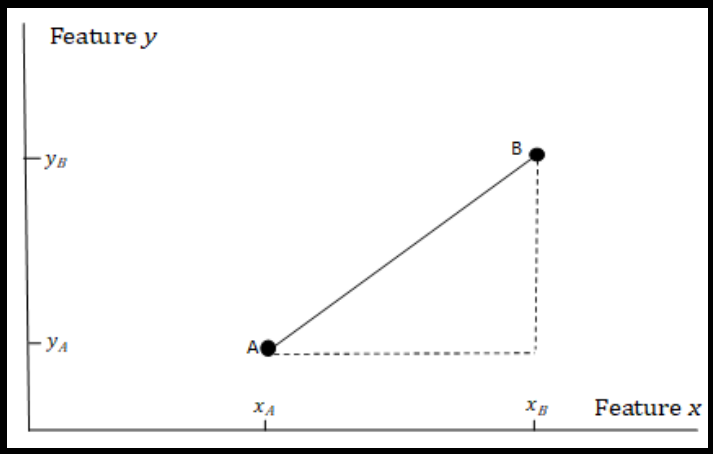
מדען נתונים, על אלגוריתם השכן הקרוב ביותר (k-Nearest Neighbors) כבר שמעת?

אחת האלטרנטיבות הפשוטות לרגרסיה לינארית או לרגסיה לוגיסטית היא אלגוריתם השכן הקרוב ביותר (k-Nearest Neighbors). אלגוריתם השכן הקרוב ביותר עוסק תחילה בבחירת ערך מסוים עבור k ולאחר מכן במציאת k התצפיות שהמאפיינים שלהן דומים ביותר למאפיינים שמתוכם אנו מבצעים חיזוי/סיווג.

נניח שאנו רוצים לחזות את השווי של בית פרטי בשכונה מסוימת מתוך גודל המגרש ושטח הבית במ"ר. אם נקבע $k = 3$, או אז יהיה עלינו לחפש את שלושת הבתים בסט האימון (Training Set) שלנו שהמאפיינים שלהם (מבחינת גודל המגרש ושטח הבית במ"ר) הם הדומים ביותר לאלו של הבית המוערך. אנו יכולים למדוד את הדימיון באמצעות כיול המאפיינים ולאחר מכן להשתמש במדד המרחק האוקלידי, שבנוסחה הבאה:

$$Distance = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}$$



ברגיל, כשישנם m מאפיינים המרחק בין נקודה P לנקודה Q הוא

$$\sqrt{\sum_{j=1}^{m} (v_{pj} - v_{qj})^2}$$

כאשר v_{pj} ו- v_{qj} הם הערכים של המאפיין ה- j עבור P ו- Q

מאחר ובחרנו $k = 3$, או אז עלינו לקחת את המחירים של שלושת הבתית הדומים ביותר מבחינת המאפיינים שלהם לאלו של הבית המוערך. נניח שהמחירים של שלושת הבתים הדומים ביותר הינם 2.30 מיליון ש"ח, 2.45 מיליון ש"ח ו- 2.18 מיליון ש"ח. לפיכך, השווי של הבית המוערך הנאמד באמצעות אלגוריתם השכן הקרוב ביותר עבור $k = 3$ שווה לממוצע האריתמטי של שלושת הבתים הללו או 2.31 מיליון ש"ח.

אילו היינו בוחרים $k = 2$, או אז עלינו לקחת את המחירים של שני הבתית הדומים ביותר מבחינת המאפיינים שלהם לאלו של הבית המוערך. נניח שהמחירים של שני הבתים הדומים ביותר הינם 2.30 מיליון ש"ח ו- 2.45 מיליון ש"ח. לפיכך, השווי של הבית המוערך הנאמד באמצעות אלגוריתם השכן הקרוב ביותר עבור $k = 2$ שווה לממוצע האריתמטי של שני הבתים הללו או 2.375 מיליון ש"ח.

אלגוריתם השכן הקרוב ביותר יכול גם לשמש לצורך סיווג. נניח שאנו רוצים לחזות האם הלוואה מסוימת תהיה טובה או לא מתוך ארבעת המאפיינים הבאים:

בעלות על בית, X_1, בעלות = 1, שכירות = 0	הכנסה (באלפי דולר), X_2	יחס החוב להכנסה, X_3	ציון אשראי (FICO), X_4	תוצאת ההלוואה, טובה = 1, חדלת פירעון = 0
1	43.304	18.47	690	0
1	136.000	20.63	670	1
0	38.500	33.73	660	0
1	88.000	5.32	660	1

אם נבחר $k = 10$, או אז יהיה עלינו לחפש את 10 ההלוואות בסט האימון שלנו שהמאפיינים שלהן הם הכי דומים לאלו של ההלוואה הנבחנת. נניח שמתוך 10 ההלוואות הדומות ביותר 8 התבררו כטובות (קרי, תוצאת ההלוואה שלהן היא 1) ו- 2 התבררו כחדלות פירעון (קרי, תוצאת ההלוואה שלהן היא 0). לפיכך, ההסתברות שההלוואה הנבחנת תגיע לחדלות פירעון הנאמדת באמצעות אלגוריתם השכן הקרוב ביותר עבור $k = 10$ שווה ל- 20% (2 מתוך 10).

