Relatório Lab 10 - CT213

Código:
A implementação da avaliação de política e da interação por valor foram baseadas na greedy police, só que guardávamos o valor o anterior para poder testar a condição de parada (usou-se np.copy para evitar que o antigo valor fosse associado ao mesmo endereço do antigo valor e a diferença ficasse sempre nula). Na interação por política, a implementação foi mais straightforward, basicamente só utilizamos as funções implementadas anteriormente, alterando os parâmetros.

Resultados:
Os primeiros resultados (os quais consideram que sempre escolhemos a “melhor” ação no momento e consideram iguais as recompensas imediatas e futuras) fazem bastante sentido, i.e. a política sugerida por eles faz com que partindo de qualquer ponto do grid o agente chegue ao objetivo.

Os resultados que consideram aleatoriedade (o agente não escolher a melhor ação) e diferenciam recompensas imediatas e futuras (usam fator de desconto) são executados em menor tempo e mostram um número menor de caminhos possíveis (ele diferencia mais os valores de cada estado), mas ainda satisfaz as mesmas condições do que o primeiro resultado

Gamma = 1.00 and Correct_Action_Probability = 1.00

Evaluating random policy, except for the goal state, where policy always executes stop:
Value function:
[-384.09, -382.73, -381.19, * , -339.93, -339.93]
[-380.45, -377.92, -374.65, * , -334.93, -334.93]
[-374.35, -368.82, -359.85, -344.89, -324.92, -324.93]
[-368.77, -358.19, -346.03,* , -289.95, -309.94]
[* , -344.12, -315.06, -250.02, -229.99, *]
[-359.12, -354.12, * , -200.01, -145.00, 0.00]
Policy:
[SURDL , SURDL , SURDL , * , SURDL , SURDL]
[SURDL , SURDL , SURDL , * , SURDL , SURDL]
[SURDL , SURDL , SURDL , SURDL , SURDL , SURDL]
[SURDL , SURDL , SURDL , * , SURDL , SURDL]
[* , SURDL , SURDL , SURDL , SURDL ,*]
[SURDL , SURDL , * , SURDL , SURDL ,S]

Value iteration:
Value function:
[-10.00, -9.00, -8.00, * , -6.00, -7.00]
[-9.00, -8.00, -7.00, * , -5.00, -6.00]
[-8.00, -7.00, -6.00, -5.00, -4.00, -5.00]
[-7.00, -6.00, -5.00, * , -3.00, -4.00]
[* , -5.00, -4.00, -3.00, -2.00, *]
[-7.00, -6.00, * , -2.00, -1.00, 0.00]
Policy:
[RD , RD , D , * , D , DL]
[RD , RD , D , * , D , DL]
[RD , RD , RD , R , D , DL]
[R , RD , D , * , D , L]
[* , R , R , RD , D , *]
[R , U , * , R , R , SURD]

Policy iteration:
Value function:
[-10.00, -9.00, -8.00, * , -6.00, -7.00]
[-9.00, -8.00, -7.00, * , -5.00, -6.00]
[-8.00, -7.00, -6.00, -5.00, -4.00, -5.00]
[-7.00, -6.00, -5.00, * , -3.00, -4.00]
[* , -5.00, -4.00, -3.00, -2.00, *]
[-7.00, -6.00, * , -2.00, -1.00, 0.00]
Policy:
[RD , RD , D , * , D , DL]
[RD , RD , D , * , D , DL]
[RD , RD , RD , R , D , DL]
[R , RD , D , * , D , L]
[* , R , R , RD , D , *]
[R , U , * , R , R , SURD]

Gamma = 0.98 and Correct_Action_Probability = 0.8

Evaluating random policy, except for the goal state, where policy always executes stop:
Value function:
[-47.19, -47.11 , -47.01 , * , -45.13, -45.15]
[-46.97, -46.81, -46.60, * , -44.58, -44.65]
[-46.58, -46.21, -45.62, -44.79, -43.40, -43.63]
[-46.20, -45.41, -44.42, * , -39.87, -42.17]
[* , -44.31, -41.64, -35.28, -32.96, *]
[-45.74, -45.28, * , -29.68, -21.88, 0.00]
Policy:
[SURDL , SURDL , SURDL , * , SURDL , SURDL]
[SURDL , SURDL , SURDL , * , SURDL , SURDL]
[SURDL , SURDL , SURDL , SURDL , SURDL , SURDL]
[SURDL , SURDL , SURDL , * , SURDL , SURDL]
[* , SURDL , SURDL , SURDL , SURDL , *]
[SURDL , SURDL , * , SURDL , SURDL ,S]

Value iteration:
Value function:
[-11.65, -10.78, -9.86, * , -7.79, -8.53]
[-10.72, -9.78, -8.78, * , -6.67, -7.52]
[-9.72, -8.70, -7.59, -6.61, -5.44, -6.42]
[-8.70, -7.58, -6.43, * , -4.09, -5.30]
[* , -6.43, -5.17, -3.87, -2.76, *]
[-8.63, -7.58, * , -2.69, -1.40, 0.00]
Policy:
[D , D , D , * , D , D]
[D , D , D , * , D , D]
[RD , D , D , R , D , D]
[R , RD , D , * , D , L]
[* , R , R , D , D , *]
[R , U , * , R , R , S]

Policy iteration:
Value function:
[-11.65, -10.78, -9.86, * , -7.79, -8.53]
[-10.72, -9.78, -8.78, * , -6.67, -7.52]
[-9.72, -8.70, -7.59, -6.61, -5.44, -6.42]
[-8.70, -7.58, -6.43, * , -4.09, -5.30]
[* , -6.43, -5.17, -3.87, -2.76, *]
[-8.63, -7.58, * , -2.69, -1.40, 0.00]
Policy:
[D, D , D , * , D , D]
[D, D , D , * , D , D]
[R, D , D , R , D , D]
[R, D , D , * , D , L]
[*, R , R , D , D , *]
[R, U , * , R , R , S]
