כאמור קודם לכן, אלגוריתם Q-Learning שניסינו לא הצליח לאור כמות המצבים הגדולה והקושי שלו להכליל (generalize) את הידע הקיים אל למצבים חדשים. גישה מקובלת להתמודד עם "בעיית הכללה" היא Q-Learning המבוסס על Value Function Approximation.

**הסבר על האלגוריתם**

בגישה זו, במקום לשערך את הערך של כתוחלת על בסיס הדגימות שאותם הסוכן ראה, הסוכן מנסה ללמוד את הערך באופן ישיר על בסיס מאפיינים של המצב הנתון. בשימוש בשיטה זו לרוב המשתמש מגדיר פיצ'רים על גבי המצב, אוסף פונקציות שמאכלסות בתוכן ידע מקצועי על הדומיין הספציפי. בהינתן פונקציות כנ"ל, מודל אפשרי אחד הוא לדוגמה:

*כלומר הערך הוא קומבינציה ליניארית של מרחב הפיצ'רים (). מודלים שונים יכולים לכלול פונקציות מורכבות יותר (במקום קומבינציה ליניארית) או אפילו שימוש בלמידה עמוקה כתחליף למציאת פיצ'רים טובים. לצורך הפשטות, בעבודה שלנו הפיצ'רים מוזנים מראש על ידינו (אנחנו מגדירים את בהינתן ) והשתמשנו במודל ליניארי.*

*הגישה הזו מאפשרת לסוכן להכליל בצורה פשוטה: המידע על מספיקים לו כדי לשערך את , ללא תלות בשאלה האם הוא ביקר קודם במצב.*

*בשלב האימון, האלגוריתם* Q-Learning with Value Approximation מנסה למצוא את וקטור המשקולות שייתן את התוצאה הטוב ביותר (מקסום ה-reward). בכל איטרציה מתבצע עדכון של המשקולות על בסיס הכלל הבא (דומה ל-Q-Learning סטנדרטי):

*הכלל הנ"ל דומה ל-*Q-Learning *הסטנדרטי: הוא "קבוע הלמידה" (שמשפיע על קצב העדכון), האיבר הוא שערוך של ה-*target *אותו אנו מנסים ללמוד (יש חלופות אחרות לייצוג האיבר הנ"ל) וה- מייצג הוא למעשה ה-*prediction *הנוכחי שלנו. העדכון הוא למעשה המרחק בין ה-target לבין ה-prediction, מנורמל ע"י הגודל של הפיצ'ר וקבוע הלמידה.*

**בחירת הפרמטרים**

בשלב הראשון בדקנו מספר אפשרויות שונות לפיצ'רים וכן פרמטרים שונים נוספים.

הגדרת ה-

ה-reward שהגדרנו לצורך הבעיה הוא כעת התשואה שהסוכן קיבל במקום הסכום האבסולוטי (כלומר במעבר מ-200 ל-210 ה-reward יהיה 5 בגלל שהשיפור הוא 5%, ולא 10 שהוא ה-reward האבסולוטי). המחשבה שמאחורי ההגדרה הזו היא שכעת הסוכן מנסה ללמוד את פונקציית המטרה באופן ישיר, ולכן אנו רוצים שערך ה-reward יהיה סטציונרי ולא תלוי ברווח הקודם (כלומר המידע על הקרנות בלבד יספיק כדי שהתהליך יהיה מרקוב).

את הקבוע בחרנו להיות 0. זה אומר שבעת ביצוע "שלב העדכון" באלגוריתם, ה-target אותו הסוכן ללמוד יהיה ה-reward ישירות. באופן עקיף, המשמעות היא שהסוכן מנסה ללמוד באופן ישיר את ה-reward – המשקולות האידיאליות הן כאלה שעבורן , כלומר ננסה לשערך את התשואה העתידית על בסיס הפיצ'רים הקיימים. זו בחירה שאינה טריוויאלית:

מצד אחד, אנחנו לא מצפים שבאמת מצליח לשערך את התשואה העתידית על בסיס מעט פיצ'רים קיימים, שכן זאת משימה בלתי אפשרית (או לפחות יש לא מעט חברות בתעשייה שמנסות להרוויח בכך). זאת נקודת ההתחלה שלנו בפרויקט על מנת להשיג שיפור ראשוני.

