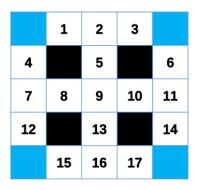
## **Final Project**

## 1. Dynamic programming (DP)



给定迷宫包含四个出口(蓝色方框),四个障碍物(黑色方块),17 个状态位。 每走一步,回报减 1 (reward = -1)。 分别用以下两种方式编写代码计算出最优价值函数 V\* 以及最优策略  $\pi*$ 

- 策略迭代
- 值迭代

## 2. TD-learning (TD)

生成二维迷宫并定义迷宫入口,每个格子内定义到达该位置可以获得的奖励 (reward),智能体在迷宫中可以上下左右四个方向移动,每次移动一格获得相邻位置的奖励.(如果碰到墙壁则反弹回原来位置再次获得奖励).通过强化学习训练智能体在有限的步数(STEP-MAX)内获得最大累计奖励. 分别实现以下方案:

- Sarsa
- Q-learning
- Sarsa(λ)