

# סטטיסטיקאי, על למידה ללא השגחה כבר שמעת?

למידה ללא השגחה משמשת לזיהוי דפוסים (ניתוח אשכולות, Clusters) בנתונים, כלומר לפירוק נתונים לקבוצות הגיוניות. יצאתי לראיין את רועי פולניצר בנושא

השגחה שמטרתו להבטיח שההתייחסות למאפיינים (Features), משתנים המשמשים באלגוריתם של למידת מכונה) תהיה זהה מבחינת רמת חשיבות, כלומר, שהמאפיינים נמדדים על בסיס אותם קני מידה (Scales).



מדען הנתונים רועי פולניצר מכהן כמנכ"ל האיגוד הישראלי למדעני נתונים מקצועיים (PDSIA). רועי בעל תואר M.B.A. במנהל עסקים (עם התמחות בבניית מודלים מתמטיים וסטטיסטיים) ותואר B.A. בכלכלה (עם התמחות בכלים ושיטות לאנליזה ותחקור מידע), שניהם מאוניברסיטת בן-גוריון בנגב ובהצטיינות. רועי מחזיק בכמה הסמכות מקצועיות רלוונטיות למדע נתונים ולמידת מכונה, ביניהן הסמכה מקצועית בינלאומית "מנהל סיכונים פיננסיים" (FRM - Financial Risk Manager) מטעם האיגוד העולמי למומחי סיכונים (Global Association of Risk Professionals) המעידה על כך שהמחזיק בה בקיא בפיתוח, יישום ותיקוף מודלים סטטיסטיים ואלגוריתמים מתמטיים כגון SVM, K-Means, PCA ו-KNN למדידה וניהול סיכונים אשראי. בשנה האחרונה רועי הקים את האיגוד הישראלי למדעני נתונים מקצועיים ולכן יצאתי לראיין אותו בנושא.

למידה ללא השגחה (Unsupervised Learning) עוסקת בזיהוי דפוסים בנתונים. המטרה המיידית 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). בנקים, לדוגמה, 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 זוגות צעירים שמשתכרים גבוה אך עדיין לא עשירים. אלו משפחות שמרוויחות בין 35 אלף ש"ח ל-70 אלף ש"ח בחודש ומחפשים גם שירותי ניהול עושר. בראיין זה רועי פולניצר יסביר נוהל פשוט לניתוח אשכולות המכונה אלגוריתם k-Means, יזכיר כמה גישות אלטרנטיביות לניתוח אשכולות ויתן דוגמה לניתוח מרכיבים עיקריים (PCA), כלי שימושי מאוד הן ללמידה בהשגחה והן ללמידה ללא השגחה.

## האם תוכל להסביר מהן שתי השיטות לביצוע קליברציה לערכי מאפיינים?

השיטה הראשונה היא קליברציה מסוג Z-score, לפיה כל אחד מהמאפיינים מכיל על ידי ניכוי התוחלת ממנו וחלוקת התוצאה המתקבלת בסטיית התקן. התוצאה היא שלמאפיינים אחרי קליברציה מסוג Z-score יש תוחלת של 0 וסטיית תקן של 1. השיטה השנייה היא קליברציה מסוג min-max, לפיה כל אחד מהמאפיינים מכיל על ידי ניכוי ערך המאפיין המינימלי ממנו וחלוקת התוצאה בהפרש שבין ערך ערך המאפיין המקסימלי לבין ערך המאפיין המינימלי. התוצאה היא שהמאפיינים אחרי קליברציה מסוג min-max נעים בין 0 ל-1.

## מדוע קליברציה לערכי המאפיינים חשובה כל כך בלמידה ללא השגחה?

למידה ללא השגחה (Unsupervised Learning) עוסקת במציאת דפוסים בנתונים, לעיתים קרובות, על ידי שימוש בניתוח אשכולות. קליברציה לערכי המאפיינים (Feature Scaling) הינה נוהל חשוב מאוד בלמידה ללא

## מהם היתרונות והחסרונות של כל אחת מהשיטות?

חסרונה של קליברציה מסוג min-max הוא שהיא לא עובדת בצורה טובה כאשר ישנם חריגים, הואיל ואז יתר הערכים המכילים הינם קרובים זה לזה וזהו בדיוק היתרון של קליברציה מסוג Z-score. יתרונה של קליברציה מסוג min-max הוא שהיא עובדת בצורה טובה כאשר המאפיינים נמדדים בקני מידה שונים עם גבול תחתון וגבול עליון וזהו בדיוק החיסרון של קליברציה מסוג Z-score.

