האקתון IML | משימת Waze | אלון ויזנר, ליאור גוברין, אראל דבל, שירה רבינוביץ'

3 ביוני 2022

משימה I

דאטה

מידול הדאטה

האתגר המרכזי בדאטה היה מידול שלו כפיצ'רים ולייבלים. כל שורה כשלעצמה אינה בעלת משמעות עבור מודל למידה כלשהו.

בחרנו לאגד את האירועים בחמישיות, לפי סדר התרחושתם, כאשר בכל חמישיית אירועים, ארבעה מהם משמשים כפיצ'רים, והמאפיינים (מיקום, סוג, תת-סוג) של המאורע החמישי הם הלייבל אותו אנו מנסים לחזות.

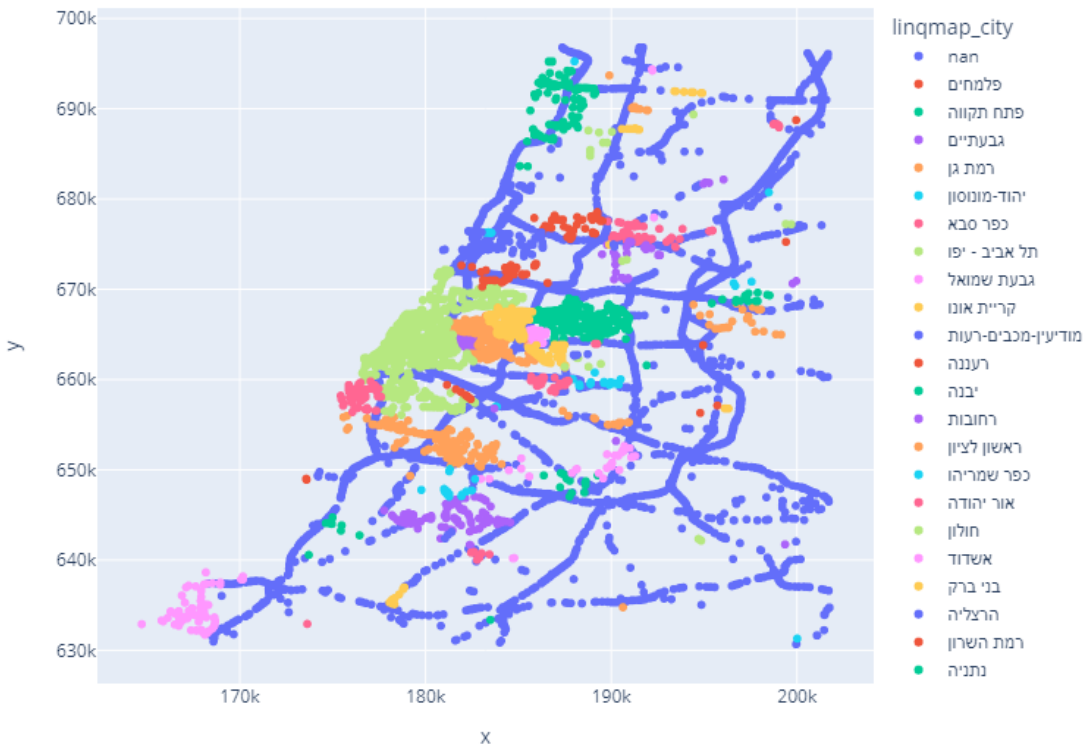
אולם, באופן זה כמות הדאטה ירדה משמעותית, מכ-$18,000$ רשומות שמתוכם $4,000$ בתל-אביב, ל-$3,600$ רשומות שמתוכן כ-900 בתל-אביב. נזהרנו לא להשתמש באותן רשומות גם עבור פיצ'רים וגם עבור לייבלים, ולכן נאלצנו לצמצם את כמות הדאטה.

כמות הדאטה

כמות הדאטה הרלוונטית למשימה הראשונה הייתה קטנה יחסית. כדי לפצות על כך, השתמשנו בדאטה מערים אחרות כדי ללמוד על תל-אביב, השתמשנו במודלים שנעזרים בשיטות כמו Bagging כדי לפצות על כמות הדאטה.

דאטה מערים אחרות

כדי לפצות על כמות הדאטה הנמוכה מתל-אביב, בחרנו ללמוד גם אירועים מערים אחרות, שלהערכתנו היו דומות מספיק בגודל ובאופי התחבורה לתל-אביב.



איור 1: מפת אירועים באזור המרכז לפי ערים

כפי שניתן לראות במפה, בערים רבות נוספות קיימת תבנית צפופה יחסית של אירועים – פתח תקווה, גבעתיים, גבעת שמואל וכו'. בחרנו ערים בהם היו לכל הפחות 100 אירועים, ולקחנו מהם דאטה לאימון. כדי לפצות על העובדה שאלו אירועים מערים שלא יופיעו בטסט,

הוספנו פיצ'ר בשם is_TLV, המציין אם האירוע התרחש בתל-אביב. אירועים שהתרחשו בכבישים בין-עירוניים לא נכללו בסט האימון.

