۹۸۱۰۳۸۶۹۷
محمدابول نژادیان

Subject:
Date

بسم خدا

ترم هفتم سری دوم

سوال ۱ | آ Q-value. دیدیم که برای ارزیابی سیاست، باید Q-value یا V-value را پیدا کنیم. اگر بخواهیم سیاست بهینه را پیدا کنیم، اگر از V-value استفاده کنیم داریم:

$$\pi(s) = \underset{a}{\text{argmax}} \sum T(s,a,s')[R(s,a,s') + \gamma \bar{V}(s')]$$

که در این حالت باید آگاهی از R و P داشته باشیم ولی چون می‌خواهیم model-free کار کنیم، این تقدیر را نداریم. از طرفی برای محاسبه T نیاز به sample های زیادی داریم که همان model based است.

ولی اگر از Q-value استفاده کنیم، به راحتی می‌توانیم سیاست را به روش زیر ارزیابی کنیم:

$$\pi(s) = \underset{a}{\text{argmax}}\, Q(s,a)$$

که برای محاسبه Q ها، sample کمتری داریم که دیگر از مقدار T (یا P) بی‌نیاز شده‌ایم و model free می‌مانیم. (محاسبه مقدار Q ها با Q-learning صورت می‌پذیرد)

Subject:
Date

ب) روش greedy ε یک روش on-Policy برای Reinforcement learning

است که بدین صورت عمل می‌کند که در ε درصد مواقع (ε بسیار کوچک) یک حرکت

Random (در اصل سیاست Random) انجام می‌دهد و در (ε-1) درصد مواقع،

حرکتی براساس (Q(s,a هایی که تا الان بدست آورده انجام می‌دهد

پس در اصل مراحل exPloration و exPlatation را در هم می‌آمیزد و

خوب است که نسبت به روش‌های off-Policy مثل Policy iteration دارد این است که

سرعت بالاتری دارد و از ε بهینه در حال یادگیری هم استفاده می‌کند.

حال اگر از greedy ε استفاده نکنیم و به جای آن از یک روش off-Policy مثل

Policy iteration استفاده کنیم، اگر یک state جدید اضافه شود در حین مسئله و یا

مقادیر reward عوض شوند، چون دیگر به مرحله exPloration بازنمی‌گردیم، در همان سیاست

می‌مانیم که ممکن است در حالت جدید بهینه نباشد.

● پس greedy ε این مزیت را دارد که همواره با احتمالی (هرچند کم) به ما اجازه

exPloration نیز می‌دهد.

PAPCO

محمد امین نیکدل

Subject: ۹۸۱۰۳۸۱۶۷

Date

ج) طبق فرض شده داریم:

$$\forall s: E_{a\sim\pi'}[Q^{\pi'}(s,a)] \geq E_{a\sim\pi}[Q^{\pi}(s,a)]$$

می‌خواهیم: $\forall s: \sum_{s'} P(s'|s,\pi'(s))\left[R(s,\pi'(s),s') + \gamma \max_{a'} Q^{\pi'}(s',a')\right]$

$$\geq \sum_{s'} P(s'|s,\pi(s))\left[R(s,\pi(s),s') + \gamma \max_{a'} Q^{\pi}(s,a')\right] \quad (*)$$

حال در هردو معادله ی داریم:

$$V^{\pi}(s') = \max_{a'} Q^{\pi}(s',a')$$

$$(*) \Rightarrow \forall s: \sum_{s'} P(s'|s,\pi'(s))\left[R(s,\pi'(s),s') + \gamma V^{\pi'}(s')\right]$$

$$\geq \sum_{s'} P(s'|s,\pi(s))\left[R(s,\pi(s),s') + \gamma V^{\pi}(s)\right] \quad (**)$$

همچنین طبق فرمول V_value برای ثابت (Policy fixed) که داریم:

$$V^{\pi} = \sum_{s'} P(s'|s,\pi(s))\left[R(s,\pi(s),s') + \gamma V^{\pi}(s')\right]$$

$$(**) \Rightarrow \boxed{\forall s: V^{\pi'}(s) \geq V^{\pi}(s)}$$