

מדען נתונים, על כיוול מאפיינים כבר שמעת?

למידה ללא השגחה (Unsupervised Learning) עוסקת בזיהוי דפוסים בנתונים. מטרתה המיידית של למידה ללא השגחה היא לא לנבא או לחזות את הערך של משתנה היעד (Target), המשתנה שאותו אנו מנסים לחזות), כי אם להבין את מבנה הנתונים ולפרק אותו לאשכולות (Clusters) הגיוניים. בנקים, לדוגמא, לעיתים קרובות משתמשים בלמידה ללא השגחה לביצוע ניתוח אשכולות ללקוחותיהם על מנת להבין אותם טוב יותר ולספק להם שירות טוב יותר. אשכול אחד יכול להיות HENRYs (זוגות צעירים שמשתכרים גבוה אך עדיין לא עשירים, High Earners Not Rich Yet). אלו משפחות שמרוויחות בין 35 אלף ש"ח ל- 70 אלף ש"ח בחודש ומחפשות גם שירותי ניהול עושר.

לפני שמשתמשים באלגוריתמים לניתוח אשכולות (כגון: ניתוח אשכולות (Cluster Analysis) עם k-Means, ניתוח אשכולות היררכי (Hierarchical Cluster Analysis), ניתוח אשכולות מבוסס-התפלגות (Distribution-Based Cluster Analysis), ניתוח אשכולות מבוסס-צפיפות (Density-Based Cluster Analysis)) ובאלגוריתמים לצמצום מימדים (Dimensionality Reduction) (כגון: ניתוח מרכיבים עיקריים (Principle Components Analysis)), יש לבצע כיוול מאפיינים (Feature Scaling, קליברציה למשתנים המסבירים), המכונה גם נורמליזציה (Normalization) או סטנדרטיזציה (Standardization) של הנתונים. כיוול מאפיינים הינו שלב ראשון והכרחי עבור אלגוריתמים רבים של למידת מכונה, לרבות אלגוריתם k-Means. מטרת כיוול המאפיינים היא להבטיח שהמאפיינים מקבלים חשיבות זהה באלגוריתם. נניח לדוגמא שאנו מבצעים ניתוח אשכולות לגברים לפי שני מאפיינים: גובה בס"מ ומשקל בק"ג. גובה יכול לנוע בין 150 ס"מ ל- 200 ס"מ בעוד שמשקל יכול לנוע בין 50 ק"ג ל- 150 ק"ג. למעשה ללא כיוול מאפיינים, שני המאפיינים הללו לא יטופלו באותה מידה של חשיבות, הואיל וטווח הגבהים הרבה יותר קטן מטווח המשקלים (50 ס"מ לעומת 100 ק"ג).



אחת הגישות לכיול מאפיינים היא לחשב את התוחלת (המסומנת באות היוונית μ) וסטיית התקן מסומנת באות היוונית σ) של כל אחד מהמאפיינים ולכייל את התצפיות של המאפיינים על ידי חיסור התוחלת מכל אחת מהתצפיות וחלוקת תוצאת הביניים (קרי, ההפרש שבין התצפית והתוחלת) בסטיית התקן. אם V הוא ערכו של מאפיין מסוים עבור תצפית מסוימת, הרי שערכו המכויל של המאפיין (Scaled Feature Value) יהיה:

$$\text{Scaled Feature Value} = \frac{V - \mu}{\sigma}$$

כאשר μ ו- σ מחושבים על בסיס כל התצפיות של אותו מאפיין. שיטה זו לכיול מאפיינים מכונה Z-score Normalization. למאפיינים המכוילים בשיטה זו יש תוחלת של 0 וסטיית תקן של 1. אם נרצה שלמאפיין מסוים תהיה השפעה גדולה יותר מאשר ליתר המאפיינים בעת קביעת פירוק הנתונים לאשכולות, הרי שעלינו לכייל אותו כך שסטיית התקן שלו תהיה גדולה יותר מ-1.

גישה אלטרנטיבית לכיול מאפיינים היא לחשב את הערך המינימלי (המסומן ב- min) והערך המקסימלי (המסומן ב- max) של כל אחד מהמאפיינים ולכייל את התצפיות של המאפיינים על ידי חיסור הערך המינימלי והערך מכל אחת מהתצפיות וחלוקת תוצאת הביניים (קרי, ההפרש שבין התצפית והערך המינימלי) בהפרש שבין הערך המקסימלי והערך המינימלי. אם V הוא ערכו של מאפיין מסוים עבור תצפית מסוימת, הרי שערכו המכויל של המאפיין (Scaled Feature Value) יהיה:

$$\text{Scaled Feature Value} = \frac{V - min}{max - min}$$

כאשר max ו- min מחושבים על בסיס כל התצפיות של אותו מאפיין. שיטה זו לכיול מאפיינים מכונה Min-Max Scaling. למאפיינים המכילים בשיטה זו יש ערכים הנעים בין 0 ל- 1.
על פי רוב, כיול בשיטת ה- Z-score עדיף על פני כיול בשיטת ה- Min-Max מאחר והראשון פחות רגיש לערכים קיצוניים, אך עם זאת יש היגיון להשתמש בכיול מסוג Min-Max כאשר הערכים נמדדים בקני מידה מוגדרים או תחומים.

פרטים אודות כתב 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.I.L.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 סיכונים 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



שלבם 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).