	1	76	4	31	71	60	36	67	17	68	18	42	72	55	21	29	10	20	28
cities	nan	תל אביב - יפו	אור יהודה	חולון	רמת גן	פתח תקווה	יהוד-מונוסון	ראשון לציון	בני ברק	רחובות	בת ים	כפר סבא	רמת השרון	נתניה	גבעתיים	הרצליה	אשדוד	גבעת שמואל	הוד השרון
counts	7850	4502	950	527	501	445	376	351	330	311	284	216	210	207	160	151	131	122	104

עיבוד הדאטה

לאחר ניקוי הדאטה, בבואנו לייצר את הדגימות, חילקנו את השורות לחמישיות: רביעייה ל-X ו-שורה אחת ל-y. כשייצרנו את הפיצ'רים חילקנו אותם ל-3 קטגוריות:

- פיצ'רים לפי שורה - פיצ'רים שמסתמכים על הדאטה של שורה אחת מהדאטה המקורי כמו קואורדינטות, סינוס וקוסינוס של הכיוון, one-hot של ה-type וכו'.
- פיצ'רים לפי רביעייה - פיצ'רים של כל הרביעייה כמו תוחלת המיקומים.
- פיצ'רים של שורה ביחס לרביעייה - פיצ'רים שמאפיינים שורה אחת אך תלויים בשורות האחרות ברביעייה כמו חלוקה לרחובות משמעותיים ברביעייה, וסיווג כל פיצ'ר לרחוב.

אחד האתגרים המרכזיים בשלב זה היה עיבוד של הפיצ'רים המרחביים – מיקום, זווית. מטרתנו הייתה לעבד אותם באופן שיקל על הלומד להבין את הקשר ביניהם, כפי שאנו תופסים אותו.

למשל, בפיצ'ר של הזווית – מובן לנו שאירועים שהתרחשו, אחד באזימוט של $359°$ עם הקוטב הצפוני, והשני עם זווית של $1°$, דומים זה לזה הרבה יותר מאירועים שהתרחשו בזוויות של $10°$ ו-$50°$.

לכן, החלפנו את פיצ'ר הזווית (linqmap_magvar) בסינוס וקוסינוס שלו, שמקבלים ערכים רציפים סביב המעבר מ-$0°$ ל-$360°$ ועם זאת תופסים את המורכבות של זווית במרחב.

דבר נוסף שעניין אותנו ברצף של ארבע אירועים, הוא עד כמה הם קשורים אחד לשני, להערכתנו. למשל, אירועים שהתרחשו במרחקים גדולים זה מזה, או בטווח זמן גדול יותר, כנראה יהיו פחות קשורים.

כיוון שברצוננו למצוא קשר בין מאורעות, ניסינו לתת משקל נמוך יותר למאורעות שאינם קשורים להערכתנו למאורעות שנמצאים איתם בשורה.

פיצ'ר לדוגמא שבא לתאר זאת, הוא שלכל מאורע, בדקנו מהו המרחק שלו מנקודת המרכז של המאורעות, ולאחר מכן בדקנו לכל מאורע, כמה הוא שונה משאר המאורעות, מבחינת מרחק מנק' המרכז (באמצעות zscore).

באופן זה, אנו מצפים שהלומד יידע לתת פחות משקלים לאירועים בודדים. בנוסף רצינו לבחון האם יש מספר של דיווחים באותה דגימה שקרו באותו רחוב, דבר שיכול להעיד על שייכות של הדיווחים המדוברים לאותו מאורע.

