

## סיכום למידה ללא השגחה

למידת ללא השגחה (Unsupervised Learning) 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) הגורסתת 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: מדען הנתונים רועי פולניצר, PDS

- מייסד ומנכ"ל האיגוד הישראלי למדעני נתונים מקצועיים (PDSIA), מייסד ויו"ר לשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (IAVFA) ובעלים של פירמת הייעוץ וההדרכה שווי פנימי.
- מחזיק בתואר M.B.A. במנהל עסקים עם התמחות בניהול סיכונים ואקטואריה ותואר B.A. בכלכלה עם התמחות במימון שניהם בהצטיינות מאוניברסיטת בן-גוריון בנגב, דיפלומה בניהול סיכונים פיננסיים (FRM) מאוניברסיטת אריאל, תואר Financial Risk Manage מארגון בינ"ל GARP, תואר Certified Risk Manage מארגון ישראלי IARM, תואר Fellow Actuary מארגון ישראלי IAVFA ותואר Professional Data Scientist מארגון ישראל PDSIA.
- בעל ניסיון אינטנסיבי של מעל עשור וחצי שנים בתחום מדע הנתונים ולמידת המכונה, הכולל ביצוע מחקרי מידע מעמיקים לשם הפקת תובנות עסקיות, ניקוי, טיוב וסידור של המידע המשמש למחקרים השונים, הפעלת אלגוריתמים שונים של מידול, כריית נתונים ו-Machine Learning על המידע ובניית תהליכי הכנת המידע והאופטימיזציה של האלגוריתמים השונים.
- מרצה לתכנות בשפות R ו-Python, לניהול סיכונים, הערכות שווי ואקטואריה והנדסה פיננסית.