

מאיקס עיגול ושחמט ועד תרגום שפות – למידת מכונה

למידת מכונה הופכת יותר ויותר לכלי חשוב בעולמות ההשקעות, הבנקאות, הביטוח והפנסיה – עד כדי כך שכמעט כל העובדים צפויים להיות מושפעים ממנה בצורה כזו או אחרת במהלך השנים הבאות. למידת מכונה עושה שימוש במאגרי נתונים גדולים Big Data על מנת ללמוד על קשרים בין משתנים, לבצע תחזיות ולפעול באינטרקציה עם סביבה משתנה. מאגרי נתונים גדולים על משתנים המתארים רכישות של צרכנים, תנועות של מחירי מניות והרבה היבטים אחרים של עסקים אינם חדשים כיום. השיפורים הטכנולוגיים האחרונים כגון העליה במהירויות המחשב והירידה בעלויות אחסון הנתונים מאפשרים לנו להגיע למסקנות מתוך מאגרי נתונים גדולים בדרכים שפשוט לא היו אפשריים לפני 20 או 30 שנה.

למידת מכונה היא ענף של בינה מלאכותית (AI). בינה מלאכותית מפתחת דרכים שבהם למידת מכונה יכולה לחקות בינה אנושית, ואפילו לחקות אותה. למידת מכונה יוצרת בינה על ידי למידה מתוך נפחים גדולים של נתונים. למידת מכונה היא ללא ספק ההתפתחות המרגשת ביותר בתוך הבינה המלאכותית והיא זו שיש לה את הפוטנציאל הרב ביותר לשנות עסקים.

על מנת להסביר כיצד למידת מכונה נבדלת מגישות אחרות של בינה מלאכותית ניקח את המשימה הפשוטה ביותר של לתכנת מחשב לשחק איקס עיגול. גישה אחת היא לספק למחשב טבלת בדיקות המציגה את הפוזיציות שיכולות לקרות ואת המהלך שהיה מבוצע על ידי שחקן אנושי מומחה עבור כל אחת מאותן פוזיציות. גישה אחרת היא להציג למחשב מספר גדול מאוד של משחקים (למשל, על ידי כך שניתן למחשב לשחק מול עצמו אלפי פעמים) וניתן לו ללמוד את המהלך הטוב ביותר. הגישה השנייה הינה יישום של למידת מכונה. למעשה ניתן להשתמש בהצלחה בכל אחת מהגישות שתוארו לצורך משחק פשוט כמו איקס עיגול. גישות של למידת מכונה הוכחו ככאלה שעובדות מצויין עבור משחקים יותר מורכבים כמו שחמט וגו (למי שלא מכיר מדובר במשחק אסטרטגיה מופשט שמקורו בסין) כאשר הגישה הראשונה בבירור אינה אפשרית.

אחת הדוגמאות הטובות לעוצמה של למידת מכונה היא תרגום שפה. כיצד ניתן לתכנת מחשב לתרגם בין שתי שפות, נגיד מעברית לגרמנית? רעיון אחד שעולה לי לראש הוא לתת למחשב מילון עברי-גרמני. למרבה הצער תרגום מילה-למילה מספק תוצאות גרועות מאוד ולכן הכרחי לנסות ולתכנת גם את כללי הדקדוק העברי והדקדוק הגרמני. זוהי מלאכה ממש לא פשוטה וגם אחרי שהיא נעשית הרי שתוצאותיה עדיין רחוקות מלהיות מושלמות. גוגל היא חלוצה בשימוש באלגוריתמים של למידת מכונה לצורך תרגום. גוגל הכריזה על כך בנובמבר 2016 והיא קוראת לטכנולוגיה החדשה בשם GNMT (תרגום המכונה העצבי של גוגל, Google Neural Machine Translation). במסגרת

טכנולוגיית ה-GNMT נותנים למחשב מיליון דפים של חומר שתורגם על ידי מומחי תרגום מעברית לגרמנית. המחשב לומד את החומר ומפתח כללי תרגום משלו. התוצאות מטכנולוגיית ה-GNMT מהוות שיפור גדול בהשוואה לגישות הקודמות.

