Université Pierre et Marie Curie

Analyse et conception

PIAD de Master1 d'Informatique en Intelligence Artificielle et Décision

---

Semi-supervised Learning Agents

Auteurs :
Lan Zhou
Matthieu Zimmer

Superviseurs :
Paolo Viappiani
Paul Weng

9 mai 2013

Version 1.1

UPMC
SORBONNE UNIVERSITÉS

Table des matières

1   Spécifications

Donner « noir sur blanc » les spécifications détaillées du système, c'est à dire une description précise de ce que fait le système : ensemble des fonctionnalités
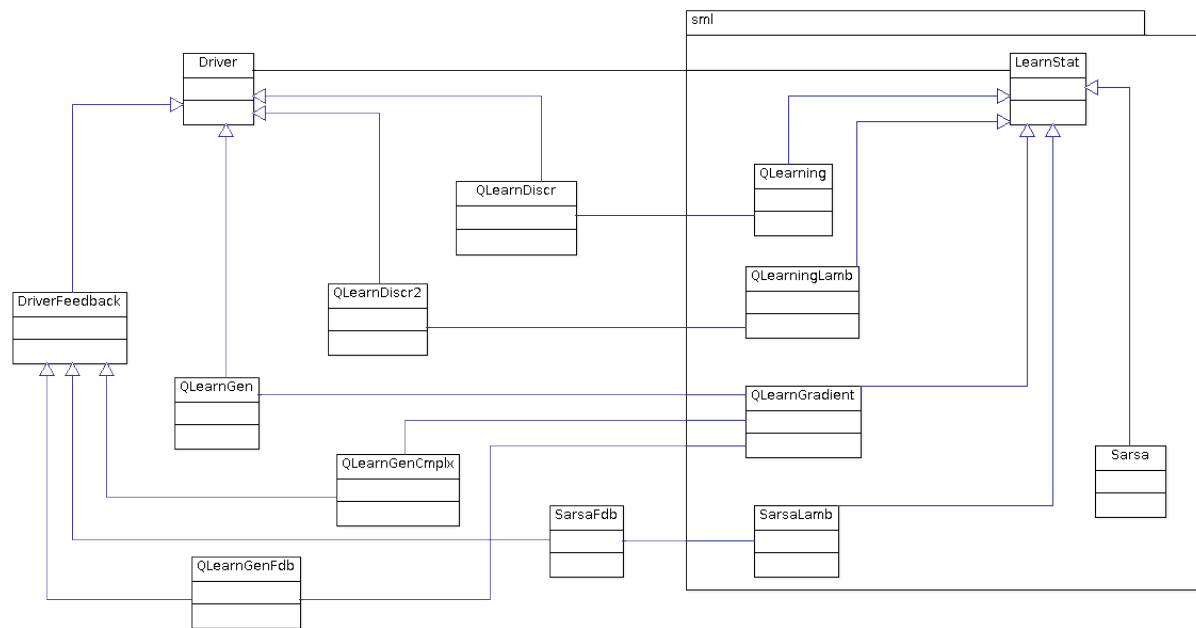
2   Diagrammes d'analyse

FIGURE 1 – Diagramme de classe

2.1   Fonctionnalités

Se référer au plan de développement.

3 Conception

3.1 Algorithmes

Initialize $Q(s, a)$ arbitrarily
Repeat (for each episode):
    Initialize s
    Repeat (for each step of episode):
        Choose a from s using policy derived from Q (e.g., ϵ-
        Take action a, observe r, s'
        $Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma \max_{a'} Q(s', a') - Q(s$
        $s \leftarrow s';$
    until s is terminal

FIGURE 2 – Q-Learning

Initialize $Q(s, a)$ arbitrarily
Repeat (for each episode):
    Initialize s
    Choose a from s using policy derived from Q (e.g., ϵ-greedy)
    Repeat (for each step of episode):
        Take action a, observe r, s'
        Choose a' from s' using policy derived from Q (e.g., ϵ-greedy)
        $Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma Q(s', a') - Q(s, a) \right]$
        $s \leftarrow s'; a \leftarrow a';$
    until s is terminal

