

סיכום מבוא ללמידת מכונה

למידת מכונה היא ענף של בינה מלאכותית העוסקת בלמידה מתוך מאגרי מידע גדולים (Big Data). למידת מכונה כרוכה בפיתוח וגיבוש אלגוריתמים לצורך ניבוי, סיווג נתונים (Cluster Data), או לקבלת סדרת החלטות רצופות (Sequential Decisions) תוך אינטראקציה אופטימלית עם הסביבה.

ניתוח סטטיסטי עוסק באופן מסורתי ביצירת השערות (מבלי להסתכל על הנתונים) ולאחר מכן בבדיקת ההשערות באמצעות נתונים. למידת מכונה שונה מניתוח סטטיסטי בכך שהיא לא יוצרת השערות אלא גוזרת את המודל לחלוטין מתוך הנתונים.

היבט חשוב של למידת מכונה הינו תיקוף ובדיקה (Validation and Testing). רוצה לומר שיש לבדוק ולתקף מודלים שנוצרו באמצעו אלגוריתמים של למידת מכונה, בעזרת סט נתונים אחר, כזה שלא שימש ליצירת המודל (Out-of-Sample). מחד גיסא, מודל מורכב מדי עשוי להתאים את עצמו יתר על המידה (Over-fit, ללמוד יותר מדי טוב את) לנתונים ששימשו לאימון המודל ובכך הוא עלול שלא להצליח להכליל (Generalize) באופן מספק דיו את הנתונים החדשים. מאידך גיסא, מודל פשוט מדי עלול שלא להצליח לתפוס היבטים חשובים של הנתונים. למידת מכונה גורסת שיש לחלק את הנתונים הזמינים ל- 3 סטים של נתונים. סט האימון (Training Set) משמש לגיבוש/פיתוח מודלים אלטרנטיביים. סט התיקוף (Validation Set) משמש לבדיקה עד כמה המודלים מכלילים טוב את הנתונים החדשים. סט הבדיקה (Testing Set) נשמר בצד לאורך כל התהליך שתואר עד כה ומשמש כמבחן סבירות סופי לרמת הדיוק של המודל הנבחר.

טרם השימוש באלגוריתם של למידת מכונה, חשוב מאוד לנקות תחילה את הנתונים. המאפיינים (Features, המשתנים המסבירים) המהווים את הנתונים יכולים להיות נומריים או קטגוריים. בכל מקרה עשויים להיות מצבים של חוסר עקביות (Inconsistencies) באופן שבו הנתונים הוכנסו למאגר הנתונים. לפיכך, יש לזהות ולתקן מצבים של חוסר עקביות. חלק מהתצפיות עשויות להיות לא רלוונטיות למשימה הנוכחית ועל כן יש להשמיטן. בנוסף, יש לבדוק שאין תצפיות כפולות או כפילויות בנתונים, דבר שעלול ליצור הטיית. יש להשמיט חריגים אשר נוצרו בוודאות כתוצאה מטעויות הקלדה או מטעויות בהכנסת הנתונים למאגר. לבסוף, יש לטפל בנתונים חסרים באופן שלא יטה את התוצאות.

משפט בייס (Bayes Theorem, נוסחת בייס) הוא תוצאה המשמשת לעתים כאשר היא נדרש לכמת את אי הוודאות. משפט בייס הוא דרך להפוך משהו להתניה. נניח שאנו רוצים לדעת מהי ההסתברות שמאורע Y יתרחש ונניח שאנו גם יכולים לדעת האם מאורע אחר שקשור למאורע Y, נקרא לו מאורע

X, התרחש או לא. עוד נניח שעל סמך ניסיון אנו יודעים את ההסתברות המותנה (Intensity) שמאורע X יתרחש בהינתן שידוע שמאורע Y התרחש. למעשה משפט בייס מאפשר לנו לחשב את ההסתברות המותנה שמאורע Y יתרחש בהינתן שידוע שמאורע X התרחש.

למידת מכונה ישנה טרמינולוגיה משלה אשר שונה מזו המסורתית המשמשת בסטטיסטיקה. במסגרת הטרמינולוגיה של למידת מכונה: מאפיין (Feature) הוא משתנה אשר לגביו יש לנו תצפיות; יעד (Target) הוא המשתנה אשר עליו אנו רוצים לבצע תחזיות; תוויות (Labels) הן תצפיות על היעד; למידה בהשגחה (Supervised Learning) היא תחום של למידת מכונה שבמסגרתה אנו משתמשים בנתונים על המאפיינים והיעדים לצורך ניבוי היעד על סמך נתונים חדשים; למידה ללא השגחה (Unsupervised Learning) היא תחום של למידת מכונה שבמסגרתה אנו מנסים למצוא דפוסים בנתונים על מנת לסייע לנו בהבנת מבנה הנתונים (בלמידה ללא השגחה אין יעד ועל כן אין גם תוויות); למידה בהשגחה למחצה (Semi-Supervised Learning) היא תחום של למידת מכונה שבמסגרתה אנו מבצעים תחזיות על היעד על סמך נתונים אשר לחלקם יש תוויות (קרי, יש להם ערכים של היעד) וליתר אין תוויות (קרי, אין להם ערכים של היעד); למידה בחיזוקים (Reinforcement Learning) היא תחום של למידת מכונה שבמסגרתה אנו יוצרים אלגוריתמים לקבלת סדרת החלטות רצופות כאשר מקבל ההחלטה פועל בתוואי של סביבה משתנה.



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.