Cheat Sheet

Step breakdown – Q learning

Q(s, a) = w1f1 + w2f2
find the max of multiple Q values

Q(s, west) = 1(weight) **(1/1) + 10**(3)
Q(s, south) = 1*(1/1) + 10*(1)

Q(s, west)(current) = 31
Q(s, south)(current) = 11

Q – learning still

Q(s, west) = 11
Q(s, south) = 11

sample = r + Γ * max Q(s', a')

sample = 9+1*11 = 20

difference = sample – current

difference = 20 – 31 = -11

update weights

Q(s, a) = w1f1 + w2f2

w1 = w1 + alpha * difference * f1
w2 = w2 + alpha * difference * f2

w1 = 1 + 0.5 * -11 * 1
w2 = 10 + 0.5 * -11 * 3

---

value iteration

(Max) Σ T(s, a, s')*((R(s, a, s') + Γ * V(s'))

0.8(0 + 1 * (-2)) + 0.2(0 + 1 * 0) = -1.6 //Going Right

0.8(0 + 1 * (0)) + 0.2(0 + 1 * (-2)) = -0.4 //Going Down

Direct Evaluation:

Episode 1 : (0 + 0 +10) = 10
Episode 2: (0 – 10) = -10
Episode 3: (0+0-10) = -10
Episode 4: (0-10) = -10

(10-10-10-10) / 4 (episodes given)

= -5

(1,2)
only episodes 2, 4

Episode 2: (0-10) = -10
Episode 4:(0-10) = -10

(-10-10)/2
= -10

prob ((1,2) | (1,1), down)

-5/-10 = ½= 0.5

<u>temporal difference learning:</u>

1 – α *V(s) + α * reward + Γ * V(s')

V(s) = direct evaluation

gamma + alpha = given

Probability:

P(A) = summation of all values that equals A
P(B| A) = P(A) / P (A, B)
P(A , B| C) = P(C) / P(A, B, C)
P(C | A, B) = P(A, B, C) / P(A, B)