עבודת בית מסכמת - למידה חישובית

שלב 1 – בחירת הנושא והדטה

(30 נקודות)

הדאטה יכול להיות

1. מהעבודה שלכם, מהחיים.
2. מתחרויות בעיקר מהאתר: <https://www.kaggle.com/>
3. מהאינטרנט (מאמרים בתחום שעבדו על דאטה ועוד)

ככל שהדאטה יהיה יותר מאתגר, יותר מעניין, עם יותר השלכות, ועם עבודה שלכם להוציא ממנו את וקטור הפיצ'רים שיכנס לאלגוריתם הלמידה כך הניקוד יהיה גבוה יותר. אתם יכולים לחפש דאטה מעניין גם באתר מסעיף ב' וגם לראות דוגמאות לפרויקטים מעולים שנעשו בקורס cs229 בלינק: <http://cs229.stanford.edu/projects.html>

שלב 2 הרצת אלגוריתם למידה על הדטה שבחרתם

המטרה שלכם היא שימוש בכמה שיותר כלים שנלמדו בקורס. עליכם להראות:

1) השוואה בין לפחות שני אלגוריתמים (יכולים להשתמש גם בכאלה שלא למדנו)

2) כוונון אלגוריתמים בהתאם לדברים שנלמדו בקורס

3) גרפיים שמראים איך השפעה של פרמטרים משפיעים על הלמידה

4) פיטצ'רים (תכונות) חשובים/לא חשובים

5) התמודדות עם overfitting .

אין צורך לכתוב אלגוריתמים! אבל בהחלט להשתמש בפרמטרים לכוונן וכו

עליכם להגיש קובץ zip שמורכב מ:

1) report -דוח עם כחמישה עמודים שמתאר את הבעיה, הדאטה התוצאות והמסקנות שלכם. בראש הדוח שמותיכם ות.ז.

2) הקוד שכתבתם

3) test- קובץ הרצה שנקרא test.py

ניפגש ב 4.3 לכרבע שעה כדי לבדוק מה עשיתם לשאול שאלות ולתת ציון. רק במקרים נדירים אסכים לשינוי היום!

**עבודת בית מסכמת - למידה חישובית**

מגישות: הלל דודיאן 318593720 יהודית פרל 311596852

מטבע הדברים, כסף הוא תחום המעסיק את רוב האנשים בחיי היומיום, אנשים שואפים להרוויח יותר במטרה לחיות בנחות וללא דאגות עלכן בחרנו להתמקד בשאלה מה הרבדים המשפעים על הכנסותיו של האדם.

הקובץ נתונים שאנחנו בחרנו נלקח מאתר UCI ,אתר המשמש ללימוד למידה חישובית . הקובץ מכיל מאגר מידע על אנשים ממדינות שונות. כולל בתוכו 14 מאפיינים ו עמודת תוצאה 32562 שורות. נרצה לבנות מודל שיחזה האם אדם מסוים ירוויח יותר או פחות 50 אלף בשנה.

בחרנו בנושא זה כי נושא שמעניין את רובנו, מכיל מידע רב, שברובו נתונים מורכבים המאפשר ביצוע של פילוחים שונים ומגוונים.

מאפייני העמודות:

1. גיל – מ גיל 17 ומעלה
2. מעמד עבודה :

* עסק פרטי (Private )
* עצמאי ( Self-emp-not-inc)
* עצמאי Self-emp-inc) )
* ממשלה פדרלית( Federal-gov )
* ממשלה מקומית( Local-gov )
* ממשלה מדינית( State-gov )
* ללא תשלום (Without-pay)
* לא עבד מעולם (Never-worked)

workclass: Private, Self-emp-

not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

1. Fnlwgt
2. השכלה –

* בוגר תואר ראשון
* בוגר מכללה (Some-college)
* 11 שנות למוד (11th)
* HS-grad
* מרצה-פרופ' (Prof-school)
* Assoc-acdm
* Assoc-vo
* 9th
* 7th-8th
* תואר שני (Masters)
* 1st-4th
* 10th
* דוקטורט
* 5th-6th
* גן (Preschool)

education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

1. מס' שנות השכלה: מ1 -16
2. מצב משפחתי :

* נשוי (Married-civ-spouse)
* גרוש ( ( Divorced
* רווק (Never-married)
* פרודים (Separated)
* אלמן (Widowed)
* נשוי – בן זוג נעדר (Married-spouse-absent)
* נשוי – בן הזוג נעדר (Married-AF-spouse)

marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

1. עסוק :

* תמיכה טכנית (Tech-suppor)
* תיקון כלי רכב (Craft-repair)
* שירות אחר (Other-service)
* מכירות (Sales)
* ניהול (Exec-managerial)
* פורפ' (Prof-specialty)
* אחראי מנקים (Handlers-cleaners)
* פקח מכונות (Machine-op-inspct)
* אדמו"ר (Adm-clerical)
* דייג (Farming-fishing)
* הובלת מטענים (Transport-moving )
* משרתת בית (Priv-house-serv)
* מגן שרות ( ( Protective-serv
* צבא (Armed-Forces)
* לא ידוע

