

למידה ללא השגחה (Unsupervised Learning)

למידת ללא השגחה עוסקת בהבנת דפוסים בתוך נתונים. באופן טבעי למידה ללא השגחה כרוכה בהתבוננות על אשכולות (Clusters), הווה אומר, על קבוצות של תצפיות דומות. לעיתים קרובות חברות עושות שימוש בלמידה ללא השגחה על מנת לנסות ולהבין טוב יותר את סוגי הלקוחות שלהן כך שהן תוכלנה לתקשר עמם בצורה יעילה יותר.

ביצוע קליברציה לערכי המאפיינים (Feature Scaling) הינו שלב מקדים (Precursor) לשלב ניתוח האשכולות (Clustering). ללא ביצוע קליברציה לערכי המאפיינים, השפעתו של מאפיין מסוים על ניתוח אשכולות תהיה בקנה המידה (Scale) אשר משמש למדידתו.

קיימות שתי גישות עיקריות לביצוע קליברציה לערכי המאפיינים. הגישה הראשונה נקראת קליברציה מסוג Z-score ובמסגרתה ערכי המאפיינים מותאמים כך שבסופו של דבר תהיה להם תוחלת של 0 וסטיית תקן של 1. הגישה השנייה נקראת קליברציה מסוג Min-Max ובמסגרתה ערכי המאפיינים מותאמים כך שבסופו של דבר הם ינועו בין 0 ל-1.

אלגוריתם ניתוח אשכולות דורש לדעת למדוד מרחקים (משהו שלמדנו מתישהו בתיכון). המדד הפופולארי ביותר עבור מרחק נקרא המרחק האוקלידי (Euclidian Distance) והוא מחושב כשורש הריבועי של סכום ריבועי ההפרשים שבין ערכי המאפיינים. מרכז הכובד של האשכול הינו נקודה המתקבלת על ידי מיצוע ערכי המאפיינים עבור כל התצפיות באשכול. אלגוריתם ניתוח האשכולות הפופולארי ביותר נקרא k-מרכזים (k-means). עבור ערך מסוים של k (מספר האשכולות), אלגוריתם ה-k-מרכזים ממזער את האינרציה (Inertia), המוגדרת כסכום ריבועי המרחקים בתוך האשכול בין התצפיות לבין מרכזי הכובד של האשכולות שלהן.

בחירת הערך הטוב ביותר עבור מספר האשכולות, k, איננה משימה קלה. גישה אחת לבחירת ה-k נקראת "שיטת המפרק" (Elbow Method) הגורסת שיש להמשיך ולהעלות את ה-k עד



להגעה לשיפור זניח יחסית באינרציה. גישה אחרת נקראת "שיטת הצללית" (Silhouette Method) והיא עורכת השוואה בין המרחק הממוצע של תצפית מסוימת מהנקודות האחרות באשכול שלה לבין המרחק הממוצע שלה מהנקודות באשכול האחר הקרוב ביותר. הגישה השלישית כרוכה בחישוב סטטיסטי הפער, אשר משווה את התצפיות שבתוך האשכולות (Clustered Observations) לתצפיות הנוצרות באופן אקראי.

ישנן מספר חלופות לאלגוריתם k-מרכזים. החלופה הראשונה נקראת ניתוח אשכולות היררכי (Hierarchical Clustering). בניתוח אשכולות היררכי אנו מתחילים ממצב שבו כל אחת מהתצפיות נמצאת באשכול שלה. לאחר מכן אנו מורידים באיטיות את מספר האשכולות על ידי צירוף אשכולות שקרובים אחת לשני לכדי אשכולות חדשים. החלופה השנייה נקראת ניתוח אשכולות מבוסס-התפלגות (Distribution-based Clustering) והיא כרוכה בהתבוננות על אזורים שבהם הנתונים צפופים/דחוסים ללא קשר למרכזי הכובד של האשכולות.

ניתוח מרכיבים עיקריים (PCA- Principal Components Analysis) הוא כלי חשוב בלמידת מכונה. ניתוח מרכיבים עיקריים כרוך בהחלפת מספר גדול של מאפיינים במספר קטן יותר של מאפיינים המכונים "המאפיינים המסבירים ביותר" (Manufactured Features) אשר תופסים את מרבית ההשתנות של היעד (Target Variability, השתנות המשתנה המוסבר). נעיר רק שהמאפיינים המסבירים ביותר אינם מתואמים האחד עם השני.

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.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 סיכוני שוק, אשראי, תפעוליים, מודל, מזילות 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 וחזוי 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).