המודלים הסטטיסטיים מחולקים לשני סוגים מרכזיים

למידה מפוקחת (Supervised learning) – מודלים ממשפחה זו מחפשים קשר בין כלל המשתנים הבלתי תלויים למשתנה תלוי כלשהו.

למידה לא מפוקחת (Unsupervised learning) – מודלים מסוג זה מנסים לאתר תבניות בנתונים, ואינם מחפשים קשר בין משתנים בלתי תלויים למשתנה תלוי כלשהו.

בעבודה זו אסקור את התוצאות של למידה מפוקחת.

- המבנה הפורמלי של למידת מכונה:

**הקלט** – המידע שיוזן למודל, כולל:

1. תחום המידע (Domain set, Data set) – כל המשתנים הבלתי תלויים באירועי העבר.
2. התוצאה (Label set) – המשתנה התלוי בתוצאות העבר.
3. סט נתוני האימון (Training data) – סט הנתונים שישמש את המודל ללמוד את הקשרים בין תחום המידע לתוצאות.

**הפלט** (Predictor) – תוצאת החיזוי של המודל עבור תצפית חדשה.

**מידת ההצלחה** (Measure of success) – מודל נמדד ביכולתו לחזות סט נתונים חדש שלא נכלל בסט הנתונים שעל בסיסו הוגדר המודל.

בדיקת מידת ההצלחה נעשית באמצעות חלוקת תחום המידע לסט נתוני אימון ולסט נתוני בדיקה, כאשר בסט האימון נעשה שימוש לבחינת המשתנים המספקים בסיס לחיזוי, לעומת סט הבדיקה שבאמצעותו בוחנים את דיוק המודל שהוגדר על סט האימון .

נהוג לחלק את סט הבדיקות ביחס של 20% מסך הרשומות בתחום המידע, והוא תלוי בעיקר באופק החיזוי הנדרש. עם זאת, בעבודה, החלוקה היא ביחס של 30%.

חשוב להבהיר את הנקודות הבאות :

* מודל שמתאים היטב לסט האימון לא בהכרח יספק חיזוי טוב
* ניתן להגיע להתאמה מושלמת באמצעות שימוש במספיק פרמטרים לחיזוי
* התאמת יתר (over-fitting) של מודל היא כישלון להבין את תבניות המידע

**השערה 0** גורסת כי 3 המחלקות (מצבי ספונטני, סינכרוני ולבד) מתפלגים בצורה אחידה

הסתברות

ציון תחום

3

1

0

1

p-value

*איור***:** התפלגות אחידה כהשערה 0

**השערה אלטרנטיבית** - מנסה להפריך את השערה 0.

p-value – עוצמת הראיות להפרכת השערה 0 ומסמלת את ההסתברות שהתצפית מהמדגם תפריך את אמיתות השערה 0.

ההשערה האלטרנטיבית גורסת כי 3 המצבים מתפלגים נורמלית -

9

ציון תחום

ציון תחום

3

1

0

הסתברות

1

p-value

התפלגות נורמלית כהשערה אלטרנטיבית

ככל שערךp-value נמוך יותר, כך נקבל הוכחה שהשערה 0 מופרכת.מודלים לחיזוי משתנים דיכוטומיים

כאשר המשתנה התלוי הוא דיכוטומי, דהיינו בעל ערכים של כן/ לא, החיזוי נעשה באמצעות מודלים מסוג "סיווג" (Classification) שמאתרים תבניות במשתנים שהוזנו במודל ומסווגים את תוצאת המשתנה הב"ת (status.)

תוצאת המודלים הללו היא וקטור של ערכים שהמודל חזה על סט הבדיקות, אותם ניתן להשוות לתוצאות האמתיות שיש בסט האימון. ההשוואה נעשית באמצעות מטריצת שגיאות (Confusion Matrix) ובאמצעות מפת חום בצורה יותר ויזואלית, כמו כן אשתמש בCLASSIFICATIONREPORT שמספקת לנו יכולת להבין תוצאות חיזוי על בסיס המטריצה. הטבלה הבאה מציגה את מבנה מטריצת השגיאות (עבור 2 מחלקות).

בעבודה זו נעשה שימוש במודלים לחיזוי כאשר המשתנים התלויים הרבה תוצאות של הניסוי (כ40) והמשתנה הב"ת הוא הסטטוס.