מצד שני, עקב כך שה-reward הוא מיידי – תלוי אך ורק ב-s,a לא תלות בפעילויות העבר או ללא delayed reward – זה הגיוני עבורנו לבחור שלא להתחשב ב- (כמה "טוב" המצב שנהיה בו לאחר מכן) כיוון שכל האינפורמציה כדי לשערך את נתונה במצב עצמו ללא תלות בעתיד.

שני הגורמים הנ"ל הובילו אותנו לבחור ב-.

קבוע הלמידה נבחר להיות . מראש ידענו שעלינו לבחור קבוע למידה קטן יחסית, מפני שכל אפיזודה שונה במהותה מאחרות (בגלל ההגרלה האקראית של 10 קרנות מתוך 8000 קיימות) ולא רצינו שהאפיזודה האחרונה (או הכמה האחרונות) "יכתיבו" יתר על המידה את התוצאה של וקטור המשקולות במידה והן יחסית חריגות. כמובן שאנחנו רצינו קבוע שהוא מספיק גדול כדי שהסוכן ילמד תוך מספר ניסיונות סביר. לקבוע הזה הגענו לאחר מספר ניסיונות של קבועים בסדרי גודל שונים. קבועים גדולים ממנו ( לדוגמה) יצרנו בעיה של : וקטור המשקולות היה מגיע בתוך מס' ריצות למספרים גבוהים שהם מעבר לתחום הייצוג של המחשב (מוצגים ב-numpy כ-inf). הקבוע שחברנו מייצג את ה-trade-off הנ"ל, הגם שבהינתן כוח חישוב גבוה יותר (כלומר אפשרות להריץ יותר אפיזודות בשלב האימון) ייתכן והיינו בוחרים קבוע קטן יותר.

מרחב הפיצ'רים

לאחר מספר שלבים של ניסוי וטעיה, בחרנו שמרחב הפיצ'רים שלנו יכלול כשישה פיצ'רים (הסעיף הראשון כולל ארבעה פיצ'רים):

1. נתון "התשואה הממוצעת ב-x הזמן האחרון" (כאשר הוא רבעון, שנה, שלוש שנים וחמש שנים). הסעיף הזה הוא למעשה ארבעה פיצ'רים שונים. בניגוד לנתון המקורי, שמופיע כנתון מצטבר (התשואה המצטברת ב- הזמן האחרון), בחרנו לנרמל את הנתונים כך שהוא יציג את התשואה הממוצעת לאורך תקופה. כך הנתון הוא באופן סקאלה ומאפשר לנו להשוות בין המשקולות: אם האלגוריתם ייתן משקולת גבוהה משמעותית לנתון של 5 שנים במקום רבעון אחרון, נוכל להסיק מכך גם על אופן הבחירה של האלגוריתם. המשימה הזו הייתה קשה יותר אילו הפיצ'רים היו נתונים בסקאלות שונות.
2. פיצ'ר של דמי הניהול הצפויים בקרן, מופיע בצורתו המקורית. אחת המחשבות שהייתנו לנו על מעט לנסות פיצ'ר שונה בהיקשר זה היא לנסות לנרמל את הנתון המקורי – לדוג' נרמול בין 0 ל-1 ולאחר מכן לקחת , כך שדמי ניהול נמוכים (שמייצגים נתון "טוב") יהיו גבוהים. לעת עתה השארנו את הפיצ'ר מופיע כנתון המקורי, וזה שיפור שאולי נכניס בעתיד.
3. פיצ'ר שמעיד על אופי הנכסים שבהם הקרן תשקיע – כזכור, קיימת לנו חלוקה של הנכסים שהקרן תשקיע ברמת האחוז מניות, אג"חים או מזומן. קרנות שמשקיעות יותר במניות צפויות להניב תשואה גבוהה יותר (באופן כללי), ועל כן הפיצ'ר הזה מבצע ממוצע משוקלל של הפרמטרים הנ"ל: 85% לכמות הנכסים במניות, 10% לכמות האג"ח ו-5% למזומן.

