

רואה חשבון, על מדע נתונים כבר שמעת?

בעיות רגרסיה אלו בעיות שבהן אנו מנסים לחזות מספר רציף, בעיות סיווג אלו בעיות שבהן אנו מנסים להגיד לאיזו קטגוריה שייכת תצפית מסוימת ובעיות ניתוח אשכולות אלו בעיות שמנסות לפרק לנו את הנתונים לקבוצות הגיוניות.

קודם כל צריך להבין שישנם מספר סוגים של בעיות שאנחנו יודעים לפתור בתחום של מדע נתונים או למידת מכונה. בגדול אני מחלק את הבעיות הללו לשלושה סוגים מרכזיים: משפחת הרגרסיה (חיזוי/ניבוי), משפחת סיווג (שיוך לקטגוריות) ומשפחת ניתוח האשכולות (פירוק לקבוצות). חשוב להבין שאין מדובר במשפחות שונות לחלוטין. כך למשל, יש הרבה מאוד קשר בין משפחת הרגרסיה למשפחת הסיווג, אבל אני לא אכנס לזה עכשיו.

מהי בעיית רגרסיה?

רגרסיה (Regression) זוהי סוג בעיה שבה אנו מנסים לחזות ערך רציף (כגון: מחיר, משקל, זמן וכו'). כאשר החיזוי שלנו הוא לחזות למשל 4.7 או 5.38%.



תוכל לתת מספר דוגמאות לבעיות ממשפחת הרגרסיה?

בהחלט. דוגמה אחת לבעיה ממשפחת הרגרסיה היא מודל ה-CLV (ערך חיי לקוח, Customer Lifetime Value). לחברות, למשל, יש חשיבות עצומה לדעת בשלב מוקדם על לקוח שלהם, האם הוא לקוח

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, כגון: הסמכה בינלאומית "מנהל סיכונים פיננסיים" (Financial Risk Manager), המעידה על כך שהמחזיק בה בקיא בפיתוח, יישום ותיקוף מודלים סטטיסטיים ואלגוריתמים מתמטיים כגון SVM, K-Means ו-KNN למדידה וניהול סיכונים (אשראי מטעם האיגוד העולמי למומחי סיכונים (Association of Risk Professionals), המעידה על כך שהמחזיק בה בקיא בפיתוח, מלא" (Fellow), המעידה על כך שהמחזיק בה בקיא בפיתוח, יישום ותיקוף מודלים סטטיסטיים ואלגוריתמים מתמטיים כגון GLM, RF ו-NN לקביעת פרמיות בביטוח כללי) מטעם לשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (Association of Valuators and Financial Actuaries), המעידה על כך שהמחזיק בה בקיא בפיתוח, יישום ותיקוף מודלים סטטיסטיים ואלגוריתמים מתמטיים כגון DT, NB ו-PCA לניהול סיכונים תפעוליים) מטעם האיגוד הישראלי למנהלי סיכונים (Israeli Association of Risk Managers). בשנה האחרונה רועי ייסד את האיגוד הישראלי למדעני נתונים מקצועיים (PDSIA - Professional Data Scientists' Israel Association) ולכן יצאתי לראיין אותו בנושא.

איפה בדיוק מדע נתונים בא לידי ביטוי?

האם תוכל לתת כמה דוגמאות לבעיות ממשפחת הסיווג?

בוודאי. דוגמא אחת לבעיה ממשפחת הסיווג היא סינון ספאם, שבמסגרתה אנו רוצים להבין האם מייל מסוים הוא ספאם, אם לאו. דוגמא אחרת לבעיה ממשפחת הסיווג היא בעיית נטישה, שעליה דיברנו קודם. כך למשל בבעיית נטישה של לקוחות, אני רוצה לדעת האם הלקוח הספציפי הזה יעזוב אותי או לא יעזוב אותי במהלך תקופת זמן מסוימת. דוגמא נוספת לבעיה ממשפחת הסיווג היא זיהוי הונאה ומדובר במונח שהיום מאוד מאוד "חם" אצל חברות כמו PayPal ו-eBay. במסגרת בעיית זיהוי הונאה אני רוצה לדעת האם סדרת עסקאות מסוימת של בנאדם כלשהו מעידה על איזשהי הונאה. למעשה בסופו של דבר אני צריך להגיד האם הייתה הונאה, אם לאו. שלוש הדוגמאות שהצגתי אלו בעיות ממשפחת הסיווג ויש להם גם מודלים אחרים או שימושים אחרים מלבעיות ממשפחת הרגרסיה.

מהי בעיית ניתוח אשכולות?