מטריצת השגיאות של מודלים לחיזוי:

|  |  |  |  |
| --- | --- | --- | --- |
| **תצפיות בפועל** | |  |  |
| SPONTAN | SYNC |  |  |
| False Positive  (שגיאה מסוג 1) | True Positive | SYNC | **תצפיות חזויות** |
| True Negative | False Negative  (שגיאה מסוג 2) | SPONTAN |

### מדדים לחישוב רמת הדיוק של המודלים הדיכוטומיים

קיימים מדדים שונים לחישוב רמת הדיוק של מודלים לחיזוי משתנים דיכוטומיים שמתבססים על מטריצת השגיאות. המדדים הבאים סוקרים את המרכזיים שבהם, ומספקים תוצאה באחוזים.

**- דיוק המודל הוא היחס בין תוצאות החיזוי הנכונות לכלל התחזיות.**

רמת דיוק החיזוי של המודל:

- ההסתברות שהמודל חזה תוצאה חיובית בהינתן שהתוצאה הייתה חיובית.

: True Positive Rate (Recall)

- כמות התוצאות החיוביות ביחס למספר הפעמים שהמודל העריך חיובי*.*

*:* Precision

ציון F (F score) – היחס המשוקלל בין ה- Recall ל- Precision.מדד זה מתאים במיוחד כאשר התצפיות אינן מתפלגות באופן אחיד בין ערכי ה- True Positive ל- True Negative ויש צורך לנרמל את משקל התצפיות הללו בציון הכללי

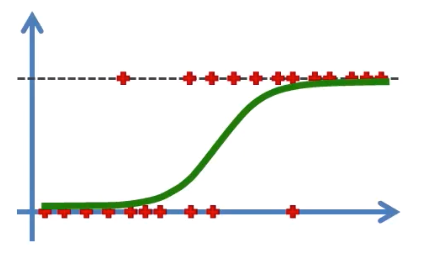
: F score

### **רגרסיה לוגיסטית**

מודל זה ממשפחת "הסיווג" מחשב את ההסתברות שערך המשתנה התלוי יהיה 0 או 1 בהתאם למשתנים הבלתי תלויים באמצעות טרנספורמציית הלוגית (Logit) המוגדרת עפ"י פונקציה

עבור משתנה ב"ת X, כאשר הוא החותך של Y כאשר X=0 ו- כשיפוע של X.

פונקציית הלוגית מסייעת לחזות מקרים מהמבנים הדומים לאיור:



מבנה הרגרסיה לוגיסטית בחיזוי משתנה סיווג

### **עצי החלטה ויער אקראי (Decision Tree & Random Forest)**

עץ החלטה היא שיטה נפוצה לסיווג המתאימה למגוון רחב של בעיות .בהינתן סט תצפיות עם מספר משתנים ב"ת ומשתנה מטרה דיכוטומי, שיטת העץ מחלקת את התצפיות למקטעים עפ"י הרקורסיה הבאה:

המודל בוחר משתנה אקראי כלשהו, נניח וערך אקראי כלשהו , ואז מחלק את התצפיות לשתי הקבוצות ו- , כאשר כל קבוצה תחומה בתוך מרובע גאומטרי.

בשלב זה, המודל בוחר כל אחד מהקבוצות שנבנו עבור המשתנה הראשון ומבצע את אותה פעולה על משתנה אקראי אחר כלשהו, וחוזר חלילה.