## האם תוכל לתת דוגמה לחישוב מרחק אוקלידי בין שתי תצפיות?

נניח שקיימים שלושה מאפיינים A, B ו-C. לתצפית אחת יש ערכים 2, 3 ו-4 עבור A, B ו-C, בהתאמה. לתצפית אחרת יש ערכים 6, 8 ו-7 עבור A, B ו-C, בהתאמה. כיצד את המרחק האוקלידי בין שתי התצפיות הללו מחשבים באופן הבא:

$$\sqrt{(6-2)^2 + (8-3)^2 + (7-4)^2} = 7.07$$

## מהו מרכז הכובד של האשכול הזה?

המרכז מתקבל ממיצוע ערכי המאפיינים. מדובר בנקודה שיש לה את הערכים 4 (ממוצע של 6 ו-2), 5.5 (ממוצע של 8 ו-3) ו-5.5 (ממוצע של 7 ו-4) עבור שלושת המאפיינים.

## אני ואתה נבחנו במבחנים הבינלאומיים להסמכה בתחום של ניהול סיכונים (FRM). אני זוכר שנבחנתי במבחנים הללו על אלגוריתם k-means, האם תוכל להסביר מהו האלגוריתם הזה?

אלגוריתם k-means, המכונה בעגה המקצועית של מדעני הנתונים בישראל k-מרכזים, הינו אלגוריתם למציאת אשכולות, הווה אומר, פירוק נתונים לאשכולות לצורך זיהוי דפוסים בנתונים עצמם.

## מהם שלבי הפעולה של אלגוריתם k-means?

שלבי הפעולה של האלגוריתם הזה הם שאנו בוחרים תחילה k נקודות בתור מרכזי כובד של אשכולות, מקצים תצפיות למרכז הכובד הקרוב ביותר, מחשבים מחדש את מרכזי הכובד של האשכולות, מקצים מחדש תצפיות למרכזי הכובד של האשכולות, וחוזר חלילה.

## כיצד בוחרים את ה-k?

קיימות שלוש שיטות לבחירת ה-k: שיטת המרפק (elbow method) ושיטת הצללית (silhouette method) ומבחן הפער (gap statistic). שיטת המרפק הינה פרוצדורה לבחירת מספר האשכולות האופטימלי על ידי צפייה על השפעת האינרציה (Inertia), סכום ריבועי המרחקים בתוך האשכול כאשר הנתונים מקובצים באשכול. בשיטת המרפק אנו מחפשים את הנקודה שעבורה השיפור השולי באינרציה, כאשר אשכול נוסף נוצר, הוא נמוך. שיטת הצללית הינה דרך לחישוב מספר האשכולות בהתבסס על המרחקים בין התצפיות. בשיטת הצללית, אנו מחשבים עבור כל ערך של k ועבור כל תצפית i:

a(i): המרחק הממוצע מכל אחת מהתצפיות באשכול שלו, ו-  
b(i): המרחק הממוצע מהתצפיות באשכול הקרוב ביותר  
הצללית של התצפית היא:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

הערך הטוב ביותר של k הוא זה שעבורו הצללית הממוצעת על פני כל התצפיות היא הגדולה ביותר. מבחן הפע הפע הוא דרך לבחירת מספר האשכולות על ידי השוואת הנתונים שבאשכולות לנתונים שמפולגים מקרית.

## האם תוכל להסביר מדוע המרחק בין תצפיות גדל ככל שמספר המאפיינים עולה?

ככל שמספר המאפיינים עולה, לסכום ריבועי המרחקים בין ערכי המאפיינים יש יותר תנאים ולכן הוא נוטה לגדול. נניח שהתחלת עם 10 מאפיינים ואז בטעות יצרת עוד עשרה מאפיינים שזהים לעשרה הראשונים, איזו השפעה יש לכך על המרחק בין שתי התצפיות? כאשר עשרת המאפיינים הנוספים נוצרים בטעות, המרחק בין שתי תצפיות גדל ב- $\sqrt{2}$  מאחר וכל הפרש ריבועי מחושב פעמיים.