בעיית ניתוח אשכולות (Clustering) היא בעיה קצת פחות אינטואיטיבית מהבעיות האחרות. מדובר בבעיה שבה אני לא מסנה שהמודל יגיד לי לאיזו קטגוריה לשייך תצפית מסוימת או מהי התחזית על סמך תצפית מסוימת. בבעיית ניתוח אשכולות אני רוצה שהמודל "יפרק" לי את הנתונים לקבוצות, בדגש על קבוצות הגיוניות.

האם תוכל לתת כמה דוגמאות לבעיות ממשפחת ניתוח האשכולות?

בוודאי. דוגמא אחת לבעיה ממשפחת ניתוח האשכולות היא פילוח לקוחות. נניח לרגע שאני חברת סלולאר ויש לי מיליון לקוחות ואני רוצה לצאת במבצעים או לתת הצעות ללקוחות שלי. די ברור שהדבר הכי טוב היה אילו הייתי יכול לתת הצעה ייחודית לכל לקוח. עם זאת, ברור שזה ממש לא פרקטי מבחינה יישומית ומצד שני אני לא יכול לתת מבצע אחד שיתאים לכל הלקוחות שלי, כי פשוט אין חיה כזו. שום מבצע לא מתאים לכלל הלקוחות. אז אני רוצה למצוא משהו באמצע, ולכן אני רוצה לפלח את הלקוחות שלי לקבוצות, כאשר כל קבוצה מורכבת מלקוחות שדומים זה לזה, במובן מסוים. באיזה מובן? זוהי שאלת מיליון הדולר. ואלו הם הדברים שהאלגוריתמים של ניתוח אשכולות, כגון: ניתוח אשכולות עם K-Means, ניתוח מרכיבים עיקריים (Hierarchical Cluster Analysis), ניתוח מרכיבים עיקריים (Principle Components Analysis) וכו' מנסים לעזור לנו להבין. אם הייתי משתמש במסד נתונים מסוים ומפלח אותו לשתי קבוצות: גברים לחוד ונשים לחוד, אז אני לא בטוח שהפילוח הזה רק על פי המגדר "מפרק" לי את הלקוחות בצורה נכונה. תחשוב על זה שלבנק מסוים לא באמת איכפת האם הלקוח הוא גבר או אישה, אלא איכפת לו יותר האם הלקוח הוא קמפן או לא, האם הלקוח גר רחוק מהבנק או לא, האם הלקוח הוא טיפוס דיגיטלי או לא. האם הלקוח הזה זה בנאדם שעכשיו כדאי לי להציע לו הצעות במייל או שהוא לעולם לא יראה אותן כי אין לו בכלל מייל. ניתוח אשכולות הוא למעשה חלוקה של הלקוחות לקבוצות שיש בהן עניין עסקי כלשהו. דוגמא נוספת לבעיה ממשפחת ניתוח האשכולות מכונה מידול נושאים. תחשוב לרגע שאתה מקבל הרבה מאוד טקסטים ואתה רוצה לחלק את הטקסטים האלה לטקסטים שעוסקים במוסיקה, טקסטים שעוסקים בפוליטיקה וטקסטים שעוסקים בספורט - אבל אתה לא יודע אלו סוגי טקסטים יש לך על השולחן. כלומר, אם אומרים לך יש פה טקסטים שעוסקים בספורט, טקסטים שעוסקים בפוליטיקה וטקסטים שעוסקים במוסיקה אז זוהי כבר בעיית סיווג כי אתה מקבל טקסט ואז אתה צריך להגיד לאיזה נושא הוא שייך. אבל תחשוב על מצב שבו לא אומרים לך אלו נושאים בכלל קיימים ואתה נדרש להחליט מהם הנושאים הכי רלוונטים שמיוצגים באוסף הטקסטים הזה שקיבלת. לכן זוהי בעיית ניתוח אשכולות ולא בעיית סיווג.

טוב או לא. אם ניקח לדוגמא חברה למשחקי אפליקציה, שנכנסתי למשחק שלה ושיחקתי בו שבוע. חברה שכזו הייתה מאוד רוצה לדעת האם אני בנאדם שמתמכר למשחק או לא, באלו שעות בדרך כלל אני משחק וכיצא באלה דברים כדי לדעת מתי לשלוח לי את ההודעה (Notification) שלה בזמן הנכון. ואם אנחנו מדברים על מודל ה-CLV, אז אותה חברה רוצה לדעת היום אם היא תשקיע בי עכשיו כלקוח שלה על מנת שאמשיך להיות שחקן אצלה באפליקציה, כמה כסף היא תראה ממני כלקוח בשנתיים הקרובות? לכן, מודל ה-CLV הוא מודל מאוד חשוב, כי על בסיסו החברה יודעת להחליט האם לשמר אותי כלקוח או לא. ישנם מקומות, למשל בעולם הבנקאות, שבהם אנו רוצים לדעת בשלב מאוד מוקדם האם הלקוח הזה הוא לקוח שאנחנו רוצים להעביר אותו למחלקת בנקאות פרטית, לבנקאות אישית או לבנקאות עסקית. ואת כל זה אומר לנו מודל ה-CLV, שכאמור כיום מאוד חברות מתעניינות בו ומיישמות אותו באמצעות מודל של הרגרסיה.