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

1. מערכת יחסים :

* אישה (Wife)
* יש ילד (Own-child)
* בעל
* לא בתוך משפחה ( Not-in-family)
* מערכת יחסים אחרת ( ( Other-relative
* לא נשוי (Unmarried)

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried

1. גזע –

* לבן ( White)
* אסיה – איסלנד (Asian-Pac-Islander)
* עאמר-הודית-אסקימו (Amer-Indian-Eskimo )
* שחור (Black)
* אחר (Other)

race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

1. מין : זכר, נקבה
2. רווח הון - רציף
3. הון הפסד - רציף
4. שעות בשבוע – רציף
5. מדינת מגורים: ארצות הברית, קמבודיה, אנגליה, פוארטו-ריקו, קנדה, גרמניה, ארצות הברית (גואם-ארה"ב וכו '), הודו, יפן, יוון, דרום, סין, קובה, איראן, הונדורס, הפיליפינים, איטליה , פולין, ג'מייקה, מקסיקו, פורטוגל, אירלנד, צרפת, הרפובליקה הדומיניקנית, לאוס, אקוודור, טייוואן, האיטי, קולומביה, הונגריה, גואטמלה, ניקרגואה, סקוטלנד, תאילנד, יוגוסלביה, אל סלבדור, טרינאדד וטובגו, פרו, הונג , הולנד.
6. עמודת תוצאה (רווח בשנה) - >50K, <=50K.

בחרנו להתמקד ב2 אלגוריתמים :

1. logistic regression
2. decision tree

כיוון שמדובר בהמון שורות הפעלנו את האלגוריתמים בחלוקה למדינות

עבור כל שיטה ביצענו את ערכות השגיאה הבאות :

הגדרנו

**טבלה 1: המתייחסת למעל 50 אלף**

**TP**- הכנסה שווה\מתחת ל50 והמודל צודק

**FP**- הכנסה מעל 50 והמודל טועה

**TN** - הכנסה מעל 50 והמודל צודק

**FN** – הכנסה שווה\ מתחת ל50 והמודל טעה

|  |  |  |
| --- | --- | --- |
| מעל 50  אלף | מתחת\שווה  ל 50 אלף | חזוי אמיתי |
| FN | TP | מתחת\שווה  ל 50 אלף |
| TN | FP | מעל 50  אלף |

**טבלה2: המתייחסת למתחת 50 אלף**

**TP**- הכנסה מעל 50 והמודל צודק

**FP**- הכנסה שווה\ מתחת ל50 והמודל טועה

**TN** - הכנסה שווה\ מתחת ל50 והמודל צודק

**FN** – הכנסה מעל 50 והמודל טעה

|  |  |  |
| --- | --- | --- |
| מתחת\שווה  ל 50 אלף | מעל 50  אלף | חזוי אמיתי |
| FN | TP | מעל 50  אלף |
| TN | FP | מתחת\שווה  ל 50 אלף |

וחשבנו את ממדי השגיאה :

* recall- מבין כל
* precision
* F-score-
* Support – מתוך הy\_test שנשלח כמה באמת נחזו כמו הערך האמיתי ב y\_test

ממדי השגיאה חושבו פעם אחת ביחס לטבלה 1 (סימון 0 בקוד ) ופעם אחת לטבלה 2 (סימון 1 בקוד)

בנוסף חושב (שיטה מובנית):

* micro avg- הממוצע הכולל של TP, FN, FP
* macro avg – ממוצע של ממוצע ללא משקל פר תווית
* weighted avg- ממוצע של support פר תווית.

**אלגוריתם logistic regression**

* נירמול עמודות מתוך 14 מאפינים 8 עמודות דרשו נירמול – נוספו 40 עמודות
* גרף לתיאור אחוז הצלחה בחיזוי ( פונקצית המחיר - L(Thets))
* חשבנו את הערכות השגיאה שלמדנו
* לצייר את ה ROC CURVE ולחשב את AUC
* ולהראות תוצאות שגיאה רק על ה test set.
* ציירו גרף שמתאר את השגיאה על ה train set וה test set לעומת מספר הדוגמאות שלמדתם עליהם (תלמדו על 20 דוגמאות, 40 דוגמאות, 80 דוגמאות, 160 דוגמאות, 210 דוגמאות) וציירו את שני הגרפים ביחד.
* יש להגיש את ערך השגיאה הממוצעת לכל .
* כמו כן להשוות את הערך האופטימלי לערך (שזה בלי רגולריזציה).
* ציירו את הגרף של השגיאה לכל ערך של .
* עבור ה  הכי טוב ציירו את הגרף של ציר ה-x מספר הדוגמאות וציר ה-y השגיאה על קבוצת הלמידה והשגיאה על קבוצת test קבועה (ללמוד על 10, 20, 30, 40, …,190,200).
* להראות הבדל בין לוגסטיק עם קפולד ובלי