגשה ייקנס ב10%- לכל יום איחור לא מוצדק. על איחורים מוצדקים (כדוגמת שירות מילואים), יש לדווח לפני מועד הגשת העבודה. ניתן יהיה להגיש עבודות באיחור אשר לא יעלה על שבוע.
- הקוד אמור להיות מסוגל לרוץ ב-Azure. יש לעקוב אחרי הנחיות מדריך ה DevOps להקמת סביבה.
- הקוד אמור לרוץ ובזמן סביר.
- עליכם לבדוק שאתם מצליחים לייצר בעצמכם סביבה חדשה על השרת מתוך environment.yml ולהשתמש בה (זה לא מספיק שיצרתם את הקובץ מתוך סביבה שרצה).
- prediction.csv צריך להיות בעל מבנה זהה למבנה שהוצג למעלה.

בדיקת התרגיל

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

1. clone ל repository שהגשתם
2. הקמת סביבה וירטואלית באמצעות קובץ הenvironment.yml
3. הרצת קובץ הpredict.py שלכם כנגד תקיית הtest וכנגד תקייה עם מבנה דומה שתכיל את סט הנתונים החבוי.

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!

טעויות נפוצות

- קובץ environment.yml שמכיל התנגשויות - עליכם לבדוק שאתם מצליחים לייצר בעצמכם סביבה חדשה על השרת מתוך הקובץ ולהשתמש בה (זה לא מספיק שיצרתם את הקובץ מתוך סביבה שרצה).
- קובץ prediction לא תקין - מכיל אינדקסים לא נכונים של הדוגמאות, סדר לא נכון של העמודות וכו'...
- קובץ predict.py שלא רץ בעזרת הפקודה python predict.py data_path.

FAQ

שאלה 1

בדרישות מהסקריפט predict.py כתוב שאנחנו צריכים להחזיר לכל מטופל פרדיקציה של 0 או 1 לגבי הSepsis. האם זה אומר להחזיר 1 רק אם אנחנו חוזים שש שעות לפני? במידה וחזינו לו Sepsis בפחות או יותר משש שעות מזמן גילוי ה-Sepsis זה נחשב אפס?

בהמשך לשאלה הנ"ל, האם בחישוב F1 יעשה שימוש בכל הרשומות של מטופל כלשהו או שכל מטופל מיוצג על ידי רשומה אחת בחישוב?

תשובה

ייתכן וההוראות לא היו ברורות - **יש חיזוי יחיד עבור כל טבלה**.
השורה הראשונה בה הSepsisLabel משתנה ל1 מייצגת 6 שעות לפני זיהוי הספסיס.
האינפוט שלגביו אתם מבצעים את החיזוי הוא **כל השורות ביחד עד לאותו השורה**.

בהתאם לכך, הF1 לא יחושב על בסיס כל רשומה בנפרד אלא על בסיס החיזוי (היחיד) לכל טבלה.

שאלה 2

מה הכוונה בבדיקת השערות בסעיף b1 בעמוד 3? מה בדיוק המבחן ומה הוא אמור לבדוק?

תשובה

כחלק מהאקפלורציה ייתכן כי תטענו כי תופעות מסויימות באות לידי ביטוי בדאטא.
תרצו להשתמש במבחני השערות כדי לבחון האם לתופעות האלה יש מובהקות סטטיסטית.
אתם יכולים להשתמש בכל מבחן שאתם מכירים עם נימוק שמסביר למה המבחן מתאים.

שאלה 3

נשמח לחידוד על המשמעות של סעיף 3.
לא ברור לנו מה החלק של ניתוח התוצאות אמור לכלול.

תשובה

יש להשתמש לפחות בשני מודלים שונים לחיזוי שמוגדר במשימה.
עבור כל מודל יש לתת הסבר על הרכיבים שלו, על האופן שבו הוא נבחר, על תהליך האימון ולנתח את התוצאות (מה שהוא למד וחזה).
לכמת ולבחון את ביצועי המודלים עם מדדים, ויזואליזציות ושיטות שונות, להתייחס לביצועי המודלים על תתי קבוצות שונות שמיוצגות בדאטא, לבחון את המשתנים שהיו משמעותיים עבור המודל, להשוות את איכות התוצאות בין שני המודלים וכו...'

שאלה 4

במהלך ניתוח הנתונים שמנו לב שטווח הערכים של העמודה HospAdmTime מכיל בעיקר ערכים שליליים אך יש גם ערכים חיוביים.
האם זה תקין או שהעמודה צריכה להכיל אך ורק ערכים שליליים?
להבנתנו העמודה מכילה את ההפרש בין הזמן בו המטופל התאשפז בבית חולים לבין הזמן בו הוא עבר למחלקת טיפול נמרץ ולכן אלו ערכים שליליים.

תשובה

מספר השעות נמדד מרגע האשפוז בטיפול הנמרץ לרגע האשפוז הכללי בבית החולים. ייתכן כי האשפוז בטיפול הנמרץ קדם לאשפוז הכללי בבית החולים וייתכן ההפך.
על כן, ייתכנו גם ערכים חיוביים וגם ערכים שליליים.

מילון משתנים:

Vital signs (columns 1-8)

HR	Heart rate (beats per minute)
O2Sat	Pulse oximetry (%)
Temp	Temperature (Deg C)
SBP	Systolic BP (mm Hg)

MAP	Mean arterial pressure (mm Hg)
DBP	Diastolic BP (mm Hg)
Resp	Respiration rate (breaths per minute)
EtCO2	End tidal carbon dioxide (mm Hg)

Laboratory values (columns 9-34)

BaseExcess	Measure of excess bicarbonate (mmol/L)
HCO3	Bicarbonate (mmol/L)
FiO2	Fraction of inspired oxygen (%)
pH	N/A
PaCO2	Partial pressure of carbon dioxide from arterial blood (mm Hg)
SaO2	Oxygen saturation from arterial blood (%)
AST	Aspartate transaminase (IU/L)
BUN	Blood urea nitrogen (mg/dL)
Alkalinephos	Alkaline phosphatase (IU/L)
Calcium	(mg/dL)

Chloride	(mmol/L)
Creatinine	(mg/dL)
Bilirubin_direct	Bilirubin direct (mg/dL)
Glucose	Serum glucose (mg/dL)
Lactate	Lactic acid (mg/dL)
Magnesium	(mmol/dL)
Phosphate	(mg/dL)
Potassium	(mmol/L)
Bilirubin_total	Total bilirubin (mg/dL)
TroponinI	Troponin I (ng/mL)
Hct	Hematocrit (%)
Hgb	Hemoglobin (g/dL)
PTT	partial thromboplastin time (seconds)
WBC	Leukocyte count (count*10^3/μL)
Fibrinogen	(mg/dL)
Platelets	(count*10^3/μL)

Demographics (columns 35-40)

Age	Years (100 for patients 90 or above)
Gender	Female (0) or Male (1)
Unit1	Administrative identifier for ICU unit (MICU)
Unit2	Administrative identifier for ICU unit (SICU)
HospAdmTime	Hours between hospital admit and ICU admit
ICULOS	ICU length-of-stay (hours since ICU admit)

Outcome (column 41)

SepsisLabel	For sepsis patients, SepsisLabel is 1 if $t \geq t_{\text{sepsis}} - 6$ and 0 if $t < t_{\text{sepsis}} - 6$. For non-sepsis patients, SepsisLabel is 0 .