דרור יצחק | צילום: עמי ארליך ||

רואה חשבון, על למידת מכונה כבר שמעת?

למידת מכונה היא ענף של בינה מלאכותית שבה הבינה נוצרת באמצעות למידה מתוך דאטה (Big Data). יצאתי לראיין את רועי פולניצר על למידת מכונה

האקטואר רועי פולניצר בעל ניסיון עשיר במתן שירותי יעוץ אקטוארי, מימוני וכלכלי ובביצוע מחקרים בתחומי הבנקאות ושוק ההון, מייסד מנכ"ל לשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (IAVFA). עמד בראשות הוועדה לקביעת מדיניות לביצוע, פיקוח וניהול הערכות שווי של תאגידים, נכסים בלתי מוחשיים ומכשירים פיננסיים מורכבים עבור רשות המסים בישראל. כמו כן מייסד הקבוצה שזכתה במכרז של רשות המסים בישראל לביצוע הערכות שווי של נכסים בלתי מוחשיים בעסקאות מקרקעין והערכות שווי בנושא שינוי מבנה עסקי. בשנה האחרונה רועי התחיל לכתוב על תחום מדע הנתונים ולמידת המכונה ולכן מצאתי לנכון לראיין אותו בנושא.

מה לך ולמדע נתונים ולמידת מכונה, אתה בכלל אקטואר ומעריך שווי?

אני אכן אקטואר מלא (Fellow) בלשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (F.IL.A.V.F.A.) ומעריך שווי תאגידים, נכסים בלתי מוחשיים ומכשירים פיננסיים מורכבים אבל בין השנים 2006-2010 שימשתי כעוזר מחקר של ד"ר שילה ליפשיץ ז"ל. במסגרת תפקידי, בניתי מאגרי נתונים גדולים Big Data ועשיתי שימוש מתקדם בדאטה לצורך בניית מודלים, יישום ופיתוח אלגוריתמים (שהיום נקראים Machine Learning) על מנת להעריך את הסיכונים והרווחיות של חמשת הבנקים הגדולים בישראל כמו גם לזהות כיוונים ומגמות בנתונים שלהם. במילים אחרות, ביצעתי מחקרים אמפיריים תוך שימוש בשיטות אקונומטריות שזה בדיוק מה שמדעני נתונים עושים כיום, רק שאז קראו להם אקונומטריקאיים.

האם תוכל להסביר למי שלא חי את התחום כמוך מה ההבדל בין למידת מכונה לבין בינה מלאכותית?

למידת מכונה היא ענף של בינה מלאכותית שבה הבינה נוצרת באמצעות למידה מתוך מאגרי נתונים גדולים. למידת מכונה מחלוקת בגדול לארבע קבוצות: למידה ללא השגחה (Unsupervised Learning), למידה בהשגחה (Supervised Learning), למידה בהשגחה למחצה (Semi-Supervised Learning) ולמידה בחיזוקים (Reinforcement Learning).

האם תוכל לתת שתי דוגמאות לתחזיות המבוצעות במסגרת למידה בהשגחה?

דוגמא אחת לתחזית שמתקבלת מלמידה בהשגחה היא אמידת ערכו של משתנה רציף (כגון: שערי ריבית, שערי חליפין, מחירי ניירות ערך, מדדי מחירים, מחירי סחורות ועוד). דוגמא אחרת היא סיווג לקבוצות (לדוגמא דירוג אשראי).

למה למשל משמשת למידה ללא השגחה?

למידה ללא השגחה משמשת לזיהוי דפוסים (ניתוח אשכולות, Clusters) בנתונים.

למה משמשת למידה בחיזוקים?

למידה בחיזוקים משמשת במצבים שבהם רצף החלטות צריך להתקבל בסביבת מידע משתנה.

למה משמשת למידה בהשגחה למחצה?