**הרצת ניסויים ותוצאותיהם**

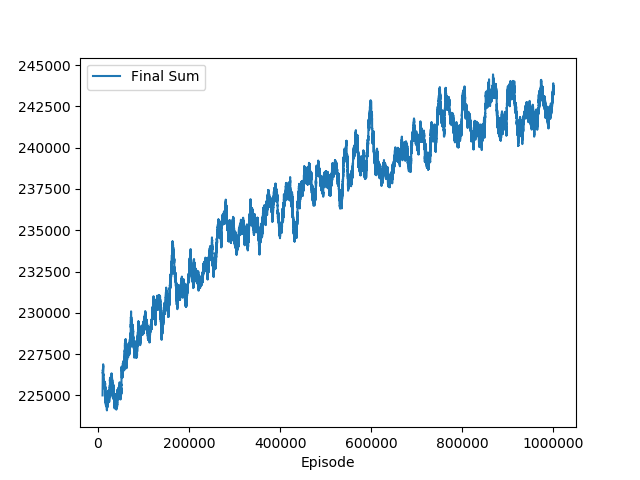
את הפרמטרים הנ"ל כיווננו באמצעות הרצה על מס' מצומצם של קרנות או הרצות על מס' אפיזודות קטן על גבי כל הקרנות. באופן מודע אין כאן חלוקה של train/test – ויש בתוצאות שלנו overfit מסוים על הנתונים הקיימים שלנו. הגישה הזו הייתה לגיטימית בעיננו לאור המטרה הקשה שלפנינו (תחזית בעולם הפיננסי), ולכן עבורנו השימוש בכלים שחדשים לנו (Reinforcement Learning) בשילוב עם בעיה קשה (עולם פיננסי) מצדיקה פגיעה מסוימת ביכולת ההכללה של המודל שנסיים איתו.

המטרה שלנו הייתה בראש ובראשונה לנצח את מדד הייחוס שהוא בחירה אקראית, וכמטרה משנית לנצח חלק מהיוריסטיקות האנושיות (לפחות את הבסיסיות שבהן).

תופעה ראשונה ששמנו אליה במהלך שלב האימון היא שוקטור המשקולות (אותו מנסים לאפטם) לא מתכנס אלא ממשיך להשתנות. ייתכן כמובן ולא הגענו להתכנסות לאורך כל שלב האימון, אך לאור השינויים אנו מעריכים שהוא מבצע תנודות ולא יגיע להתכנסות. בדיעבד, זה הגיוני בגלל אופי הבעיה: בכל אפיזודה נבחרות קרנות שונות, וכיוון שלא יהיה ניתן באמת להעריך את התשואה (שזה ה- באופן בלתי מפורש) כפונקציה ליניארית של הפיצ'רים שלנו, סביר שבכל איטרציה נמצא עדכון כלשהו של המשקולות כל עוד קבוע הלמידה נשאר קבוע. כהערת צד, יכולנו לכפות התכנסות באמצעות שימוש ב- כטור ולא קבוע (כלומר להשתמש ב- כאשר הוא מס' האפיזידות), אך בשלב הראשון נמנענו מכך.

את הסוכן כפי שתואר למעלה (עם הגדרת הפרמטרים ומרחב הפיצ'רים הנ"ל) אומן למשך מיליון אפיזודות. כזכור, בכל אפיזודה נבחרות עשר קרנות באופן אקראי ומשך כל אפיזודה כ-43 תורות. כלומר בסה"כ הסוכן חווה כ-43000000 תורות משחק בהם הוא גם מקבל זוג של מצב-פרס ומעדכן את המשקולות בהתאם.

הגרף הבא מתאר את ה-return הכולל (בסוף כל אפיזודה) כפונקציה של מס' האפיזודות שהסוכן התאמן עליהם:



נזכיר כי הסוכן שבוחר באופן אקראי סיים בממוצע של 223,000 ₪. בגרף הנ"ל הנתון שמוצג הוא ממוצע מתגלגל (rolling mean) של אלפי אפיזודות בכל פעם, אחרת הגרף היה נראה "רועש" מדי. בצורה זו ניתן לראות באופן יותר ברור את המגמה.