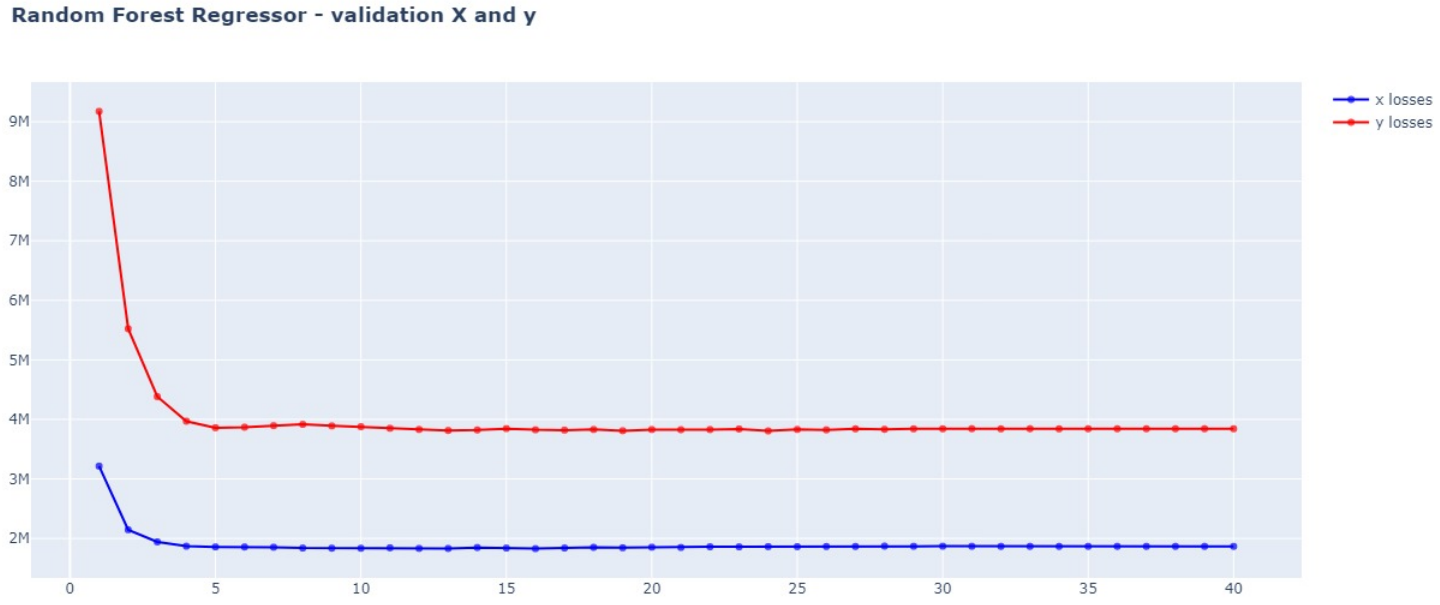
מערכת הלומדים

בחירת הלומדים

לשם תחזיות הלייבלים (linqmap_type,linqmap_subtype,x,y), השתמשנו ב-4 לומדים שונים, כל אחד עבור כל לייבל בנפרד (גילינו שחיזוי של כל קורדינטה בנפרד נותנת חיזוי מדוייק מחיזוי משותף של הקורדינטות). חילקנו את הדאטה ל-train ו-dev, והתנסינו עם לומדים שונים.

לתחזיות הרגרסיה של המיקום (x,y), בדקנו לומדים כמו Linear Regression ו-Decision Tree Regressor; Lasso; Ridge; SGDRegressor. עבור אלו שעניינו אותנו, בדקנו את הביצועים שלהם עם פרמטרים שונים, וניסנו לכייל את ההיפר-פרמטרים.

המודל RandomForestRegressor הראה ביצועים טובים באופן יחסי, ולכן בחרנו לחקור את הפרמטרים שלו.



איור 2: Loss of Random Forest Regressor as function of num. learners, on validation set

עבור הלומדים שבחרנו למשימת הקלסיפיקציה ביצענו תהליך דומה. הסתכלנו בעיקר על וריאציות של Decision Tree, כיוון שלשיטתנו יש למודל זה את הסיכוי הטוב ביותר ללמוד פיצ'רים גאוגרפיים, כי במרחב רב–ממדי, נקודות שקרובות במציאות יהיו גם קרובות מבחינת מודל העץ.

בחנו וריאציות כמו Random Forest, Bagging Classifier ובדקנו את הפרמטרים שלהם. להלן גרף המייצג את ביצועי המודל Extra Tress Classifier על ה-Validation Set:

איור 3: Loss of Extra Trees Classifier as function of num. learners, on validation set

כשלים במודל

במהלך החיזוי של המיקום, נתקלנו בקשיים להוריד את הסטייה הריבועית של הקורדינטות, ולאחר שהצגנו על גרף את נק' האמת אל מול נק' התחזית גילינו כמה תובנות מעניינות על סוגי הכשלים במודל שלנו.

מראש בגלל המבנה המוארך של תל אביב, השגיאה בציר הy הייתה גדולה משמעותית מהשגיאה בציר ה-x. אך ניתן לראות בגרף הבא שיש נטייה של נק' החיזוי להתכווץ אל מול נק האמת. רואים בבירור שהקיבוץ של הנק יחסית נכון, אך בחיזוי גם בציר האופקי וגם בציר האנכי יש כיווץ של נק' התחזית אל מרכז הקבוצה.

בתמונות למעלה ניתן לראות (משמאל לימין): בכחול ציר ה-x הוא של הדגימות החזויות ו-y הוא האמיתי, באדום הדגימות האמיתיות, ובירוק - ציר ה-x הוא של הלייבלים האמיתיים ו-y הוא של החיזוי. אפשר לראות שתופעת ההתכווצות קיימת בשני הצירים.