מדע נתונים (Data Science) הוא תחום שכולל בחובו למידת מכונה ולעיתים נחשב לרחב יותר וכולל גם משימות כמו למשל קביעת מטרות, יישום והטמעת מערכות ותקשורת עם בעלי עניין. נציין כי יש שרואים במונחים "למידת מכונה" ו-"מדע נתונים" כמילים נרדפות, מאחר וקשה לראות כיצד מומחי למידת מכונה יכולים להיות יעילים לארגון כלשהו אם רק עושים שימוש בדאטה (Big Data) באמצעות בניית מודלים, פיתוח ושימוש באלגוריתמים וניתוח תהליכים לצורך זיהוי כיוונים ומגמות במגוון תחומים לרוחב הארגון מבלי: (1) להיות אחראים על זיהוי אתגרים עסקיים בהם DATA יכול להוות גורם מכריע בשיפור קבלת החלטות; (2) להיות אחראים על איתור ואיסוף מקורות מידע פנים ארגוניים וחיצוניים, הגדרה ואיפיון של שימושי המידע בארגון; (3) לבנות מאגרי המידע שאינם מובנים ונמצאים במסגרת הארגון, אפיון והגדרת הצגת המידע ותוצריו לדרג מקבלי ההחלטות בארגון; (4) לפתח כלים, מודלים, תהליכים ומערכות לרוחב הארגון בתחום האנליזה; ו-(5) לנהל מעקב אחר התקדמות הפרויקטים בהתאם לתכנית העבודה אל מול מנהלי התחומים והעובדים. רוצה לומר- אף ארגון לא באמת יעסיק מומחי למידת מכונה בשביל שאלו רק יעבדו עם מאגרי נתונים גדולים תוך פיתוח ושימוש באלגוריתמים של למידת מכונה על מנת לזהות כיוונים ומגמות בעולמות תוכן שונים.

ניתן לתאר את למידת המכונה או מדע הנתונים כעולם החדש של סטטיסטיקה. הסטטיסטיקה המסורתית נוגעת לנושאים כמו הסתברויות, התפלגויות, רווחי סמך, מבחני מובהקות ורגרסיה לינארית. נאמר מראש ידע בנושאים הללו הוא חשוב, אבל כיום אנו יכולים ללמוד מתוך מאגרי נתונים גדולים בדרכים שלא היו אפשריות לפני כן. לדוגמא, אנו יכולים לפתח מודלים לא-לינאריים לניבוי ושיפור קבלת החלטות, לחילופין אנו יכולים לחפש דפוסים בנתונים על מנת לשפר את ההבנה של החברה את לקוחותיה והסביבה שבה היא פועלת או לחילופין חילופין אנו יכולים לפתח כללי החלטה כאשר אנו פועלים באינטרקציה עם סביבה משתנה.

כאמור, יישומים אלו של למידת מכונה אפשריים היום הודות לשיפורים הטכנולוגיים האחרונים כגון העליה במהירויות המחשב והירידה בעלויות אחסון הנתונים.

ונחתום בעדות אישית. כסטטיסטיקאי וכאקונומטריקאי בעת טבילת האש הראשונה שלי בתחום למידת המכונה נחשפתי לטרמינולוגיה שנראתה לי משונה. לדוגמא, אני כסטטיסטיקאי וכאקונומטריקאי מדבר על משתנים בלתי תלויים ומשתנים תלויים בעוד שמדעני נתונים מדברים על

מאפיינים (Features) ויעדים (Targets). למעשה למדתי שפה חדשה ואני לא מתכוון ל- Python ולספריות הסטטיסטיות שלה כמו NumPy (שעובדת עם מערכים וקוראת קבצי טקסט), Matplotlib (שיוצרת תרשימים), Scipy (שמבצעת אופטימיזציה ופותרת משוואות א-לינאריות) ו- Pandas (תיקיית האלגוריתמים של למידת מכונה).



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.