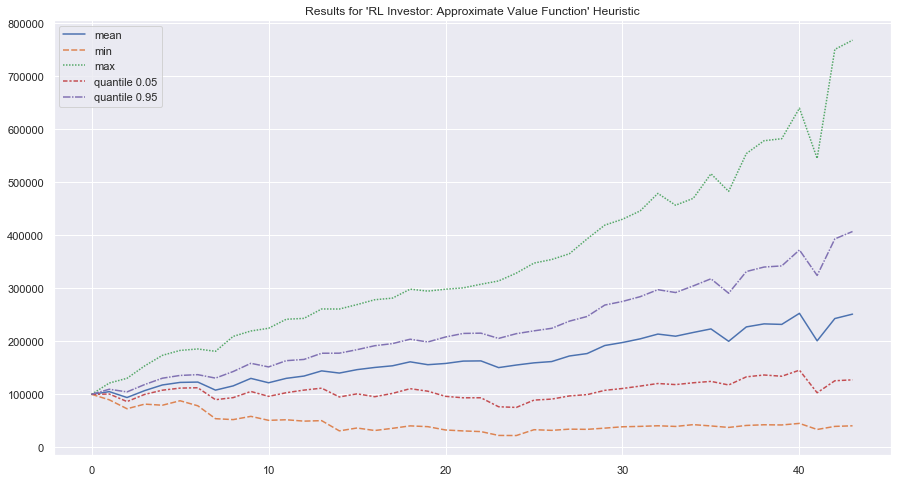
מהגרף ניתן לראות את מגמת הלמידה של הסוכן: ההתחלה היא דומה לתוצאות שהשיג הסוכן שבוחר באופן אקראי, וישנו שיפור עקבי ומינורי לאורך הזמן. העקומה המתקבלת דומה לצפי של עקומות למידה (fitting with plateau) שבה יש שיפור שהולך ודועך עם הזמן. ייתכן ולו היינו מאמנים על מס' אפיזודות גבוה יותר היה מתקבל שיפור נוסף, אבל מהעקומה נראה שאת מרבית השיפור (שמאפשר לנו להעריך את ביצועי הסוכן) כבר השגנו.

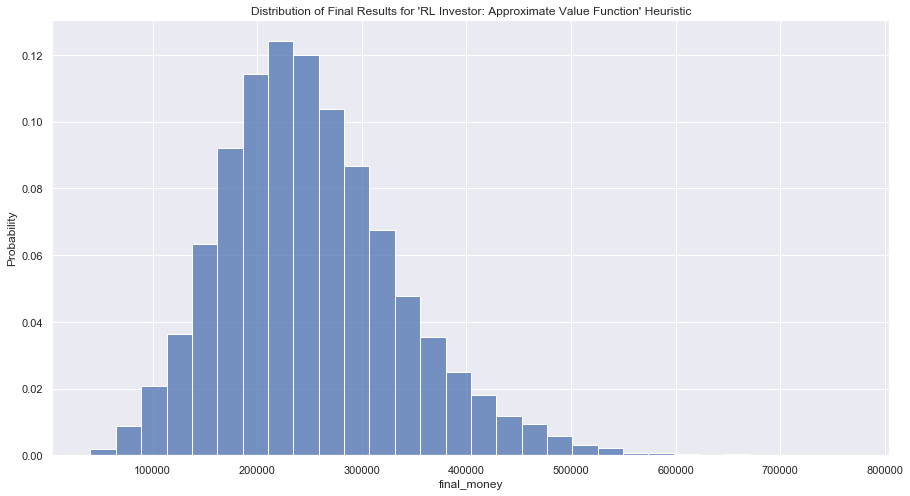
הגרף הנ"ל מראה את עקומת הלמידה של הסוכן, אך אין בו מספיק כדי להעריך באופן מלא את ביצועי הסוכן. לכן לאחר התוצאות הסופיות, נתנו לסוכן בשלב "המבחן" לרוץ על כ-25,000 אפיזודות (בדומה להרצה של סוכנים מבוססי היוריסטיקות), ללא למידה, ומדדנו את התוצאות.

