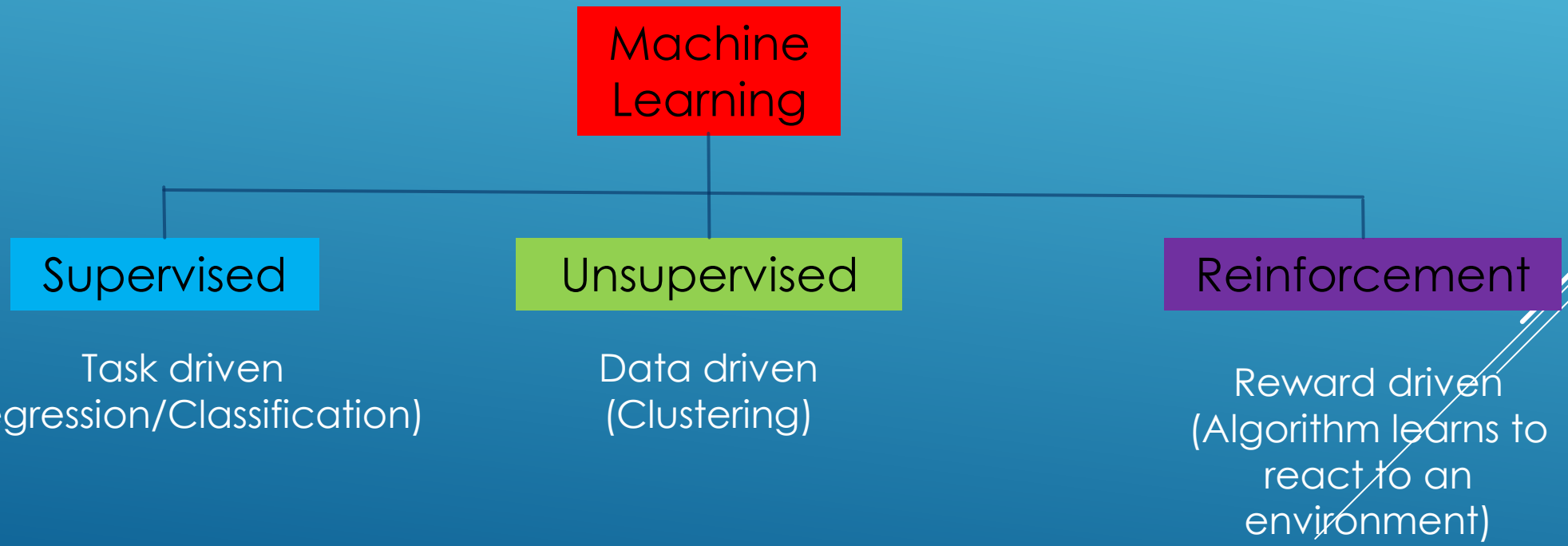
REINFORCEMENT LEARNING IN QUANTITATIVE TRADING

Kingsley (Mingyang) Di

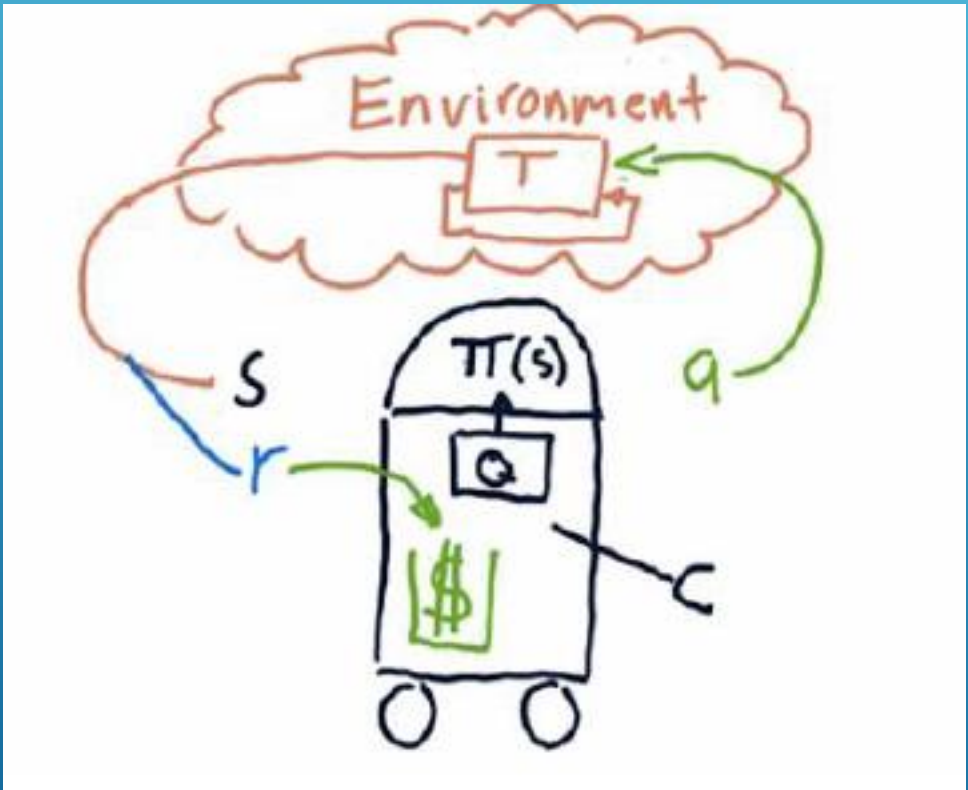
05/18/2017

MACHINE LEARNING

Types of Machine Learning



RL PROBLEM



AlphaGo

RL IN STOCK TRADING

Robotics Example

Stock Training

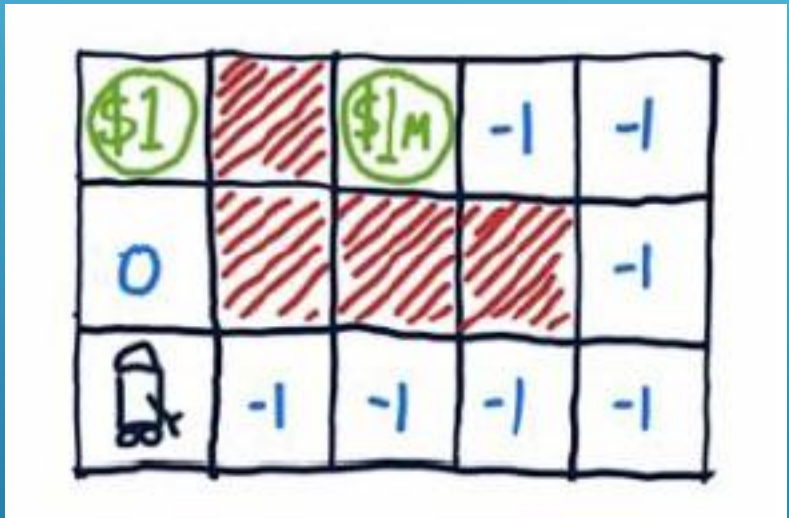
MARKOV DECISION PROCESS

a. Set of States S
b. Set of actions A
c. Transition function $T[s, a, s']$
d. Reward function $R[s, a]$

Find:

Policy $\pi^*(s)$ that will maximize reward

WHAT TO OPTIMIZE?



Scenario 1: infinite horizon $\sum_{i=1}^{\infty} r_i$

Scenario 2: finite horizon $\sum_{i=1}^{n} r_i$

Scenario 3: infinite horizon with discounted reward $\sum_{i=1}^{\infty} \lambda^{i-1} r_i$

$$E[R[s_1, a_1] + \lambda R[s_2, a_2] + \lambda^2 R[s_3, a_3] + \cdots]$$

HOW TO SOLVE?

- With known transitions and rewards
  - Policy iteration
  - Value iteration

Practical ???

