אלגוריתמים לגילוי חריגות

במסמך זה נגדיר את האלגוריתמים לגילוי חריגות שבאים עם האפליקציה. אלגוריתמים אלו צריכים להיות ממומשים כפלאג-אין לאפליקציה (ראו נספח כיצד יוצרים פלאג-אין בג'אווה). המשתמש יוכל באמצעות האפליקציה לטעון את האלגוריתמים הבאים או אחרים שימומשו בעתיד.

אלגוריתם ראשון – Z Score

אלגוריתם זה הוא אלגוריתם Univariate כלומר אלגוריתם המחפש חריגות על כל משתנה ביחס לעצמו בלבד, ולא ביחס למשתנים אחרים.

z-Score:

ציון Z מוגדר כמרחק בין נקודה x לבין אוסף של נקודות המיוצגות ע"י הממוצע שלהן $\bar{x}$ ביחידות של סטיות תקן, כלומר:

$$z(x, X) := \frac{|x - \bar{x}|}{\sigma_X}$$

האלגוריתם פועל בצורה מאד פשוטה. יש לו שני שלבים - למידה, וגילוי.

בשלב הלמידה האלגוריתם מקבל time-series data ומחזיר עבור כל עמודה X בקלט ערך מספרי המייצג סף t_X. הסף מחושב כך:

$$t_X \leftarrow \arg\max_x z(x, X)$$

כלומר הסף לעמודה X מחושב כ z-Score המקסימלי שנתגלה ע"י השוואת כל ערך $x \in X$ לשאר הערכים ב X מזמן 0 ועד לזמן הופעתו של x (לא כולל x). ההיגיון הוא שבקלט תקין זה ערך ה z-Score הגבוה ביותר שמותר לנו לראות ועדיין ייחשב כתקין.

בשלב הגילוי האלגוריתם מקבל time-series data שצריך למצוא בה חריגות. לכן האלגוריתם מודד לכל ערך $x \in X$ ה z-Score שלו מול כל הערכים ב X מזמן 0 ועד לזמן הופעתו של x (לא כולל x). אם ערך ה z-Score גבוה מסף העמודה t_X אז תוכרז חריגה הקשורה לעמודה X.

תצוגה ויזואלית:

כאשר נבחר עמודה X נרצה לראות גרף המייצג את כל מרחקי ה z-Score מתחילת הטיסה ועד נקודת הזמן הנוכחית עבור העמודה שנבחרה. למשל כך:

אלגוריתם שני – רגרסיה ליניארית

זהו האלגוריתם שמימשנו בפת"מ 1. לבחירתכם הפלאג-אין יכול פשוט להתחבר לשרת שכתבתם או לממש את האלגוריתם בצורה מקומית.

למען הנוחות מצורף שוב תיאור האלגוריתם:

בשלב המקדים ניקח קובץ של טיסה תקינה ונבדוק לכל מאפיין מי מהמאפיינים האחרים הכי קורלטיבי אליו ע"פ Pearson. כלומר לכל מאפיין f_i נמדוד באמצעות Pearson את הקורלציה בין וקטור הערכים של f_i לבין וקטור הערכים של כל $f_{j\neq i}$. כך לכל מאפיין f_i נחזיר את המאפיין f_j שעבורו קבלנו את תוצאת ה Pearson בערך מוחלט הגבוהה ביותר.

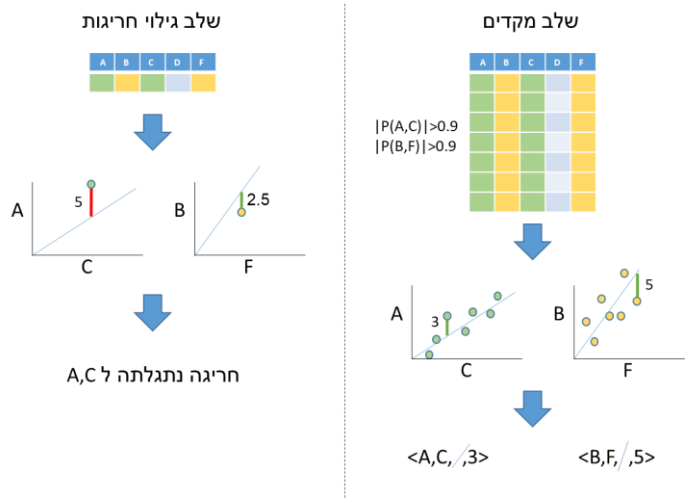
אך לא די בכך, נצטרך גם להגדיר איזשהו סף (threshold) שמעליו המאפיינים ייחשבו בכלל כקורלטיביים. לדוגמה אם נמצא שהקורלציה המקסימלית בין שני מאפיינים כלשהם היתה 0.2 זה לא יכול לעזור לנו לגלות חריגה שכן מראש שני המאפיינים הללו בלתי תלויים. לעומת זאת על קורלציה של 0.9 אפשר לסמוך הרבה יותר. הסף הזה הוא פרמטר של האלגוריתם וככזה הוא ישפיע מאד על הדיוק שלו.