למידה בהשגחה למחצה משמשת ליצירת תחזיות כאשר לחלק מהנתונים הזמינים יש ערכים עבור ה- Target (המשתנה שאותו אנו רוצים לחזות) ולחלק לא.

מקריאת המאמרים שלך הבנתי שאחת הבעיות בתחום למידת המכונה היא התאמת יתר (Over-Fitting). איך אתה יודע לומר האם מודל מסוים של למידת מכונה מתאים יתר על המידה לנתונים, אם לא?

אם מודל מסוים אשר פותח על בסיס סט אימון (Training Set) מסוים איננו מכליל בצורה טובה את סט התיקוף (Validation Set, רוצה לומר שהתחזיות שאותו מודל מייצר על בסיס סט התיקוף הן הרבה יותר גרועות מאלו שהוא מייצר על סמך סט האימון) אז קיימת בעיה של התאמת יתר.

הזכרת סט אימון וסט תיקוף אבל לא הזכרת את סט הבדיקה, האם תוכל להסביר את תפקידם של סט התיקוף וסט הבדיקה?

סט התיקוף משמש להשוואה בין מודלים על מנת לבחור את המודל בעל רמת הדיוק הטובה ביותר ושמכליל בצורה הטובה ביותר את נתוני סט התיקוף. סט הבדיקה נשמר בצד על מנת לספק מבחן סופי לרמת הדיוק של המודל הנבחר.

מה פירוש מאפיין קטגורי?

מאפיין (Feature) בטרמינולוגיה של למידת מכונה שקול אפקטיבית מכל הבחינות המהותיות למשתנה מסביר בטרמינולוגיה של אקונומטריקה. למעשה מדובר במשתנה אשר לגביו יש לנו תצפיות. מאפיין קטגורי (Categorical Feature) הוא מאפיין לא נומרי כאשר נתונים מוקצים או מיוחסים לאחד ממספר קטגוריות.

ממה שהבנתי קיימים חמישה סוגים של ניקוי נתונים (Data Cleaning), האם תוכל למנות אותם?

ניקוי נתונים מתחלק לחמש קבוצות: 1) תיקון רישום לא עקבי (Inconsistent Recording); 2) הסרת תצפיות שאינן רלוונטיות; 3) הסרת תצפיות כפולות; 4) טיפול בחריגים; ו- 5) טיפול בנתונים חסרים.

ראיתי באחד המאמרים שלך שכתבת ש- "נוסחת בייס (Bayes Theorem) מאפשרת להפוך משהו להתניה", אתה מוכן להסביר למה התכוונת?

נוסחת בייס מטפלת במצב שבו אנו יודעים מהי ההסתברות המותנה לקרות מאורע X מותנה בכך שמאורע Y קרה ואנו רוצים לדעת מהי ההסתברות המותנה לקרות מאורע Y מותנה בכך שמאורע X קרה. לשם הדוגמא, נניח ש- 25% מהאימיילים הם ספאם ונמצא כי ספאם כולל בתוכו מילה מסוימת (נניח "רוצה") 40% מהזמן.

כעת בוא נניח, שבסך הכל 12% מהאימיילים כוללים את המילה "רוצה". נשאלת השאלה מהי ההסתברות שאימייל מסוים הוא ספאם אם ידוע שהוא מכיל בוודאות את המילה "רוצה"? אז על פי נוסחת בייס ההסתברות המותנה שהמילה "רוצה" תופיע במייל מסוים מותנה בכך שידוע שאותו מייל הוא ספאם היא 40%, ההסתברות שמייל מסוים הוא ספאם היא 25% וההסתברות שהמילה "רוצה" תופיע במייל מסוים היא 12%. לפיכך, ההסתברות המותנה שאימייל מסוים הוא ספאם מותנה בכך שידוע שהוא מכיל את המילה "רוצה" היא 83.33% (40% כפול 25% חלקי 12%). לאמור- ישנו סיכוי של 83.33% שאימייל שכולל את המילה "רוצה" הוא ספאם.

