

למידה בהשגחה (Supervised Learning): עץ החלטה (Decision Tree)

עץ החלטה הינו אלגוריתם לסיווג או לניבוי ערכו של משתנה. המאפיינים מסודרים לפי סדר החשיבות. לצורך סיווג, קיימים שני מדדים אלטרנטיביים לאי וודאות: מדד אנטרופי ומדד ג'יני. כאשר ערכו של משתנה נחזה, אי הוודאות נמדדת באמצעות השגיאה הריבועית הממוצעת (MSE- Mean Squared Error) של התחזית. חשיבותו של מאפיין הינו רווח המידע הצפוי שלו (Expected Information Gain). רווח המידע הצפוי נמדד על ידי הירידה באי הוודאות הצפויה אשר תתרחש כאשר יתקבל ידע אודות המאפיין.

במקרה של מאפיינים קטגוריים, הידע המתקבל הינו על פי רוב אודות התווית (Label) של המאפיין (למשל, האם יש בבעלותו של לווה פוטנציאלי מסוים בית או שמא הוא שוכר דירה). במקרה של מאפיין נומרי יש לקבוע ערך סף (Threshold) אחד (או יותר) המגדיר שני טווחים (או יותר) עבור ערך המאפיין. ערכי סף אלו נקבעים באופן שממקסם את רווח המידע הצפוי.

אלגוריתם עץ ההחלטה קובע תחילה את צומת השורש (Root Node) האופטימלי של העץ באמצעות מקסום "קריטריון רווח המידע" שכרגע הגדרנו. לאחר מכן, אלגוריתם עץ ההחלטה ממשיך לעשות אותו הדבר עבור הצמתים העוקבים. הקצוות של הענפים הסופיים של העץ מכונים צמתי עלים (Leaf Nodes). כאשר העץ משמש לצורך סיווג, הרי שצמתי העלים מכילים את ההסתברויות של כל אחת מהקטגוריות להיות הנכונה. כאשר ערך נומרי נחזה, או אז צמתי העלים נותנים את הערך התוחלתי של היעד. הגיאומטריה של העץ נקבעת באמצעות סט האימון (Training Set), אך הסטטיסטיקה שעוסקת ברמת הדיוק של העץ צריכה (כמו תמיד בלמידת מכונה) לבוא מתוך סט הבדיקה (Test Set) ולא מתוך סט האימון.

פרטים אודות כותב המאמר: האקטואר רועי פולניצר, FRM

רועי בעל תואר שני במימון (התמחות בניהול סיכונים ואקטואריה) ותואר ראשון בכלכלה (התמחות במימון), שניהם מאוניברסיטת בן-גוריון בנגב, בעל דיפלומה בניהול סיכונים פיננסיים (FRM®) מאוניברסיטת אריאל בשומרון ולמד בתוכנית ללימודי תעודה באקטואריה באוניברסיטת חיפה. כמו כן, רועי אקטואר מלא



(Fellow) בלשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (F.I.L.A.V.F.A.), מוסמך כמעריך שווי מימון תאגידי (CFV) מטעם לשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (IAVFA), מוסמך כמנהל סיכונים פיננסיים (FRM) מטעם האיגוד העולמי למומחי סיכונים (GARP) ומוסמך כמומחה לניהול סיכונים (CRM) מטעם האיגוד הישראלי למנהלי סיכונים (IARM).

לרועי ניסיון של מעל ל-15 שנה בביצוע ניתוחים כמותיים במכשירים פיננסיים, בהערכת שווי תאגידים ונכסים בלתי מוחשיים, באמידה וכימות סיכונים כמו תמותה, אריכות ימים, תחלואה, ביטולים והחלמה מנכות, ובמידול ומדידת סיכוני שוק, אשראי, תפעוליים, מודל, נזילות והשקעות לצורכי יישום הוראות רגולטוריות ותקינה חשבונאית, פיתוח, יישום ותיקוף מודלים בתחומים של הערכות שווי, ניהול סיכונים, אקטואריה והנדסה פיננסית, קביעת תעריפי ביטוח חיים, הערכת פרמיות סיכון והערכת עתודות ביטוח, קביעת עלות תנאי פנסיות (צוברות ותקציביות) והכנת מאזנים אקטואריים לקרנות פנסיה, ניתוח וחזוי מצבים פיננסיים מורכבים וכן העברת סמינרי הדרכה והשתלמויות בתחומי התמחותו: מימון, אקטואריה, הערכות שווי, בנקאות, ניהול סיכונים, אופציות והנדסה פיננסית.

ניסיונו של רועי בתחום ה- Data Analysis, כולל: עבודה עם מאגרי מידע גדולים Big Data תוך שימוש ב- Statistical Learning (כגון: סטטיסטיקה תיאורית, הסתברות, הסקה סטטיסטית, סטטיסטיקה א-פרמטרית, חלוקת נתונים, נרמול נתונים, Fitting ו- Bayes Theorem) ובאלגוריתמים מסוג Unsupervised Learning (כגון: Hierarchical Clustering, k-means Clustering, Density-based Clustering, Distribution-based Clustering ו- Principle Components Analysis) למציאת דפוסים וזיהוי מגמות ואנומליות בעולמות ניהול הסיכונים, ההשקעות, האקטואריה, הביטוח והפנסיה, פיתוח תשתית לצורך ניתוח נתונים, שילוב והטמעת כלים לצורך גישה ושלפה עצמאית של נתונים ממאגרי מידע, פיתוח דוחות, ממשקים ומסכים באמצעות כלי ויזואליזציה.

ניסיונו של רועי בתחום ה- Data Science, כולל: עבודה עם מסדי נתונים גדולים Big Data תוך שימוש באלגוריתמים מסוג Supervised Learning (כגון: Linear Regression, Ridge Regression, Lasso Regression, Elastic Net Regression, Logistic Regression, Maximum Likelihood Estimation, k-Nearest Neighbors, Decision Tree, Random Forest, Ensemble, Bagging, Boosting, Naïve Bayes Classifier, Linear Separation, Support Vector Machine, Non-Linear Separation, SVM Regression, Artificial Neural Network, Convolutional Neural Network ו- Recurrent Neural Network) לניבוי וסיווג בעולמות ניהול הסיכונים, ההשקעות, האקטואריה, הביטוח והפנסיה ובמודלים מסוג Reinforcement Learning (כגון: Q-learning, Monte Carlo, Simulation, Temporal Difference Learning ו- n-Step Bootstrapping) לקבלת החלטות מרובות שלבים בעולמות ניהול הסיכונים, ההשקעות, האקטואריה, הביטוח והפנסיה, זיהוי אתגרים עסקיים שבהם DATA יכול להוות גורם מכריע בשיפור קבלת החלטות, איתור ואיסוף מקורות מידע, הגדרה ואיפיון של שימושי המידע, בניית מסד המידע, אפיון והגדרת הצגת המידע



ותוצרין, פיתוח כלים, מודלים, תהליכים ומערכות בתחום האנליזה, תוך שימוש בכלי אנליזה מתקדמים (EXCEL, VBA ו-R).