חישוביות וקוגניציה - תרגיל 8

להגשה עד: 30/12/2021

שימו לב: שאלה 1 היא שאלה שמערבת חישובים אנליטיים וסימולציות נומריות. הגישו קובץ ובו תשובות מילוליות, גרפים, וחישובים שנתבקשתם לערוך, וכן את קבצי הקוד שכתבתם.

שאלה 1

נתון Markov Decision Process (MDP) ובו שני מצבים - Home, Out ושתי פעולות Stay,Switch. סיכויי המעברים מוגדרים באופן הבא:

Stay	H	O
H	1	0
O	0	1

Switch	H	O
H	0.2	1
O	0.8	0

כאשר כל תא בטבלה מתאר את סיכויי המעבר מהמצב שמצויין בעמודה למצב שמצויין בשורה, כאשר לוקחים את הפעולה הרלבנטית. למשל, ההסתברות להגיע החוצה אם נמצאים בבית ובוחרים להחליף, היא:

$$P\left[O|H,\text{Switch}\right]=0.8$$

וכו'.

הגמול מוגדר באופן הבא (בצורה דטרמיניסטית):

$$\begin{aligned} r\left(H,\text{Stay}\right)&=0\\ r\left(H,\text{Switch}\right)&=1\\ r\left(O,\text{Stay}\right)&=2\\ r\left(O,\text{Switch}\right)&=0 \end{aligned}$$

פרמטר ה discounting הוא $\gamma=0.5$

שימו לב: השאלות מתחילות בעמוד הבא.

חלק א' - שיטות מבוססות מודל ומשוואות בלמן

1. נתון סוכן אשר מקבל החלטות באקראי, כלומר בכל מצב בוחר כל אחת מהפעולות בסיכוי שווה של 0.5. כתבו את משוואות בלמן עבור פונקציית הערך V^{π} של הסוכן, ופתרו אותן כדי למצוא את הפונקציה V^{π}.

2. מבלי לבצע חישובים, מהי לדעתכם המדיניות (Policy) האופטימלית?

3. וודאו שה Policy שניחשתם בסעיף 2 מקיימת את משוואות האופטימליות של בלמן.

4. כתבו פונקציה שמממשת את אלגוריתם Value Iteration והשתמשו בה כדי לפתור את ה MDP הנתון. מהי V^*? האם המדיניות האופטימלית π^* שמצאתם בעזרת הפונקציה זהה לזו שמצאתם בסעיף 2?

<u>הערות</u>:

- באלגוריתם ה Value Iteration, אתחלו את האלגוריתם לפונקציית ערך אקראית ($V(s) = 0$ לכל s), ובצעו איטרציות כל עוד יש s עבורו $\left|V^{(t+1)}(s) - V^{(t)}(s)\right| > \epsilon$ עבור ϵ קטן (למשל 1e-10), וכל עוד מספר האיטרציות קטן מ $T_{\max}$ עבור $T_{\max}$ גדול (למשל $T_{\max} = 5000$). שימו לב שבפועל ההתכנסות צפויה להיות מהירה בהרבה).
- השתמשו בכלל עדכון סינכרוני, כלומר חשבו את הערכים החדשים $V^{(t+1)}$ של כל המצבים בהתבסס על הערכים הנוכחיים $V^{(t)}$.

חלק ב' - שיטות Model-Free ולמידת הפרשים זמניים

1. כתבו פונקציה שתאפשר לכם להריץ סימולציה של סוכן בסביבה - כלומר, פונקציה שמקבלת מצב s ופעולה a ומחזירה (באופן הסתברותי במידת הצורך) את המצב הבא s', ואת הגמול שהתקבל מהמעבר (כלומר $R(s,a)$, שימו לב שבמקרה שלנו הגמול הוא דטרמיניסטי)

2. השתמשו באלגוריתם TD-Learning ע`מ ללמוד את V^{π} עבור הסוכן האקראי מסעיף 1 בחלק א'. אתחלו את $\hat{V}$ בצורה שרירותית (למשל $\hat{V}(s) = 0$ לכל s), ובצעו סימולציה של הסוכן בסביבה כאשר בחירת הפעולות נעשית לפי המדיניות שהוגדרה לעיל (בכל מצב, כל פעולה נבחרת באקראי בהסתברות שווה של 0.5). כאשר הסוכן עובר ממצב s למצב s' ומקבל גמול r עדכנו את הערך $V(s)$ לפי כלל הלימוד:

$$\hat{V}(s) \leftarrow \hat{V}(s) + \eta\left(r + \gamma\hat{V}(s') - \hat{V}(s)\right)$$

כאשר η קצב הלימוד (השתמשו ב $\eta = 0.01$). הריצו את הלמידה עד $T = 3000$ צעדים. הציגו גרף של $\hat{V}(\text{Home})$ ו $\hat{V}(\text{Out})$ כפונקציה של מספר הצעדים בסביבה שביצע הסוכן. סמנו בקווים מקווקווים על אותו גרף את הערכים האמיתיים של V^{π} שמצאתם בסעיף 1 של חלק א'. האם הלמידה מתכנסת לערך הנכון? מהי ההשפעה של קצב הלימוד על ההתכנסות?

3. השתמשו באלגוריתם Q-Learning ע`מ ללמוד את המדיניות האופטימלית באופן Off-Policy, בהתבסס על ההתנהגות של הסוכן האקראי. אתחלו את $\hat{Q}$ בצורה שרירותית (למשל $\hat{Q}(s,a) = 0$ לכל s,a) ובצעו סימולציה של אותו סוכן. כאשר הסוכן במצב s, בוחר פעולה a, מקבל גמול r ועבור למצב s' עדכנו את $\hat{Q}(s,a)$ לפי כלל הלמידה הבא:

$$\hat{Q}(s,a) \leftarrow \hat{Q}(s,a) + \eta\left(r + \gamma\max_{a'}\hat{Q}(s',a') - \hat{Q}(s,a)\right)$$

כאשר η קצב הלימוד (השתמשו ב $\eta = 0.1$). הריצו את הלמידה עד $T = 3000$ צעדים. הציגו גרף של $\hat{V}^*(s) \equiv \max_a \hat{Q}(s,a)$ עבור שני המצבים כפונקציה של מספר הצעדים בסביבה שביצע הסוכן. סמנו בקווים מקווקווים על אותו גרף את הערכים האמיתיים של V^* שמצאתם בסעיף 4 של חלק א'. האם הלמידה מתכנסת לערך הנכון?