כמדען נתונים, איזה מודל עדיף יותר לצורך ניבוי או ביצוע תחזיות, מודל מסדר גבוה (מודל מורכב) או מודל מסדר נמוך (מודל פשוט יחסית)?

לאחרונה פיתחתי באחד מהמאמרים שלי מודל לניבוי משכורת שנתית על סמך הגיל של מקבל המשכורת. אז לקחתי סט אימון של 10 עובדים באותו מקצוע ובאותו אזור אך בגילאים שונים ועל סמך סט האימון בניתי מודל ניבוי מסוג פולינום מסדר חמישי. למי שלא מכיר רמת הדיוק של של מודל ניבוי נמדדת על ידי פרמטר שנקרא סטיית התקן של השגיאה (Standard Deviation of Error). סטיית התקן של השגיאה של מודל הפולינום מסדר חמישי נאמדה על ידי בכ- 12,902 עבור סט האימון ו- 38,794 עבור סט הבדיקה, בהתאמה. על מנת לבדוק שמודל הפולינום מסדר חמישי שבניתי באמת מייצר תחזיות טובות בניתי על סמך אותו סט אימון מודל ניבוי נוסף, הפעם מסוג פולינום מסדר שני. סטיית התקן של השגיאה של מודל הפולינום מסדר שני נאמדה על ידי בכ- 32,932 עבור סט האימון ו- 33,554 עבור סט הבדיקה,

האם בדקת מה קורה עבור המודל הכי פשוט, המודל הלינארי או מה שאתה וודאי קורא לו מודל פולינום מסדר ראשון?

סטיית התקן של השגיאה של מודל הפולינום מסדר ראשון נאמדה על ידי בכ- 49,731 עבור סט האימון ו- 49,990 עבור סט הבדיקה, בהתאמה. אני לומד מכך שלושה דברים. הדבר הראשון הוא שמודל הפולינום מסדר חמישי, במקרה דנן שלפנינו, מתאים יתר על המידה (Overfits) לנתוני סט האימון ועל כן הוא פחות מצליח לייצר תחזיות טובות (כפי שעולה מישומו על נתוני סט הבדיקה). הדבר השני הוא שהמודל שעליו אתה שאלת, מודל הפולינום מסדר ראשון (המודל הלינארי)' במקרה דנן שלפנינו, מתאים חסר על המידה (Underfits) לנתוני סט האימון מאחר והוא מייצר את התחזיות הגרועות ביותר מבין שלושת המודלים (כפי שעולה מישומו על נתוני סט הבדיקה). הדבר השלישי הוא שמודל הפולינום מסדר שני (המודל הריבועי), במקרה במקרה דנן שלפנינו, הוא ככלל הנראה המודל הטוב ביותר מבין שלושת המודלים וזאת מאחר והוא מייצר את התחזיות הטובות ביותר מבין השלושה (כפי שעולה מישומו על נתוני סט הבדיקה).

האם בדקת מהן התוצאות עבור פולינומים מסדר שלישי ורביעי, לשם השלמת התמונה כמובן?

סטיית התקן של השגיאה של מודל הפולינום מסדר שלישי נאמדה על ידי בכ- 31,989 עבור סט האימון ו- 36,986 עבור סט הבדיקה, בהתאמה, בעוד שסטיית התקן של השגיאה של מודל הפולינום מסדר רביעי נאמדה על ידי בכ- 21,824 עבור סט האימון ו- 32,932 עבור סט הבדיקה, בהתאמה. מה ניתן ללמוד מכך? ניתן לראות שככל שמודל הפולינום הוא מסדר גבוה יותר, כך אנו מקבלים רמת דיוק גבוהה יותר עבור סט האימון אבל יש הבדל גדול מאוד בין ביצועי המודל עבור סט האימון לבין ביצועי המודל עבור סט הבדיקה.