דוגמה למתחילים: כך מחשבים ערך חיי לקוח *

כשמתחילים על הערך ארוך-הטווח מלקוח (Customer Lifetime Value - CLV או בקיצור CLV) ניתן להשתמש בנוסחה הבאה:

$$CLV = \frac{P}{1+d-R}$$

P הוא הרווח לתקופה (חודש, רבעון, שנה)
d הוא מקדם ההיוון לתקופה (פרמטר שמביא בחשבון את ערך הזמן של כסף)
R הוא אחוז השימור לתקופה (הסיכוי שלקוח קיים יישאר עד לתקופה הבאה) למשל, אם הרווח השנתי (הכנסות פחות הוצאות) מלקוח הוא 100 שקל, מקדם ההיוון הוא 10% לשנה, והסיכוי שלקוח קיים יישאר לשנה הבאה הוא 80%, ערך חיי הלקוח הוא:

$$CLV = \frac{100}{1+0.1-0.8} = 333 \text{ שקלים}$$

* דוגמה פשטנית - מודל ערך חיי לקוח מלא, שמביא בחשבון את כלל נתוני הפירמה, הוא מורכב יותר. כאשר מדובר בלקוח פוטנציאלי, יש להפחית את עלויות הרכישה כדי להבין כמה הוא שווה

מתוך מאמרם של ה"ר"ח שלומי שוב ופרופ' ברק ליבאי בדה-מרקר: "מאמזון ועד פפר: כך צריך להעריך שווי של חברות בעידן הדיגיטלי", 3 בדצמבר 2017.

אז מודל ה-CLV הוא דוגמא אחת לבעיה ממשפחת הרגרסיה כי אנו מנסים לחזות מחיר של משהו. תוכל לתת דוגמא נוספת?

דוגמא נוספת לבעיה ממשפחת הרגרסיה היא מדד ה-TTF (זמן ההגעה לכשל, Time To Failure). נניח שיש מכונה מסוימת והמכונה הזו מייצרת קבצי יומן (Logs) ואני רוצה על בסיס אותם קבצי יומן לדעת מתי המכונה הזו הולכת להתקלקל. אז תחשוב לרגע שקבצי היומן הללו מתארים כל מיני דברים שקרו, כל מיני אירועים שקרו במכונה, כל מיני חיישנים שקרו במכונה ולאט לאט מצטברים נתונים שאומרים מתי היו תקלות במכונה. ואם אני מנתח אותם נכון אז אני למעשה יכול להגיד, על פי קבצי יומן הרישום שאני רואה עכשיו, הרכיב הזה והזה במכונה הולך להתקלקל בעוד 3 ימים. החשיבות של זה היא שאני יכול הלכה למעשה לפתור בעיות במכונה עוד לפני שהן קורות. אני יודע להחליף רכיב לפני שהוא בכלל התקלקל. זה מה שנקרא "טיפול מנע" ולא "טיפול שבר" וכמובן שיש לכך חשיבות כלכלית עצומה. תחשוב על מקומות כמו חדרי שרתים, שיש בהם אלף שרתים ואני רוצה על פי קבצי היומן של השרתים הללו לדעת מתי שרת מסוים עומד להתקלקל. הלא ברור לך שהרבה יותר טוב להחליף אותו לפני שהוא מתקלקל ולא אחרי שהוא מתקלקל.

מהי בעיית סיווג?

בעיית סיווג (Classification) היא בעיה שבה המודל שלנו מנסה לסווג תצפית מסוימת לאחת מכמה קטגוריות. מודל הסיווג רוצה להגיד לנו התצפית הזו היא צהובה, התצפית הזו היא ורודה, התצפית הזו היא סגולה. למעשה יש לי מספר סופי של אפשרויות שהמודל יכול לבחור מביניהן והמטרה שלי היא לשייך כל תצפית לקטגוריה מסוימת (צבעים בדוגמא שלי).

אז אלו הן בגדול שלוש סוגי הבעיות הגדולות במדע נתונים?