- Unknown transitions and rewards
  - Experience tuple: $\langle s_1, a_1, s_1', r_1 \rangle$

    $\langle s_2, a_2, s_2', r_2 \rangle$

    ...
  - Model-based: build a model of $T[s, a, s']$ and $R[s, a]$, then policy/value iteration
  - Model-free: Q-learning

$$T[s, a, s'] = \frac{\#\text{ times took we action } a \text{ state } s \text{ and got to } s'}{\#\text{ times we took action } a \text{ in state } s}$$

Q-LEARNING

a. A model-free approach: it does not know or use models of the transitions T or rewards R
b. Builds a table of utility values as the agents interacts with the world
c. Guaranteed to provide an optimal policy (!!!)

Q-LEARNING

What is Q

$$Q[s,a] = \text{immediate reward} + \text{discounted future reward}$$

How to use Q?

$$\pi(s) = \arg\max_a(Q[s,a])$$

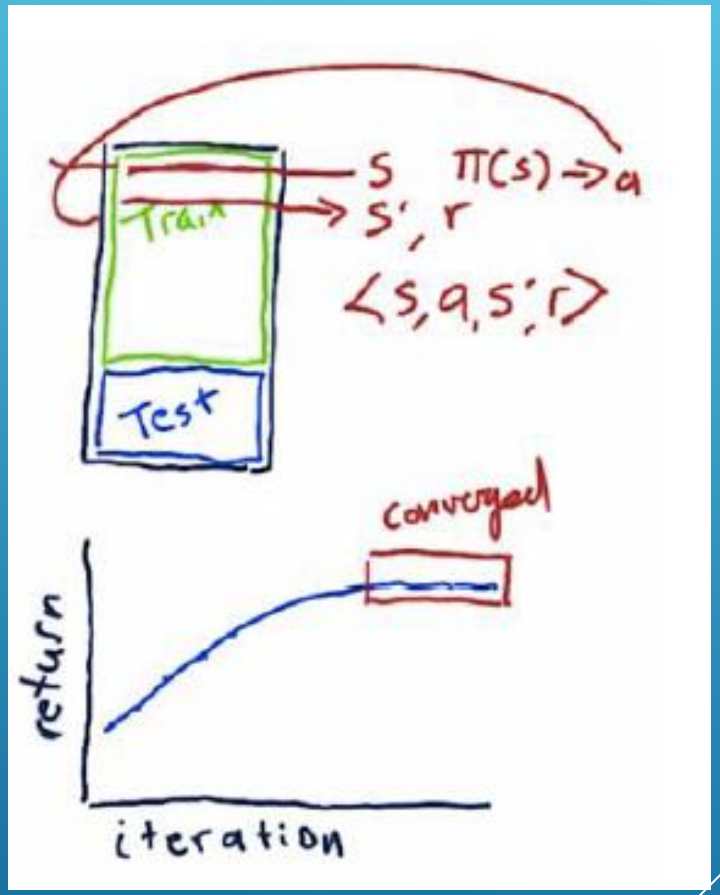
$\pi^*(s)$ $Q^*[s,a]$

Q-LEARNING

Big Picture:
- Select training data
- Iterate over time $\langle s, a, s', r \rangle$
- Test policy π (backtest)
- Repeat until converge

Details:
- Set start time, initialize $Q[\quad]$
- Compute s
- Select a
- Observe r, s'
- Update Q

$\langle s, a, s', r \rangle$



Q-LEARNING

Update Rule

α: learning rate 0 to 1.0 (in practice, around 0.2)

λ: discount rate 0 to 1.0

$$Q'[s,a] = (1-\alpha)Q[s,a] + \alpha \cdot \text{improved estimate}$$

$$Q'[s,a] = (1-\alpha)Q[s,a] + \alpha(r + \lambda \cdot \text{later rewards})$$

$$Q'[s,a] = (1-\alpha)Q[s,a] + \alpha\left(r + \lambda \cdot Q\left[s', \arg\max_{a'}(Q[s',a'])\right]\right)$$

THANK YOU