FIGURE 3 – Sarsa

Initialize $Q(s, a)$ arbitrarily and $e(s, a) = 0$, for all s, a
Repeat (for each episode):
    Initialize s, a
    Repeat (for each step of episode):
        Take action a, observe r, s'
        Choose a' from s' using policy derived from Q (e.g., ϵ-gr
        $a^* \leftarrow \arg\max_b Q(s', b)$ (if a' ties for the max, then $a^* \leftarrow$
        $\delta \leftarrow r + \gamma Q(s', a^*) - Q(s, a)$
        $e(s, a) \leftarrow e(s, a) + 1$
        For all s, a:
            $Q(s, a) \leftarrow Q(s, a) + \alpha \delta e(s, a)$
            If $a' = a^*$, then $e(s, a) \leftarrow \gamma \lambda e(s, a)$
                    else $e(s, a) \leftarrow 0$
        $s \leftarrow s'; a \leftarrow a'$
    until s is terminal

FIGURE 4 – Q-Learning(λ)

Initialize $Q(s, a)$ arbitrarily and $e(s, a) = 0$, for all s, a
Repeat (for each episode):
    Initialize s, a
    Repeat (for each step of episode):
        Take action a, observe r, s'
        Choose a' from s' using policy derived from Q (e.g., ϵ-greedy)
        $\delta \leftarrow r + \gamma Q(s', a') - Q(s, a)$
        $e(s, a) \leftarrow e(s, a) + 1$
        For all s, a:
            $Q(s, a) \leftarrow Q(s, a) + \alpha \delta e(s, a)$
            $e(s, a) \leftarrow \gamma \lambda e(s, a)$
        $s \leftarrow s'; a \leftarrow a'$
    until s is terminal

FIGURE 5 – Sarsa(λ)

Initialize $\vec{\theta}$ arbitrarily and $\vec{e} = \vec{0}$
Repeat (for each episode):
    $s \leftarrow$ initial state of episode
    For all $a \in \mathcal{A}(s)$:
        $\mathcal{F}_a \leftarrow$ set of features present in s, a
        $Q_a \leftarrow \sum_{i \in \mathcal{F}_a} \theta(i)$
    Repeat (for each step of episode):
        With probability $1 - \epsilon$:
            $a \leftarrow \arg\max_a Q_a$
            $\vec{e} \leftarrow \gamma \lambda \vec{e}$
        else
            $a \leftarrow$ a random action $\in \mathcal{A}(s)$
            $\vec{e} \leftarrow 0$
        For all $i \in \mathcal{F}_a$: $e(i) \leftarrow e(i) + 1$
        Take action a, observe reward, r, and next st
        $\delta \leftarrow r - Q_a$
        For all $a \in \mathcal{A}(s')$:
            $\mathcal{F}_a \leftarrow$ set of features present in s', a
            $Q_a \leftarrow \sum_{i \in \mathcal{F}_a} \theta(i)$
        $a' \leftarrow \arg\max_a Q_a$
        $\delta \leftarrow \delta + \gamma Q_{a'}$
        $\vec{\theta} \leftarrow \vec{\theta} + \alpha \delta \vec{e}$
    until s' is terminal

FIGURE 6 – Q-Learning descente de gradient

Initialize $\vec{\theta}$ arbitrarily and $\vec{e} = \vec{0}$
Repeat (for each episode):
    $s \leftarrow$ initial state of episode
    For all $a \in \mathcal{A}(s)$:
        $\mathcal{F}_a \leftarrow$ set of features present in s, a
        $Q_a \leftarrow \sum_{i \in \mathcal{F}_a} \theta(i)$
    $a \leftarrow \arg\max_a Q_a$
    With probability ϵ: $a \leftarrow$ a random action $\in \mathcal{A}(s)$
    Repeat (for each step of episode):
        $\vec{e} \leftarrow \gamma \lambda \vec{e}$
        For all $\bar{a} \neq a$:   (optional block for replacing traces)
            For all $i \in \mathcal{F}_{\bar{a}}$:
                $e(i) \leftarrow 0$
        For all $i \in \mathcal{F}_a$:
            $e(i) \leftarrow e(i) + 1$ (accumulating traces)
            or $e(i) \leftarrow 1$ (replacing traces)
        Take action a, observe reward, r, and next state, s'
        $\delta \leftarrow r - Q_a$
        For all $a \in \mathcal{A}(s')$:
            $\mathcal{F}_a \leftarrow$ set of features present in s', a
            $Q_a \leftarrow \sum_{i \in \mathcal{F}_a} \theta(i)$
        $a' \leftarrow \arg\max_a Q_a$
        With probability ϵ: $a' \leftarrow$ a random action $\in \mathcal{A}(s)$
        $\delta \leftarrow \delta + \gamma Q_{a'}$
        $\vec{\theta} \leftarrow \vec{\theta} + \alpha \delta \vec{e}$
        $a \leftarrow a'$
    until s' is terminal

FIGURE 7 – Sarsa descente de gradient