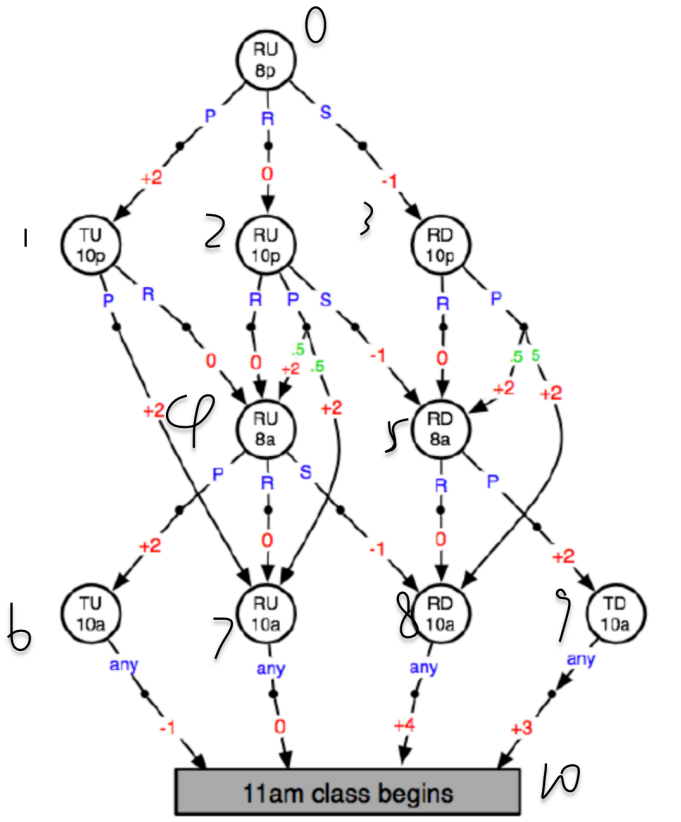
Project 3: Reinforcement learning

Yue Hu

Oct. 29 2019



The numbers indicate the index for the states.

Part 1: run 50 episodes and see experience sequence.

```
(curent state, action) --> (rewards, next state)
```

```
 (RU8p S-->1 RD10p)  (RD10p P-->2 RD10a)  (RD10a R-->4 11am class begins)
 (RU8p R-->0 RU10p)  (RU10p P-->2 RU8a)  (RU8a S-->-1 RD10a)  (RD10a S-->4
11am class begins)
 (RU8p R-->0 RU10p)  (RU10p P-->2 RU8a)  (RU8a P-->2 TU10a)  (TU10a P-->-1
11am class begins)
 (RU8p S-->1 RD10p)  (RD10p R-->0 RD8a)  (RD8a P-->2 TD10a)  (TD10a P-->3
11am class begins)
 (RU8p R-->0 RU10p)  (RU10p P-->2 RU8a)  (RU8a R-->0 RU10a)  (RU10a S-->0
11am class begins)
 (RU8p P-->2 TU10p)  (TU10p R-->0 RU8a)  (RU8a P-->2 TU10a)  (TU10a P-->-1
11am class begins)
 (RU8p S-->1 RD10p)  (RD10p P-->2 RD10a)  (RD10a R-->4 11am class begins)
 (RU8p R-->0 RU10p)  (RU10p P-->2 RU8a)  (RU8a P-->2 TU10a)  (TU10a R-->-1
11am class begins)
 (RU8p P-->2 TU10p)  (TU10p R-->0 RU8a)  (RU8a R-->0 RU10a)  (RU10a S-->0
11am class begins)
```