אילו היינו בוחרים $k = 9$, או אז עלינו לחפש את 9 ההלוואות בסט האימון שלנו שהמאפיינים שלהן הם הכי דומים לאלו של ההלוואה הנבחנת. נניח שמתוך 9 ההלוואות הדומות ביותר 8 התבררו כטובות (קרי, תוצאת ההלוואה שלהן היא 1) ו- 1 התבררה כחדלות פירעון (קרי, תוצאת ההלוואה שלהן היא 0). לפיכך, ההסתברות שההלוואה הנבחנת תגיע לחדלות פירעון הנאמדת באמצעות אלגוריתם השכן הקרוב ביותר עבור $k = 9$ שווה ל- 11.11% (1 מתוך 9).

פרטים אודות כותב המאמר: האקטואר רועי פולניצר, FRM

רועי בעל תואר שני במימון (התמחות בניהול סיכונים ואקטואריה) ותואר ראשון בכלכלה (התמחות במימון), שניהם מאוניברסיטת בן-גוריון בנגב, בעל דיפלומה בניהול סיכונים פיננסיים (FRM®) מאוניברסיטת אריאל בשומרון ולמד בתוכנית ללימודי תעודה באקטואריה באוניברסיטת חיפה. כמו כן, רועי אקטואר מלא (Fellow) בלשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (F.IL.A.V.F.A.), מוסמך כמעריך שווי מימון תאגידי (CFV) מטעם לשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (IAVFA), מוסמך כמנהל סיכונים פיננסיים (FRM) מטעם האיגוד העולמי למומחי סיכונים (GARP) ומוסמך כמומחה לניהול סיכונים (CRM) מטעם האיגוד הישראלי למנהלי סיכונים (IARM).

לרועי ניסיון של מעל ל- 15 שנה בביצוע ניתוחים כמותיים במכשירים פיננסיים, בהערכת שווי תאגידים ונכסים בלתי מוחשיים, באמידה וכימות סיכונים כמו תמותה, אריכות ימים, תחלואה, ביטולים והחלמה מנכות, ובמידול ומדידת סיכוני שוק, אשראי, תפעוליים, מודל, נזילות והשקעות לצורכי יישום הוראות רגולטוריות ותקינה חשבונאית, פיתוח ,יישום ותיקוף מודלים בתחומים של הערכות שווי, ניהול סיכונים, אקטואריה והנדסה פיננסית, קביעת תעריפי ביטוח חיים, הערכת פרמיות סיכון והערכת עתודות ביטוח, קביעת עלות תנאי פנסיות (צוברות ותקציביות) והכנת מאזנים אקטואריים לקרנות פנסיה, ניתוח וחיזוי מצבים פיננסיים מורכבים וכן העברת סמינרי הדרכה והשתלמויות בתחומי התמחותו: מימון, אקטואריה, הערכות שווי, בנקאות, ניהול סיכונים, אופציות והנדסה פיננסית.

ניסיונו של רועי בתחום ה- Data Analysis, כולל: עבודה עם מאגרי מידע גדולים Big Data תוך שימוש ב- Statistical Learning (כגון: סטטיסטיקה תיאורית, הסתברות, הסקה סטטיסטית, סטטיסטיקה א-פרמטרית, חלוקת נתונים, נרמול נתונים, Fitting ו- Bayes Theorem) ובאלגוריתמים מסוג Unsupervised Learning (כגון: Hierarchical Clustering,k-means Clustering, Density-based Clustering, Distribution-based Clustering ו- Principle Components Analysis) למציאת דפוסים וזיהוי מגמות ואנומליות בעולמות ניהול הסיכונים, ההשקעות, האקטואריה, הביטוח והפנסיה, פיתוח תשתית לצורך ניתוח נתונים, שילוב והטמעת כלים לצורך גישה ושליפה עצמאית של נתונים ממאגרי מידע, פיתוח דוחות, ממשקים ומסכים באמצעות כלי ויזואליזציה.

ניסיונו של רועי בתחום ה- Data Science, כולל: עבודה עם מסדי נתונים גדולים Big Data תוך שימוש באלגוריתמים מסוג Supervised Learning (כגון: Linear Regression, Ridge Regression, Lasso Regression ,Elastic Net Regression, Logistic Regression, Maximum Likelihood Estimation, k-Nearest Neighbors, Decision Tree, Random Forest, Ensemble, Bagging,

Boosting, Naïve Bayes Classifier, Linear Separation, Support Vector Machine, Non-Linear Separation, SVM Regression, Artificial Neural Network, Convolutional Neural Network ו- Recurrent Neural Network) לניבוי וסיווג בעולמות ניהול הסיכונים, ההשקעות, האקטואריה, הביטוח והפנסיה ובמודלים מסוג Reinforcement Learning (כגון: Q-learning, Monte Carlo Simulation, Temporal Difference Learning ו- n-Step Bootstrapping) לקבלת החלטות מרובות שלבים בעולמות ניהול הסיכונים, ההשקעות, האקטואריה, הביטוח והפנסיה, זיהוי אתגרים עסקיים שבהם DATA יכול להוות גורם מכריע בשיפור קבלת החלטות, איתור ואיסוף מקורות מידע, הגדרה ואיפיון של שימושי המידע, בניית מסד המידע, אפיון והגדרת הצגת המידע ותוצריו, פיתוח כלים, מודלים, תהליכים ומערכות בתחום האנליזה, תוך שימוש בכלי אנליזה מתקדמים (EXCEL, VBA ו- R).