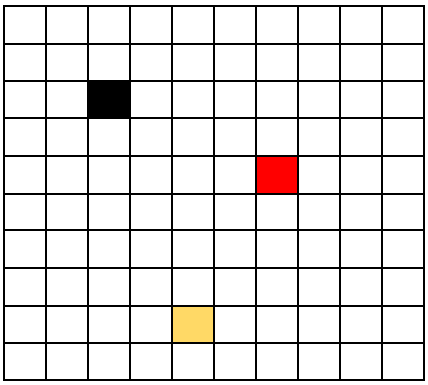
Learning the Water Maze:

As an example of generalized reinforcement learning, we consider the water maze task. This is a navigation problem in which rats are placed in a large pool of milky water and have to swim around until they find a small platform that is submerged slightly below the surface of the water. The opaqueness of the water prevents them from seeing the platform directly, and their natural aversion to water (although they are competent swimmers) motivates them to find the platform. After several trials, the rats learn the location of the platform and swim directly to it when placed in the water.

We are going to simulate a simple model of navigation problem.



Simulation parameters:

1. 15x15 map
2. Fixed target (black)
3. Fixed cat (red)
4. A random starting point in each trial (yellow)
5. 4 direction to move
6. Each square have 4 probability for 4 directions to move (wall = 0)
7. At the beginning equal probability of movement = 0.25 (except wall restriction)
8. Start step by step movement with the probability of each direction.
9. End of trial = being in red or black

Each trial will end with achieving to red or black. If you arrive in black you have to increase the probability of the last move with small amount ϵ (learning rate), and with arriving in red you should decrease the probability of the last move with the same amount. Be aware that summation of all probabilities must be equal to 1. Also, you need to find the algorithm to propagate the update of probabilities. For example: If you arrive in black you have to increase the probability of last move a little (ϵ). And if you arrive to square with dominant probability (= 0.6), you have to increase the probability of last move a little (ϵ).

1. Plot the paths before and after training. 1.1 assign demo files (video format).
2. Plot the gradients and contour plot of the learned values.
3. What is the effect of the learning rate (ϵ) and discount factor (γ) in this problem? Hint: consider the minimum number of iterations needed for reaching the optimal paths for each ϵ and γ in a heatmap. (you may need to plot the heatmap in logarithmic scale to observe the differences)
4. Consider two target squares with different positive values; what is the effect of ϵ and γ in the learning procedure? Compare different values and explain your observations.
5. Implement TD(λ) algorithm and compare with previous section.