כן. בנוסף, אני רוצה להזכיר את בעיות ממשפחת האפליקציות שמאוד מאוד נפוצות היום. דוגמא אחת לבעיה ממשפחת האפליקציות נקראת מערכות המלצה. הסיבה שמערכות המלצה אינן עומדות בפני עצמן נעוצה בכך שהן עושות שימוש בכלים הן ממשפחת הרגרסיה, הן ממשפחת הסיווג והן ממשפחת ניתוח האשכולות. ולכן על פי רוב ראשית ממפים את שלוש המשפחות הגדולות ורק לאחר מכן מדברים על בעיות ממשפחת האפליקציות שמשמשות בכל הכלים הללו.

מהן מערכות המלצה? אני מודה שמעולם לא שמעתי עליהן.

אני אסביר לך למה אני בטוח שדווקא אתה כן מכיר מערכות המלצה מחיי היומיום שלך. כאשר אתה גולש באתר Ynet בסוף הכתבה מציעים לך "אם התעניינת בכתבה הזאת, אולי תעניין אותך גם הכתבה הזאת", זוהי מערכת המלצה. לחילופין, כאשר אתה גולש באתר של אמזון, אתה מסתכל על איזשהו ספר ומיד למטה כתוב לך "מי שקנה את הספר הזה, קנה בסוף גם את הספר הזה", זוהי גם מערכת המלצה. כלומר, מאחורי מערכות המלצה הללו ישנן הרבה מאוד לוגיקות. סוג אחד של מערכות המלצה היא הצעה למוצר חלופי. למשל, אם המחשב רואה שאתה מסתכל עכשיו על ספר מסוים, אז הוא מציע לך לקנות ספר אחר שהוא למעשה ספר חלופי – כי המחשב יודע שאתה לא תקנה את שני הספרים, אלא רק אחד מהם. סוג אחר של מערכות המלצה היא הצעה למוצר משלים. למשל, הסיפור של טרגט עם הבירה והחיתולים.

מה זה הסיפור הזה של טרגט עם הבירה והחיתולים?

הנהלה של טרגט, רשת הכלבו הגדולה בארה"ב, הגיעו למסקנה שכדאי למקם בחנויות שלהם את הבירות ממש ליד החיתולים. הסיבה לכך נעוצה במחקר סטטיסטי שהם קנו, אשר ממנו עלה שהרבה מאוד גברים שמגיעים לקנות חיתולים באמצע הלילה מחפשים גם לקנות בירה. אלו למעשה שני סוגים שונים של מערכות המלצה שמשמשות בלוגיקות שונות.

זה מעניין מאוד. יש לך דוגמאות נוספות לבעיות ממשפחת האפליקציות?

כן. דוגמא נוספת לבעיה ממשפחת האפליקציות נקראת זיהוי אנומליות. זיהוי אנומליות היא אחת הבעיות הקשות ביותר להיות וזוהי בעיה ששייכת לתחום הלמידה ללא השגחה. נסביר שאנומליה זה משהו שאתה לא יודע שהוא אנומליה עד שלא זיהית אותו כאנומליה. לכן בהרבה מאוד מהבעיות של זיהוי אנומליות אין לנו בכלל מידע מוקדם לגבי מהי אותה אנומליה. תחשוב למשל על זיהוי אנומליות בתקשורת פנימית של רכבים. יש לך המון המון תקשורת תקינות בתוך הרכב והרכיבים של הרכב מתקשרים אחד עם השני. ועכשיו אתה צריך לזהות בתוך כל התקשורת המורכבת הזאת, משהו שלא מסתדר לך או שלא בדיוק מאפיין את התקשורת הזאת. האם זה אומר שאיזשהו האקר פרץ לרכב? אולי, ואולי לא. אבל אין לך מושג מה זה יכול להיות ואתה מתחיל לחקור אלו סוגים של אנומליות ישנם ואתה מתחיל להפעיל כל מיני מודלים רלוונטיים לזיהוי אנומליות.

אותי הכי מעניין ניתוח שפה טבעית וניתוח טקסט (Natural Language Processing and Text Analytics). האם גם זה דוגמא לבעיה ממשפחת האפליקציות?