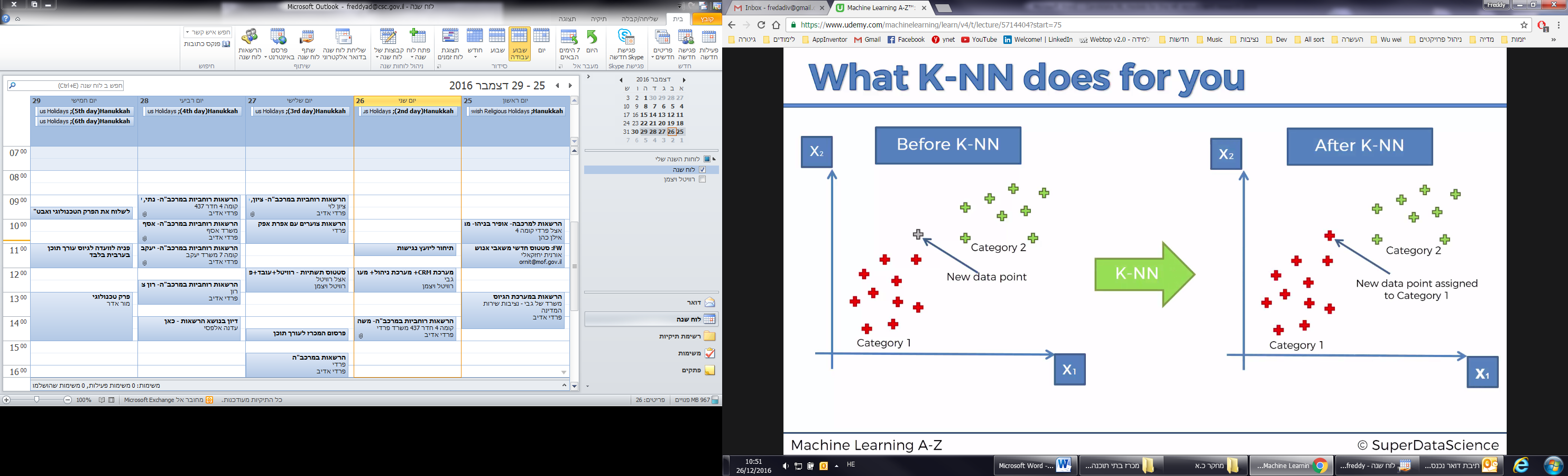
התוצר הסופי של המודל היא חלוקה של כל מרחב התצפיות למרובעים הומוגניים ככל האפשר המייצגים את המשתנים והערכים שנבחנו. האיור מפרט בצורה גרפית את שלבי החישוב של המודל:

שלבים מרכזיים באלגוריתם ה- Random Forest

### **KNN**

מודל זה מסווג קבוצות של ערכי המשתנים התלויים עפ"י ערכי משתנה המטרה, ומשלבים תצפיות חדשות בהתאם לקרבתם לכל אחת מהקבוצות שחושבו

הקלט למודל מכיל, מעבר לסט הנתונים, גם את פרמטר K המגדיר מה המרחק המגדיר שכנים. האיור מציג דוגמא של בעיה הניתנת לפתרון באמצעות מודל ה- KNN.



חיזוי באמצעות מודל KNN

### **רשת נוירונים (Neural Networks)**

במודל זה נעשה שימוש ביחידה אטומית הנקראת נוירון הקולטת שניים או יותר מקורות מידע, מבצעת חישוב מסוים ומוציאה את תוצאתה כקלט לנוירון אחר

רשת הנוירונים היא למעשה סט גדול של נוירונים הבנוי בשכבות, כך שהשכבה הראשונה (Input Layer) קולטת את המשתנים הב"ת ומעבירה לשכבות המרכזיות (Hidden Layers) מחשבת את משקלם של המשתנים הב"ת ביחס לחיבורים של המשתנים הב"ת שהוזנו אליהן, ושכבת הפלט (Output Layer) מוציאה את תוצאות המודל והמשקלות של המשתנים הב"ת.

בתהליך האימון, הנוירונים מחשבים אלגוריתם אלגברי מתוך מגוון אלגוריתמים אפשריים, בהתאם לבעיה העומדת על הפרק, ולבחון את השפעת החישוב על ערך משתנה המטרה, ולשנות את משקלות הערכים שנכנסו כקלט לכל נוירון כדי לבחון את השפעתם על משתנה המטרה.

האיור מציג דוגמא לרשת נוירונים עבור שלושה משתנים ב"ת

|  |
| --- |
|  |

סיכום העבודה:

העבודה התחלקה לשלושה שלבים עקריים:

1. ארגון המידע
2. דימות המידע
3. תחזית עבור משתנה ב"ת (סטטוס)
4. ארגון המידע:

בכדי לארגן את המידע בצורה שיטתית השתמשתי בפנדס עבור תתי מחברות. המטרה בשימוש בכמה מחברות הייתה, שבמצב בו המודלים לא מספקים תוצאות רצויות, אוכל לחקור את טבלאות המידע השונות בכל שלבי הארגון. ישנם 4 שלבי ארגון מידע, לעיתים שימוש בEXCEL ולרוב שימוש בPANDAS.

במחברת הסופית אותה אציג, מכל frame נמחקו 7 השניות הראשונות, ונלקחו 5 דגימות בשניה.