(RU8p R-->0 RU10p) (RU10p P-->2 RU10a) (RU10a S-->0 11am class begins)
(RU8p R-->0 RU10p) (RU10p S-->-1 RD8a) (RD8a R-->0 RD10a) (RD10a S-->4 11am class begins)
(RU8p P-->2 TU10p) (TU10p P-->2 RU10a) (RU10a S-->0 11am class begins)
(RU8p R-->0 RU10p) (RU10p P-->2 RU8a) (RU8a S-->-1 RD10a) (RD10a P-->4 11am class begins)
(RU8p P-->2 TU10p) (TU10p R-->0 RU8a) (RU8a P-->2 TU10a) (TU10a R-->-1 11am class begins)
(RU8p R-->0 RU10p) (RU10p P-->2 RU8a) (RU8a R-->0 RU10a) (RU10a P-->0 11am class begins)
(RU8p R-->0 RU10p) (RU10p P-->2 RU10a) (RU10a R-->0 11am class begins)
(RU8p S-->1 RD10p) (RD10p P-->2 RD10a) (RD10a P-->4 11am class begins)
(RU8p S-->1 RD10p) (RD10p R-->0 RD8a) (RD8a P-->2 TD10a) (TD10a R-->3 11am class begins)
(RU8p P-->2 TU10p) (TU10p P-->2 RU10a) (RU10a P-->0 11am class begins)
(RU8p P-->2 TU10p) (TU10p R-->0 RU8a) (RU8a P-->2 TU10a) (TU10a S-->-1 11am class begins)
(RU8p P-->2 TU10p) (TU10p P-->2 RU10a) (RU10a S-->0 11am class begins)
(RU8p R-->0 RU10p) (RU10p P-->2 RU8a) (RU8a R-->0 RU10a) (RU10a R-->0 11am class begins)
(RU8p R-->0 RU10p) (RU10p S-->-1 RD8a) (RD8a P-->2 TD10a) (TD10a S-->3 11am class begins)
(RU8p P-->2 TU10p) (TU10p P-->2 RU10a) (RU10a S-->0 11am class begins)
(RU8p R-->0 RU10p) (RU10p S-->-1 RD8a) (RD8a R-->0 RD10a) (RD10a S-->4 11am class begins)
(RU8p P-->2 TU10p) (TU10p P-->2 RU10a) (RU10a P-->0 11am class begins)
(RU8p R-->0 RU10p) (RU10p S-->-1 RD8a) (RD8a P-->2 TD10a) (TD10a R-->3 11am class begins)
(RU8p S-->1 RD10p) (RD10p P-->2 RD10a) (RD10a R-->4 11am class begins)
(RU8p R-->0 RU10p) (RU10p S-->-1 RD8a) (RD8a P-->2 TD10a) (TD10a P-->3 11am class begins)
(RU8p S-->1 RD10p) (RD10p R-->0 RD8a) (RD8a P-->2 TD10a) (TD10a S-->3 11am class begins)
(RU8p R-->0 RU10p) (RU10p S-->-1 RD8a) (RD8a R-->0 RD10a) (RD10a S-->4 11am class begins)
(RU8p R-->0 RU10p) (RU10p S-->-1 RD8a) (RD8a P-->2 TD10a) (TD10a S-->3 11am class begins)
(RU8p P-->2 TU10p) (TU10p P-->2 RU10a) (RU10a P-->0 11am class begins)
(RU8p S-->1 RD10p) (RD10p R-->0 RD8a) (RD8a R-->0 RD10a) (RD10a R-->4 11am class begins)
(RU8p S-->1 RD10p) (RD10p R-->0 RD8a) (RD8a P-->2 TD10a) (TD10a S-->3 11am class begins)
(RU8p R-->0 RU10p) (RU10p S-->-1 RD8a) (RD8a R-->0 RD10a) (RD10a S-->4 11am class begins)
(RU8p R-->0 RU10p) (RU10p P-->2 RU8a) (RU8a R-->0 RU10a) (RU10a R-->0 11am class begins)
(RU8p S-->1 RD10p) (RD10p R-->0 RD8a) (RD8a R-->0 RD10a) (RD10a R-->4 11am class begins)
(RU8p P-->2 TU10p) (TU10p P-->2 RU10a) (RU10a R-->0 11am class begins)
(RU8p R-->0 RU10p) (RU10p S-->-1 RD8a) (RD8a R-->0 RD10a) (RD10a P-->4 11am class begins)
(RU8p S-->1 RD10p) (RD10p P-->2 RD10a) (RD10a S-->4 11am class begins)
(RU8p P-->2 TU10p) (TU10p P-->2 RU10a) (RU10a R-->0 11am class begins)
(RU8p R-->0 RU10p) (RU10p P-->2 RU10a) (RU10a S-->0 11am class begins)

```
 (RU8p S-->1 RD10p)  (RD10p P-->2 RD8a)  (RD8a R-->0 RD10a)  (RD10a P-->4
11am class begins)
 (RU8p S-->1 RD10p)  (RD10p P-->2 RD10a)  (RD10a P-->4 11am class begins)
 (RU8p R-->0 RU10p)  (RU10p R-->0 RU8a)  (RU8a S-->-1 RD10a)  (RD10a R-->4
11am class begins)
 (RU8p R-->0 RU10p)  (RU10p R-->0 RU8a)  (RU8a R-->0 RU10a)  (RU10a S-->0
11am class begins)
 (RU8p R-->0 RU10p)  (RU10p S-->-1 RD8a)  (RD8a P-->2 TD10a)  (TD10a S-->3
11am class begins)
 (RU8p R-->0 RU10p)  (RU10p P-->2 RU8a)  (RU8a S-->-1 RD10a)  (RD10a R-->4
11am class begins)
 (RU8p S-->1 RD10p)  (RD10p R-->0 RD8a)  (RD8a P-->2 TD10a)  (TD10a P-->3
11am class begins)
```