## בוא תסביר לי כיצד עובד ניתוח אשכולות היררכי? ומהן היתרונות והחסרונות שלו בהשוואה לאלגוריתם k-means?

ניתוח אשכולות היררכי (Hierarchical Clustering) הינו דרך לבניית אשכולות תצפית אחת בכל פעם. בניית אשכולות היררכי אנו מתחילים לשים כל תצפית באשכול שלה. בכל שלב אנו מוצאים את שני האשכולות הקרובים ביותר ומצרפים אותם יחד לכדי אשכול חדש. החסרון הוא שניתוח אשכולות היררכי הוא איטי. היתרון הוא שניתוח אשכולות היררכי מזהה אשכולות בתוך אשכולות.

## קראתי במאמר שלך שיש עוד שני סוגים של ניתוחי אשכולות, האם תוכל להסביר מהם?

ניתוח אשכולות מבוסס-התפלגות (Distribution-Based Clustering) הינו דרך לניתוח אשכולות באמצעות התאמת תצפיות לתמהיל של התפלגויות. ניתוח אשכולות מבוסס-התפלגות מניח שתצפיות נוצרות מתמהיל של שתי התפלגויות ומשתמש בשיטות סטטיסטיות להפריד ביניהם. ניתוח אשכולות מבוסס-צפיפות (Density-Based Clustering) הינו דרך ליצירת דפוסים של אשכולות לא סטנדרטיים. ניתוח אשכולות מבוסס-צפיפות כרוך בהוספת נקודות חדשות לאשכול שקרובות למספר נקודות שכבר נמצאות באותו אשכול.

## גם על ניתוח מרכיבים עיקריים נבחנו במבחנים הבינלאומיים להסמכה בתחום של ניהול סיכונים (FRM). האם תוכל להסביר מהי אותה שיטה סטטיסטית?

ניתוח מרכיבים עיקריים (Principals Components Analysis), המכונה בעגה המקצועית של מדעני הנתונים PCA, הינו דרך להחליף נתונים עם מאפיינים מתואמים במספר קטן של מאפיינים לא מתואמים.

## באלו נסיבות PCA שימושי ביותר להבנת נתונים?

לשאלתך ה- PCA שימושי ביותר כאשר קיימים מספר מאפיינים שמתואמים מאוד. ל- PCA יש את הפוטנציאל להסביר את מירב ההשתנות בנתונים באמצעות מספר קטן של מאפיינים חדשים (Manufactured Features) שאינם מתואמים האחד עם השני.

## האם תוכל לתת לנו דוגמה ליישום של PCA?

כאשר רוצים למדוד רגישויות של אג"ח לריבית יש לזכור שגם שיעורי התשואה משתנים בצורה מורכבת. מקובל לייחס את השינויים בריביות הספוט לשלושה מרכיבים עיקריים. המרכיב הראשון הוא שינוי בחותך (שינוי במקביל של עקום התשואות), לפיו מסתכלים על מה שקורה לשיעור הריבית ל-3 חודשים. המרכיב השני הוא שינוי בשיפוע, לפיו מסתכלים על מה שקורה להפרש שבין ריבית ל-10 שנים לבין הריבית לשנה. המרכיב השלישי הוא שינוי בקמירות, לפיו מסתכלים על מה שקורה במשך שנה להפרש שבין שיעור הריבית ל-5 שנים לבין ממוצע שיעורי הריבית לשנה ול-10 שנים. מניתוח PCA שביצעתי על מנת לראות כמה כל אחד משלושת המרכיבים הללו מסביר את השינויים בריביות הספוט, עולה שהשינויים בחותך מסבירים 71% מהשינויים בריביות הספוט, השינויים בשיפוע מסבירים 26% מהשינויים בריביות הספוט והשינויים בקמירות מסבירים רק 2% מהשינויים בריביות הספוט. לפיכך, שלושת הגורמים העיקריים הללו מסבירים יחדיו בסך הכל 99% מהשינויים בריביות הספוט. זו דוגמה קלאסית של PCA.