1. דימות המידע: (השתמשתי במידע של שבע השניות הראשונות אותן מחקתי בכדי לברר עד כמה 7 השניות הראשונות אינן רלוונטיות, וכדי לא לפגום במידע איתו אעבוד עבור התחזיות, לציין שהמידע ברור יותר עבור מעט שניות.)

דימות דו מימדי: השתמשתי בפונקציה בכדי להציג ויזואלית מידע על גבי גרף בקלות. בעבודה, הדגמתי את מה שהיא עושה עבור תתי הטבלאות (של שבע השניות הראשונות) בכדי שהתוצאות תהיינה ברורות יותר.

דימות תלת מימדי: נעשה בשביל גיוון והצגת יכולות של matplotlib .

1. תחזיות: חלק זה מחולק לשלושה חלקים. במהלך כל שלב הפרדקציה, העלתי כמה השערות והבאתי אותן לידי ביטוי במחברת. המטרה הייתה שיפור אחוזי הדיוק. החלקים הנ"ל בצורה הבאה: א. הרצת מודלים נבחרים. ב. הרצת מודלים נבחרים לאחר PCA. ג. הרצת מודלים נבחרים עבור 7 השניות הראשונות (שמחקתי מהפריים המקורי) בכדי למצוא האם ניתן לשפר את אחוזי הדיוק עבור מצב ספונטני וסינכרוני באמצעותם.

**סיכום תחזית:**

**ככלל, קל מאוד לזהות מצב ALONE. כלל המודלים מציגים תוצאות טובות מאוד בזיהוי. האתגר היה להפריד בין מצב SPONTAN לבין SYNC כפי שניתן לראות ויזואלית במפת החום. להפתעתי, מודלים מודרניים עליהם למדנו הסמסטר, לא הציגו רמות דיוק גבוהות יותר בהכרח מאשר המודלים הבסיסים.**

**עבור א:**

**Knn לא מציג תוצאות טובות.**

**Naïve bayas, Logistic reg**

**מציגים תוצאות טובות מאוד. 86 אחוזי דיוק עבור שלושת המצבים. (חישוב ממוצע)**

**עבור כל שאר המודלים, התוצאות הן טובות, משתנה עבור סוגי שגיאה וכו'. המודלים הם עץ החלטות, אדאבוסט, יער, רשת נוירונים.**

**עבור ב: לאחר PCA**

**Logistic reg מציג תוצאות טובות מאוד 88 אחוזי דיוק.**

**כל השאר שנבחרו, תוצאות טובות.**

**עבור ג: פרידקציה עבור 7 שניות ראשונות וללא מצב ALONE.**

**Logistic reg מציג תוצאות טובות של 82 אחוז.**

**לבסוף, כדאי לבחור את LOGISTIC REG עבור תוצאות טובות באופן כללי. עם זאת, בוודאי שיש מודלים שמתפקדים טוב יותר, עבור ערכים מסויימים, למשל, עבור TYPE 2 נכון להציג את NAÏVE BAYAS .**

**Apple-DF**

העבודה התחלקה גם כאן לשלושה שלבים + חצי, כאשר מטרתי האישית הייתה לשפר טכניקת קוד, והכרות מעמיקה יותר עם SEABORN ו SKLEARN

מטרתי הייתה לבחון את משתנה הרייטינג (דירוג כוכבים) לשם כך בחרתי שלושה DF, אפל, גוגל-סטור, טריפאדוויסר.

מהר מאוד התמקדתי באפל-סטור, מתוך רצון לערוך ולדיציה עם גוגל-סטור (ישאר לקיץ). מטרתי הייתה חיזוק הידע בספריות שנלמדו.

הצגתי פונק' לדימות הנתונים, תומכת במספר סוגי גרפים, ניתן להוסיף עוד. הצגתי לאורך העבודה נתונים סטטיסטים מעניינים בנוגע לדירוג הכוכבים ודירוג באופן כללי.

לגבי חיזוי, התוועדתי לספריית SKLERAN בצורה טובה מאוד, על אף שהתוצאות לא טובות, (ככל הנראה הפריים קטן מדי, וייתכן באמת שלא ניתן לחזות-אך עדיין לא אמרתי נואש) אני מרוצה מאוד מהקוד, ובייחוד מתוספת הידע שמחברת זו סיפקה לי.