התוצאות הסופיות שהסוכן השיג לאחר אימון של מיליון אפיזודות, כאשר המדידה היא על 25,000 אפיזודות:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| היוריסטיקה | סכום סיום ממוצע | סכום סיום בניסוי הכי מוצלח | סכום סיום בניסוי הכי גרוע | הירידה הבודדת הכי גדולה | תקופת הדימום הכי ארוכה |
| סוכן Reinforcement Learning | 251.27 | 768.06 | 40.61 | 233.174 | 7 |

והגרפים:





בהשוואה למדד הייחוס (בחירה אקראית), "הסוכן החכם" השיג תוצאה ממוצעת טובה יותר ב-10%, וכן היה טוב יותר גם במדדי הסיכון ("הירידה הבודדת הכי גדולה" זהה, ואילו "תקופת הדימום הכי ארוכה" השתפרה מ-10 ל-7).

למרות זאת, "הסוכן החכם" נופל מחלק מהיוריסטיקות שהצגנו בפרק הקודם, לדוגמה מהיוריסטיקה שבוחרת את "התשואה הגבוהה ביותר בחמש השנים האחרונות" וכן "דמי הניהול הכי נמוכים". שתי היוריסטיקות הנ"ל טובות יותר מהסוכן החכם רוחבית כמעט בכל המדדים שלקחנו.

לסיכום, הגרסה הזו של הסוכן החכם הביאה לשיפור על פני מדד הייחוס האקראי, אך היא נפלה מחלק מהיוריסטיקות האנושיות הפשוטות.

וקטור המשקולות הסופי

אחד היתרונות בשיטת ה-Value Function Approximation שבה בחרנו היא שניתן לראות באופן ישיר את וקטור המשקולות הסופי, ובכך לקבל מושג כיצד פיצ'רים מסוימים משפיעים על התוצאה הסופית – לדוגמה החשיבות של כל פיצ'ר (feature importance). בעולם כיום עולה החשיבות ליכולת להסביר מודלים (explainability) של בינה מלאכותית ואת אופן קבלת ההחלטות שלהם, ואחת הדרכים לעשות זאת היא על ידי הבנת החשיבות של כל פיצ'ר בתהליך קבלת ההחלטות של הסוכן.

וקטור המשקולות שקיבלנו:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| פיצ'ר | תשואה ממוצעת בחמש השנים האחרונות | תשואה ממוצעת בשלוש השנים האחרונות | תשואה ממוצעת בשנה האחרונה | תשואה ברבעון האחרון | דמי ניהול בקרן | המדד המשוקלל של אופי הנכסים בקרן |
| משקל | 0.07 | -0.07 | -0.81 | -0.21 | -10.37 | 5.49 |

הנתון המעניין הוא ההשוואה בין המשקולות בפיצ'רים של "התשואה הממוצעת ב- הזמן האחרון": באופן מובהק המשקל המרכזי ניתן לתשואה בשנה האחרונה וברבעון האחרון. זה נתון מפתיע בעיננו, לאור העובדה שדווקא היוריסטיקות שבוחרות את הקרן על בסיס התשואות הללו השיגו ביצועים פחותים ביחס ליוריסטיקות שמבוססות על התשואה בשלוש וחמש השנים האחרונות. לא הצלחנו לחשוב על הסבר מספק לפער הנ"ל.

יש להיזהר מעריכת השוואה בין הפיצ'רים של "התשואה הממוצעת..." לבין הפיצ'רים של "דמי הניהול בקרן" וה-"אופי הנכסים", לאור הסקאלות השונות שבהם הנתונים המקוריים נמצאים. בעוד מדדי ה-"התשואה הממוצעת" הם לרוב בסד"ג של (תשואה חד ספרתית), "דמי הניהול בקרן" לדוגמה נעים בסקאלה שבין 0 לבין 3.

לו היינו מעמיקים יותר בחלק זה (מה התועלת של כל פיצ'ר) יכולנו לבדוק למשל מה הייתה ההשפעה על הסוכן במידה והיינו מורידים פיצ'ר מסוים לחלוטין, או הערכת מקורבת של הגרדיאנט כדי לבדוק כיצד תזוזה קטנה במשקולות משפיעה על התוצאה הסופית.