האיגוד הישראלי למדעני נתונים מקצועיים
Professional Data Scientists' Israel Association

סיכום למידה בחיזוקים

למידה בחיזוקים (Reinforcement Learning) עוסקת בקבלת סדרת החלטות רצופות (Sequential Decisions). למידה בחיזוקים קשורה לפעולות ומצבים (States). הפעולות המבוצעות מובילות לתגמולים (Rewards) ועלויות. פונקציית ה- Q אומדת את התגמול הצפוי (בניכוי עלויות) מביצוע פעולה פרטנית כאשר הסביבה מתוארת באמצעות מצב פרטני. הפעולה הטובה ביותר במצב פרטני היא זו שעבורה פונקציית ה- Q היא הגדולה ביותר.

היבט חשוב של למידה בחיזוקים הוא התחלופה שבין ניצול (Exploitation) לחקר (Exploration). כאשר לומדים מתוך נתונים מסומלצים (Simulated, מתורחשים) או היסטוריים, אז די מפתה לבצע פעולה שנראית כטובה ביותר בהתבסס על הנתונים שנאספו עד כה. למרות זאת, אם האלגוריתם תמיד עושה זאת (קרי, בוחר את הפעולה שנראית כטובה ביותר בהתבסס על הנתונים שנאספו עד כה), אז הוא למעשה מפסיק ללמוד מאחר והוא לעולם לא מנסה לבצע פעולה חדשה. לפיכך, אלגוריתם של למידה בחיזוקים מקצה הסתברות נמוכה כלשהי של ε לפעולה מסוימת שנבחרת אקראית והסתברות של $1 - \varepsilon$ לפעולה הטובה ביותר שזוהתה עד כה.

מקובל להמחיש את התחלופה שבין ניצול לחקר באמצעות שתי דוגמאות קלאסיות. הדוגמא הראשונה נקראת בעיית ריבוי ידיות של מכונות מזל (Multi-Armed Bandit Problem), בעיה המוכרת היטב בסטטיסטיקה שבה מהמר מנסה ללמוד איזו מתוך מספר ידיות של מכונות מזל (Slot Machines) בקזינו מספקת את התמורה (Payout) הגבוהה ביותר. זוהי דוגמא פשוטה יחסית של למידה בחיזוקים היות והסביבה (קרי, המצב) לעולם לא משתנה. הדוגמא השנייה היא משחק הנים (Game of Nim) שבו המצב מוגדר באמצעות מספר הגפרורים שנותרו והפעולה היא מספר הגפרורים להרמה. בשני המקרים הללו, למידה בחיזוקים מספקת דרך ללמוד את האסטרטגיה הטובה ביותר.

הערך של ביצוע פעולה פרטנית במצב פרטני מכונה ערך ה- Q. קיימות מספר דרכים שונות לעידכון ערכי ה- Q. אחת הדרכים היא לבסס את העידכון על התגמול הכולל נטו (עם או בלי היוון) שבין הזמן הנוכחי ומועד האופק (Horizon Date). דרך אחרת מסתכל רק על פעולה אית קדימה ומבססת את העידכון על הערך שחושב עד כה עבור הימצאות במצב שקיים בזמן הפעולה הבאה. נהלי עידכון נוספים נופלים בין שתי הדרכים הקיצוניות שהוזכרו לעיל כאשר אנו מסתכלים על מספר פעולות קדימה בעת חישוב ההשלכות של פעולה מסוימת.

ברגיל, ביישומי 'עולם אמיתי' של למידה בחיזוקים ישנם מספר רב של מצבים ופעולות. דרך אחת להתמודד עם זה היא להשתמש בלמידה בחיזוקים ביחד עם רשת נוירונים מלאכותית (ANN). למידה בחיזוקים מייצרת את ערכי ה- Q עבור חלק מהקומבינציות של מצב-פעולה וה- ANN משמשת לאמידת פונקציה מורכבת יותר.

<u>פרטים אודות כותב המאמר: מדען הנתונים רועי פולניצר, PDS</u>

- 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.