

ניקוי נתונים

ניקוי נתונים הינו היבט חשוב ביותר, שלא לומר מרתק, של למידת מכונה. מסקרים שנעשו בארה"ב על הנושא עולה כי מדעני נתונים מקדישים לא פחות מ- 80% מזמנם לניקוי נתונים.

ברגיל, מאגרי מידע גדולים טומנים בחובם בעיות הטעונות תיקון. ניקוי נתונים טוב הוא זה שעושה את כל ההבדל בין למידת מכונה מוצלחת לכזו שאינה מוצלחת. הביטוי "זבל נכנס, זבל יוצא" (garbage-in, garbage-out) רלוונטי וישים לגבי למידת מכונה ממש באותה מידה שהוא תקף ונכון לגבי ניתוחים אחרים.

בשלב זה, נציין כי קיימים שני סוגים של נתונים: נומריים (Numerical) וקטגוריאליים (Categorical).

נתונים נומריים מכילים מספרים, בעוד שנתונים קטגוריאליים הינם נתונים שיכולים 'ליפול' לתוך מספר קטגוריות שונות. לדוגמא, נתונים לניבוי מחירי דירות יכולים לקטלג דירה מסוימת כמשופצת או כלא משופצת, עם או ללא מעלית, עם מזגן או בלי מזגן וכיוצא באלה קטגוריות (שותפה מעשנת, שומרת שבת, נקיה, אוהבת חיות ומסיבות וכו'). למעשה ניתן להמיר נתונים קטגוריאליים למספרים לצורך הניתוח.

הסוגיה הראשונה הקשורה לניקוי נתונים הינה "תיעוד או רישום לא עקבי" (Inconsistent Recording) של נתונים. לאמור- נתונים נומריים וקטגוריאליים כאחד עלולים להיות חשופים לתיעוד לא עקבי. לדוגמא, נתונים נומריים עבור שטח של דירה מסוימת עשויים להיות רשומים במאגר נתונים כלשהו כ- 100, מאה, 1.0 ממ"ר, 100 מ"ר, +100, וכך הלאה. לפיכך, חשוב ביותר ראשית לבחון את הנתונים על מנת לקבוע איזו וריאציות יכולים לקבל הנתונים ואז להחליט על הגישה הטובה ביותר לניקוי. נתונים קטגוריאליים עשויים לציין דירה מסוימת כ-: "משופצת", "אחרי שיפוץ", "כמו חדשה" וכו'. הגישה הפשוטה ביותר היא להכין רשימה של כל החלופות הקיימות במאגר עבור מאפיין מסוים (משופצת או לאו במקרה דנן שלפנינו) ולאחר מכן לאחד אותם בצורה הולמת.

הסוגיה השנייה הקשורה לניקוי נתונים מכונה "תצפיות לא רצויות" (Unwanted Observations). לדוגמא, אם אתה מפתח מודל לניבוי מחירי דירות באזור מסוים, אז חלק מהנתונים שלך עשויים להתייחס בכלל למחירי דירות או מחירי בתים שאינם באזור הנבדק. על כן חשוב מאוד למצוא דרך לזהות את הנתונים הללו ולהסירם טרם התחלת הניתוח.

הסוגיה השלישית הקשורה לניקוי נתונים נקראת "תצפיות כפולות" (Duplicative Observations). לדוגמא, כאשר נתונים ממספר מקורות מידע שונים "מוכנסים" לתוך מאגר נתונים אחד, הרי שאך טבעי שתהייה תצפיות כפולות אשר עלולות כידוע להביא להטיית תוצאות הניתוח. לפיכך, חשוב לעשות שימוש באלגוריתמים של חיפוש על מנת למצוא ולהסיר כפילויות ככל שרק שניתן.

הסוגיה הרביעית הנוגעת לניקוי נתונים נקראת "חריגים" (Outliers). במקרה של נתונים נומריים ניתן לזהות חריגים על ידי העלאת הנתונים על תרשים (תרשים פיזור או היסטוגרמה) או לחילופין באמצעות חיפוש נתונים אשר מרוחקים שש סטיות תקן (Six Sigmas) מהתוחלת. לעיתים די ברור שהחריגים הינם תוצאה של טעות הקלדה. לדוגמא, אם שטחה של דירת 3 חדרים מופיע במאגר כ- 1,000 מ"ר אז די ברור שמדובר בטעות הקלדה. עם זאת, כפי שלימד אותי מורי ורבי המלומד ה"ה ד"ר שילה ליפשיץ ז"ל (שהייתה לי הזכות להיות עוזר מחקר שלו במשך 4 שנים) "חריגים מסירים רק אם קיימת סיבה מספיק טובה לעשות זאת". נתונים נומריים שגדולים באופן יוצא דופן או שקטנים בצורה לא רגילה, אם הם אכן נכונים, הרי שהם מכילים מידע שימושי. השפעתם של חריגים על התוצאות של למידת מכונה תלויה במודל שבו נעשה שימוש. כך למשל לחריגים ישנה השפעה רבה על מודלים מסוג רגרסיה, בעוד שעל מודלים מסוג עצי החלטה (Decision Trees) כמעט ואין להם כל השפעה.

הסוגיה החמישית הנוגעת לניקוי נתונים מכונה "נתונים חסרים" (Missing Data). כמעט בכל מאגר נתונים גדול חסרים ערכים. הגישה הפשוטה לטיפול בסוגיה זו היא פשוט להסיר את הנתונים שלגביהם חסרים ערכים עבור מאפיין אחד או יותר. אבל זה ממש לא רצוי הואיל ופעולה שכזו מקטינה אפקטיבית את גודל המדגם ועשויה ליצור הטיית. במקרה של נתונים קטגוריאליים, הפתרון הפשוט ביותר הוא ליצור קטגוריה חדשה שנקראת "חסרים" (Missing).

במקרה של נתונים נומריים, אחת הגישות היא פשוט להחליף את הנתונים החסרים בתוחלת או בחציון של ערכי הנתונים שאינם חסרים. לדוגמא, אם שטחה של דירה מסוימת חסר ומצאנו שחציון שטחי הדירות ישנם נתונים זמינים נאמד בכ- 115 מ"ר, הרי שאנו יכולים לעשות שימוש בערך שמצאנו עבור כל יתר הדירות שחסרים לנו נתונים אודות גודל שטחן.

גישות קצת יותר מתוחכמות כוללות הרצת רגרסיות על היעד (Target), מחירי הדירות במקרה דנן שלפנינו) ביחס לערכים שאינם חסרים ולהשתמש בתוצאות הרגרסיה עבור הערכים שחסרים. לעיתים מקובל להניח שנתונים חסרים באופן מקרי (פשוט כי ככה זה) ולעיתים עצם העובדה

שחסרים נתונים היא כשלעצמה אינפורמטיבית. במקרה האחרון, מקובל לייצר משתנה אינדיקטור חדש שמקבל את הערך 0 אם קיימים נתונים או את הערך 1 אם חסרים נתונים.



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.