בול. תיארתי כרגע דוגמא נפוצה מאוד לבעיה ממשפחת האפליקציות המכונה Text Analytics & NLP. האמת שלא רציתי להיכנס אליה אבל אתה שאלת. תראה, בגדול, מדובר על ניתוח של שפה טבעית. נניח שאתה קורא טקסט ואתה רוצה לדעת האם מי שכתב אותו כועס או לא. לחילופין, אתה מקבל שתי עבודות של סטודנטים, טקסטים, ואתה רוצה לדעת האם אחד מהם העתיק מהשני או האם הם כתבו אותם ביחד. עכשיו על פי רוב,

סטודנטים הם לא מפגרים ולכן כאשר הם לוקחים טקסט של חבר שלהם הם משנים את הניסוחים. אז תחשוב שעכשיו יש לך משימה לקחת שני טקסים שעושים שימוש במילים שונות ולבדוק האם שני הטקסטים הללו אומרים בדיוק את אותו דבר, או לא. אז זה למשל מדוע משימת NLP היא כל כך מעניינת. לבוא ולהגיד מהי רמת הקרבה בין שני טקסטים. דברים מהסוג הזה, תרגומים ועוד.

הנכדה שלי ביקשה לשאול אותך על זיהוי תמונות (Image Recognition) וזיהוי אובייקטים בתמונה (Object Recognition)?

עיבוד תמונה הוא תחום חדש יחסית בעולם למידת המכונה מאז הופעת הלמידה העמוקה (Deep Learning). הלמידה העמוקה הביאה עמה כל מיני יכולות מאוד מאוד חדשות. אני רק אומר שעיבוד תמונה זה משהו שקיים כבר המון שנים אבל הכלים של למידת מכונה התחילו להיכנס לתחום הזה רק בתחילת שנות העשרה של המאה ה-21.

דיברנו עד כה על יישומים של מדע נתונים. שאלה קצת אחרת, האם אתם ב"שווי פנימי" מבצעים גם חישובי שווי הוגן של מכשירים פיננסיים, לרבות נגזרים משובצים (IFRS)?

הצגת שוויים של מכשירים פיננסיים במונחי שווי הוגן היא אחת הדרישות המרכזיות של התקינה הבינלאומית (IFRS). דרישה לחישוב שווי הוגן מופיעה, בין היתר, ב-IFRS 7 וב-IFRS 9. לצד חישובי שווי הוגן של מכשירים פשוטים כגון איגרות חוב, פרוורדים ואופציות ונילה, אנו ב"שווי פנימי" מבצעים הערכות שווי של מכשירים אקזוטיים ולא סטנדרטיים. תמחור מכשירים אלו כרוך בניית מודלים ייחודיים, המתבססים על טכניקות נומריות מתקדמות. כבר לפני כמעט עשור פיתחנו ב"שווי פנימי" מודל ממוחשב לתמחור איגרות חוב להמרה – אחד המכשירים המורכבים לתמחור.

אז אתם ב"שווי פנימי" גם מפתחים מודלים להערכת שווי מכשירים מורכבים במסגרת IFRS?

על פי כללי התקינה החשבונאית הבינלאומית IFRS, יש לחשב שווי הוגן של מכשירים נגזרים הגלומים במכשירים פיננסיים אחרים. דוגמא לכך היא אופציית המרה הגלומה באיגרות חוב להמרה. איגרת חוב להמרה הינה מוצר מורכב, המאופיין בתכונות רבות המקשות על תמחורה. וכן בנוסף, הפרמטרים הדרושים לתמחור אינם חד משמעיים (לדוגמה מרווח סיכון אשראי באג"ח לא מדורגת). אנו ב"שווי פנימי" מבצעים הערכת שווי של נכסים מורכבים שונים לפי מספר מודלים המקובלים בתחום. בין היתר, אנו משתמשים במודל של Tsiveriotis & Fernandes (1998) לתמחור איגרות חוב להמרה. מודל זה ניתן ליישום הן על ידי שימוש ב-Finite Difference Method והן באמצעות שימוש בעצים בינומיים/טרינומיים. מדובר במודל הנפוץ ביותר בעולם ועל פי מחקרים הוא מספק את התוצאות הקרובות ביותר למחירי השוק.

מודלים כל כך מורכבים טומנים בחובם סיכון מודל, לא כך?

אתה צודק ולכן השימוש במספר מודלים בתהליך החישוב מאפשר לנו ב"שווי פנימי" לבצע ניתוח השוואתי של התוצאות טרם הגשת חוות הדעת האקטוארית שלנו ללקוח, וכל זאת על מנת להקטין את סיכון המודל. לפיכך, חוות הדעת שלנו מורכבת מארבעה פרקים: (1) רקע (תיאורטי); (2) תיאור המודלים לתמחור; (3) תיאור הפרמטרים שבהם נעשה שימוש; ו- (4) ניתוחי רגישות לתוצאות בהתאם לרמות שונות של פרמטרים, אשר לגביהם לא קיימת וודאות מלאה (כגון: מרווח סיכון אשראי, סטיית תקן של המניה וכיוצא באלה פרמטרים).