Average return for 50 epoches is 4.100

State values for random policy, which are calculated by policy evaluation using Bellman equations.

```
state values:  (RU8p : 4.18)  (TU10p : 1.67)  (RU10p : 2.50)  (RD10p : 5.38)
(RU8a : 1.33)  (RD8a : 4.50)  (TU10a : -1.00)  (RU10a : 0.00)  (RD10a : 4.00)
(TD10a : 3.00)  (11am class begins : 0.00)
```

We can see that the Monte Carlo estimation for start state (4.1) is quite near the actual start state value (4.18).

Part 2: learning optimal policy by policy iteration

```
iteration:  0
state values:  (RU8p : 4.00)  (TU10p : 2.00)  (RU10p : 2.50)  (RD10p : 6.50)
(RU8a : 1.00)  (RD8a : 5.00)  (TU10a : -1.00)  (RU10a : 0.00)  (RD10a : 4.00)
(TD10a : 3.00)  (11am class begins : 0.00)
Policy:  (RU8p: S)  (TU10p: P)  (RU10p: S)  (RD10p: P)  (RU8a: S)  (RD8a: P)
(TU10a: P)  (RU10a: P)  (RD10a: P)  (TD10a: P)
iteration:  1
state values:  (RU8p : 7.50)  (TU10p : 2.00)  (RU10p : 4.00)  (RD10p : 6.50)
(RU8a : 3.00)  (RD8a : 5.00)  (TU10a : -1.00)  (RU10a : 0.00)  (RD10a : 4.00)
(TD10a : 3.00)  (11am class begins : 0.00)
Policy:  (RU8p: S)  (TU10p: R)  (RU10p: S)  (RD10p: P)  (RU8a: S)  (RD8a: P)
(TU10a: P)  (RU10a: P)  (RD10a: P)  (TD10a: P)
iteration:  2
state values:  (RU8p : 7.50)  (TU10p : 3.00)  (RU10p : 4.00)  (RD10p : 6.50)
(RU8a : 3.00)  (RD8a : 5.00)  (TU10a : -1.00)  (RU10a : 0.00)  (RD10a : 4.00)
(TD10a : 3.00)  (11am class begins : 0.00)
Policy:  (RU8p: S)  (TU10p: R)  (RU10p: S)  (RD10p: P)  (RU8a: S)  (RD8a: P)
(TU10a: P)  (RU10a: P)  (RD10a: P)  (TD10a: P)
```

We can see that the policy converges quite fast, in just 3 iterations. The optimal policy says that the student should study and finish the homework first, then party afterwards night to day.