לכל זוג מאפיינים שימצאו קורלטיביים דיים (קורלציה ישירה או הפוכה) נלמד את קו הרגרסיה שהנתונים שלהם מייצרים.

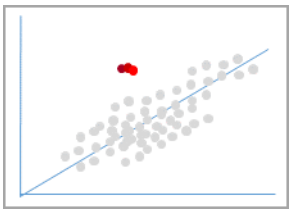
כעת נעבור על הטיסה התקינה שוב, ולכל זוג מאפיינים, ולכל נקודה דו-ממדית שלהם ב data נמדוד את ההיסט המקסימאלי שראינו ביחס לקו הרגרסיה שלהם. מכיוון שזו טיסה תקינה, אז נניח שהמרחקים שראינו מותרים וצפויים. אם בטיסת המבחן נראה נק' שמייצרת היסט גדול יותר, נתריע על כך כחריגה ואף נוכל לומר מיהם המאפיינים המעורבים בחריגה.

בשלב גילוי החריגות נקבל את הקלט תוך כדי טיסה – כלומר שורה אחר שורה. מכל שורה שכזו ניצור אוסף של נקודות דו-ממדיות ע"פ המאפיינים שלמדנו שהם קורלטיביים בשלב המקדים. לכל נק' דו-ממדית שכזו נמדוד את המרחק שלה מקו הרגרסיה שלמדנו עבור המאפיינים שיצרו אותה. אם המרחק הזה גדול דיו מהמרחק המקסימאלי שראינו עבור המאפיינים האלה נתריע על חריגה ונכלול בדיווח את המאפיינים המעורבים.

התרשים הבא מתאר אילוסטרציה של גלאי החריגות הפשוט שלנו:



תצוגה ויזואלית:



עבור העמודה הנבחרת נרצה לראות ייצוג של כל הנקודות הדו-ממדיות שנוצרות מהעמודה הנבחרת ובת זוגה הקורלטיבית ביותר אליה (ציר X וציר Y). ברקע נרצה לראות את הנקודות מתוך הטיסה התקינה ואת קו הרגרסיה שיצרו. את הנקודה הדו-ממדית שנוצרת מהטיסה הנוכחית נרצה לראות באופן בולט - למשל ע"י שובל של 30 הערכים האחרונים. כאשר ישנה חריגה האלגוריתם יציג זאת באופן בולט, למשל הוא יצבע את הרקע \ את הנק' הנוכחית באדום.

אלגוריתם שלישי – היברידי

כאשר לעמודה כלשהי קיימת קורלציה חלשה בלבד או לא קיימת קורלציה עם אף עמודה אחרת, האלגוריתם יכול לנהוג איתה כמו האלגוריתם הראשון – כלומר למדוד את ציון ה z-Score. אם יש קורלציה גבוהה עם עמודה אחרת אז האלגוריתם יכול לנהוג כמו האלגוריתם השני, כלומר עם רגרסיה ליניארית.

הסף לקורלציה גבוהה או נמוכה הוא **פרמטר** של האלגוריתם כאשר נגדיר את הערכים הדיפולטיבים הבאים:

- קורלציה הגדולה או שווה בערך מוחלט ל 0.95 תיחשב גבוהה (נשתמש באלגוריתם 2).
- קורלציה הנמוכה ממש מ 0.5 תיחשב חלשה (נשתמש באלגוריתם 1).

ומה לגבי עמודה שהקורלציה הגבוהה ביותר עבורה לעמודה אחרת היא בערך מוחלט בין 0.5 ל 0.95?

במקרה זה האלגוריתם יפעל כך:

בעת הלמידה האלגוריתם ימצא לכל זוג עמודות עם קורלציה כזו את המעגל המינימלי (מרכז + רדיוס) שמקיף את כל הנקודות הדו-ממדיות שנוצרו מהן בטיסה התקינה.

בעת הגילוי לכל זוג כזה של עמודות נמדוד האם הנקודה הדו-ממדית שנוצרת מהקלט הנוכחי נמצאת בתוך או מחוץ למעגל שלמדנו עבור זוג העמודות הללו בטיסה התקינה.

ההיגיון הוא שהקפנו את ענן הנקודות באזור שנחשב נורמלי. כל נקודה שתימצא מחוץ לאזור זה תיחשב כחריגה.

כדי למצוא את המעגל המינימלי באופן יעיל תוכלו להשתמש באלגוריתם של Welzl

https://en.wikipedia.org/wiki/Smallest-circle_problem

תצוגה ויזואלית:

כאשר נבחרת עמודה עם קורלציה חלשה תוצג תצוגה הדומה לזו של האלגוריתם הראשון, ואילו כאשר נבחרת עמודה עם קורלציה חזקה תוצג תצוגה הדומה לזו של האלגוריתם השני. כאשר יש נבחרה עמודה עם קורלציה לא חזקה ולא חלשה האלגוריתם יציג ברקע את הנקודות של שנוצרו מקלט הטיסה התקינה, את המעגל שמקיף אותם, ונקודה בצבע בולט (או עם שובל) שמשקפת את הקלט הנוכחי. ניתן למשל לצבוע אותה בירוק כל עוד היא בתוך המעגל ובאדום אם היא מחוצה לו. למשל כך: