מבוא ללמידת מכונה | האקתון 2022

רועי קלנר 316480193 | רני טוחי 212359905
דניאל אזולאי 311119895 | אלון לוי 313163958
2 ביוני 2022

ניתוח הבעיות ועיבוד מוקדם

לאחר בחינה של הפיצ'רים השונים, החלטנו על תת-קבוצה של פיצ'רים משמעותיים שאליהם התייחסנו באימון, תוך ויתור על פיצ'רים שהערכנו כלא רלוונטים / עם הרבה תאים ריקים, לדוגמא nearby, expected_end_date וכו'. עוד עיבודים שנעשו:

- עבור למידת הקואורדינטות השתמשנו רק בדגימות מתל אביב-יפו.
- לשם הנוחות שינינו את שמות העמודות מהצורה linqmap_<col> לצורה <col>.
- אופציה להמיר את התאריכים מפורמט epoch ל-datetime.
- הוספת עמודה של timeslot עבור המשימה השנייה.
- נירמול ערכי עמודות הזמן / קואורדינטות.
- השלמת ערכים חסרים בקובץ הטסט לפי ערכים נפוצים.

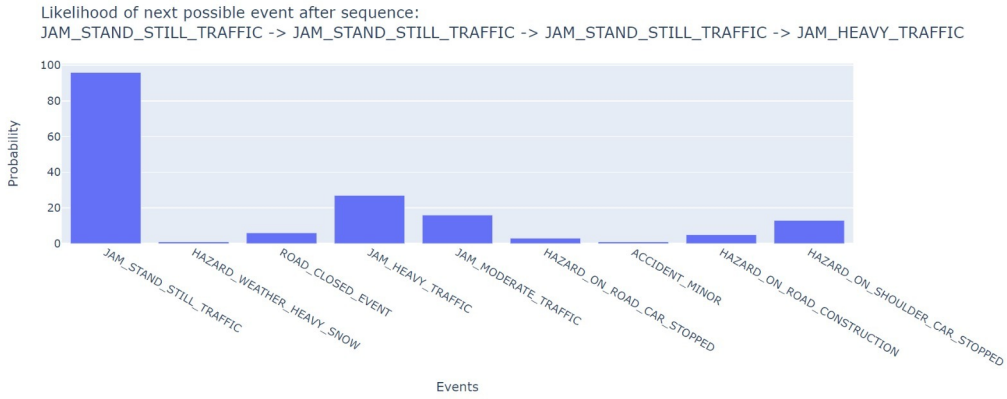
משימה ראשונה - חיזוי אירוע על-פי 4 אירועים קודמים

חילקנו את המשימה לשני חלקים:

- חיזוי ה-Type, Subtype - קלאסיפיקציה
- חיזוי קואורדינטות x,y - רגרסיה

חיזוי Type, Subtype

על מנת לחזות את האירוע הבא בהינתן סדרה נתונה של ארבעה אירועים, יצרנו מודל שבהינתן קבוצת דגימות, מייצר לכל רצף של ארבעה אירועים רשימת הסתברויות עבור כל אירוע שעלול לבוא אחרי רצף כזה. המודל מייצר את רשימה זו על ידי מעבר על כל הדגימות והתבוננות בכל רצף של ארבעה אירועים. לכל רצף כזה, המודל סופר כמה פעמים כל אירוע נוסף הופיע אחרי רצף זה. על מנת לחזות את האירוע הבא בהינתן רצף של ארבעה אירועים, המודל שולף מרשימת ההסתברויות של הרצף שהתקבל את האירוע בעל ההסתברות הגבוהה ביותר. דוגמה להסתברויות של רצף אירועים:

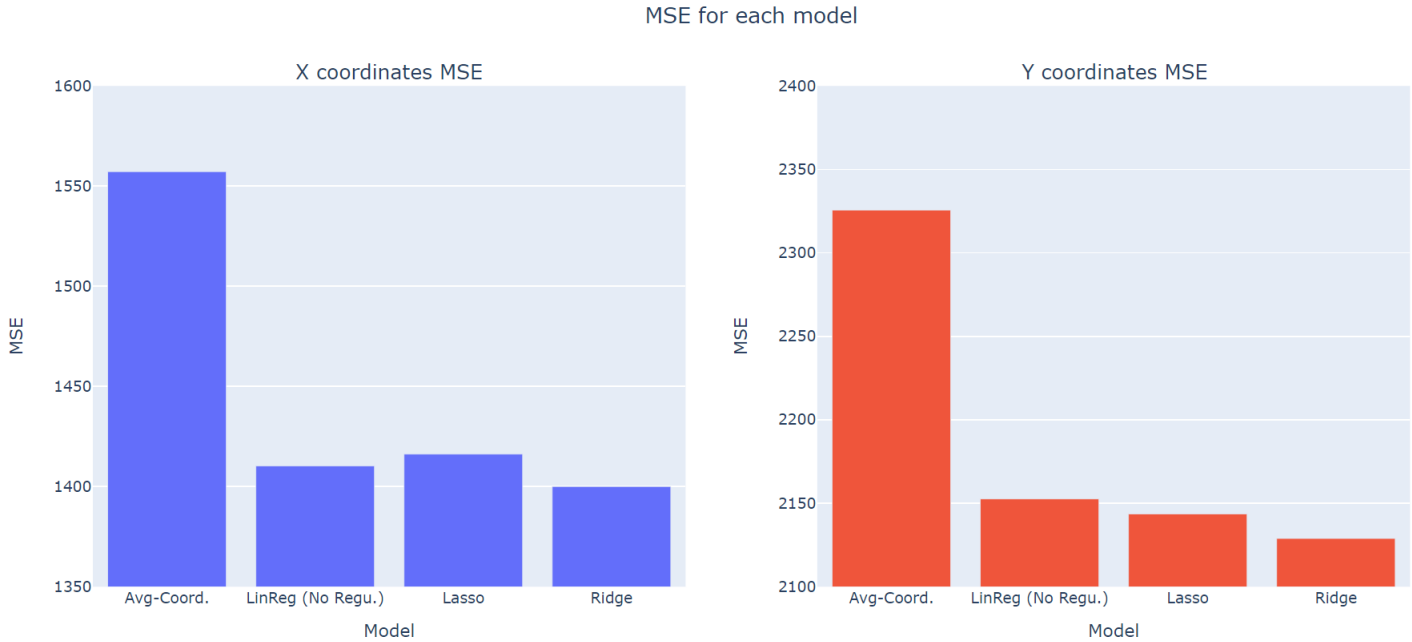


חיזוי קואורדינטות

במבט ראשון הבנו כי זו בעיה שונה מבעיות רגרסיה אחרות שנתקלנו בהן בקורס. זאת משום שעד כה לא התעסקנו בניבוי על פי רצף (sequence) של מאורעות/דגימות. על מנת להתמודד עם זה, יצרנו פסאודו-דגימות, שכל אחת מורכבת מרביעיה של דגימות עוקבות בדאטה (על פי ה-update time).

הבייסליין שאיתו התחלנו היה לקחת את ממוצע ארבע הקואודנטות עבור כל חיזוי. לאחר מכן עברנו לנסות מודלים שונים של רגרסיה לינארית.

על-מנת להשוות ביניהם בצורה הוגנת, פיצלנו את הדאטה ל-train ו-test (ביחס של 1:5). תחילה ביצענו רגרסיה ללא רגולריזציה, וראינו שהתוצאות אכן משתפרות ביחס לניבוי על-פי ממוצע. על מנת למנוע התאמת-יתר השווינו בין רגולריזציה עם Ridge, Lasso, וערכים שונים לפרמטר הרגולריזציה. לשם בחירת הפרמטר תוך חיסכון בדאטה, השתמשנו ב-10-Cross-Validation. לבסוף מצאנו כי הרגולריזציה שמזערה את ה-MSE (על סט המבחן שיצרנו) היא Ridge עם פרמטר רגו' 197, ולכן זה המודל שבחרנו להשתמש בו לניבוי הטסט.



משימה שניה - חיזוי התפלגות אירועים

עבור המשימה השנייה, יצרנו מהדגימות מטריצת פיצ'רים על ידי חילוץ היום בשבוע בו הדגימה התעדכנה, ה-timeslot של העדכון, וסוג האירוע שקרה בדגימה. לכל שורה במטריצה החדשה, בנינו response שמתאר את מספר הפעמים שהאירוע המתואר קרה באותו היום ב-timeslot התואם. לאחר מכן, אימנו מודל רגרסיית Ridge על המטריצה ועמודת התוצאות שיצרנו. בשביל לחזות את המידע הנדרש עבור יום מסוים, המרנו את התאריך ליום בשבוע והעברנו אותו לפונקציית הפרדיקציה. פונקציית הפרדיקציה שלנו קיבלה יום בשבוע (ערך בין 0 ל-6), כך שלכל timeslot אפשרי (לדוגמה, 12-10), ולכל אירוע אפשרי (לדוגמה CLOSED_ROAD), יצרנו וקטור דגימה וקיבלנו תחזית מהמודל עבור היום, סוג האירוע, וה-timeslot. כך מילאנו 12 ערכים בטבלה הרצויה ליום הנתון.