גרפים אלה מראים בבירור את סוג השגיאה, ואת הנטייה של המודל להתכווץ. היכולת שלנו לאפיין את השגיאה היא קריטית ועוזרת מאד.

הסיבה הראשונה היא כמובן ההבנה של איכות התחית והיכולת לתווך ללקוח את הכשלים של המודל. אך הסיבה השנייה חשובה בהרבה, אנו מעריכים שמכיוון שכבר השקענו זמן בהבנה וניתוח של הכשלים של המודל, ייתאפשר בעתיד, בהינתן עוד זמן, לחקור יותר את הדאטה, ולנסות לנטרל את אותו גורם שמושך את התחזיות לכיוון מרכז ההתרחשות. כבר בזמן העבודה המועט שהיה לנו, הצלחנו עקב הבנה זו לבצע מספר התאמות (הגדלה של המשתנה הגלובלי שמחליט על הגודל המרחק שייגרום להוצאה של דיווח מחישוב המרכז) שהראו שיפור בחיזוי של המודל.

יתר על כן, ההבנה שתופעות המירכוז קוראת גם בציר הy וגם בציר הx מאפשרת לחדד את המחקר העיתידי על שיפור החיזוי של הדאטה.

משימה II

בבואנו לבצע את המשימה השנייה, התחלנו בניתוח כמות הדאטה שברשותנו. מצאנו שברשותנו המידע הבא. שורת המספרים למעלה מייצגת את התאריך בחודש מאי, כלומר התאריכים הם 15.5 עד 24.5:

linqmap_type	ACCIDENT					JAM					ROAD_CLOSED					WEATHERHAZARD				
day	15	16	17	18	24	15	16	17	18	24	15	16	17	18	24	15	16	17	18	24
0-2	0	0	0	4	0	0	0	0	23	0	0	0	0	523	0	0	0	0	115	0
2-4	0	0	0	4	0	0	0	0	28	0	0	0	0	1060	0	0	0	0	220	0
4-6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6-8	0	0	0	5	0	0	0	0	963	0	0	0	0	288	0	0	0	0	206	0
8-10	13	0	11	0	15	301	0	946	659	877	139	0	324	148	330	103	0	181	76	170
10-12	3	14	7	0	10	86	274	266	0	390	139	330	336	0	337	91	173	173	0	176
12-14	7	0	11	0	11	99	0	235	0	194	139	0	344	0	169	96	0	165	0	77
14-16	10	0	20	0	0	336	0	755	0	0	137	0	346	0	0	105	0	170	0	0
16-18	10	0	25	0	0	320	0	1057	0	0	134	0	335	0	0	98	0	200	0	0
18-20	0	0	7	0	0	0	0	428	0	0	0	0	313	0	0	0	0	151	0	0
20-22	0	0	3	0	0	0	0	78	0	0	0	0	329	0	0	0	0	137	0	0
22-24	0	0	2	0	0	0	0	72	0	0	0	0	485	0	0	0	0	140	0	0

איור 4: כמות אירועים לפי סוג, יום וטווח שעות

שמנו לב שהרבה מאוד מטווחי השעות חסרים. למשל, עבור טווח השעות 18-20, קיים רק יום אחד בדאטה (ה-17.5, יום שלישי) עם מידע לגבי אותן שעות.

בנוסף, כמות הימים השונים בשבוע לגביהם יש לנו מידע היא חלקית (ראשון, שני, שלישי ורביעי), וגם בשעות בהם יש מידע יותר מלא, יש לכל היותר $3-4$ ימים שונים מהם אנו יכולים ללמוד.

לכן, עקב מחסור בדאטה, החלטנו לחזות את הממוצע: לכל יום חול, חזינו את ממוצע האירועים ביום חול, לפי המידע הידוע לנו. בנוסף, יום ראשון הקרוב, ה-05.06 הוא חג השבועות, שבו אנו מצפים שכמות האירועים בכבישים וסוגים יהיו שונים מביום חול.

לכן, כפלנו את ממוצעי האירועים שראינו בשבוע הקודם בקבועים כלשהם שקבענו מראש, קבועים אלו ניסנו לחזות ע"י למידה על התוכן (היסטוריית תנועה בערים בארץ בשבתות בחגים) ממקורות מידע שונים באינטרנט, בשילוב עם החישבה שלנו (למשל אנו מעריכים שבשבתות וחגים יש פחות עומסים בשעות הבוקר, אך בשעות הצהריים סביר שאנשים כבר יוצאים מהבתים בוודאי בשבתות קיץ ולכן אמנם תהיה פחות תנועה אך ביחס נמוך יותר). בנוסף בדקנו את שעת יציאת החג, לוודא שהיא לא מתנגשת עם אחד הסלוטים ולא תייצר יחסית הרבה תנועה.