Introduction to Reinforcement Learning

Subrahmanya Swamy Peruru

Photos provided by Pexels

Paradigms of Machine Learning

- Supervised Learning ✓ 1
- Unsupervised Learning 2
- Reinforcement Learning 3

Supervised Learning

function fitting

$x, f(x)$

(x, y)

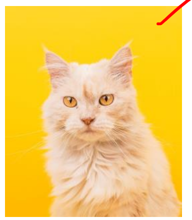
(x_1, y_1)

(x_2, y_2)

(x_3, y_3)

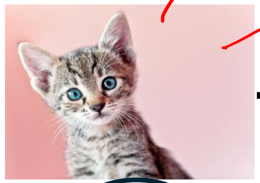
(x_4, y_4)

Labeled Training Data



x_1

Cat y_1



x_2

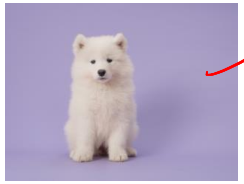
x

Cat y_2

$\rightarrow x$

x_3

Dog y_3



x_4

Dog y_4

$f(x)$

Labels

Training

New image

x_5

$f(x_5) = y_5$

Trained model

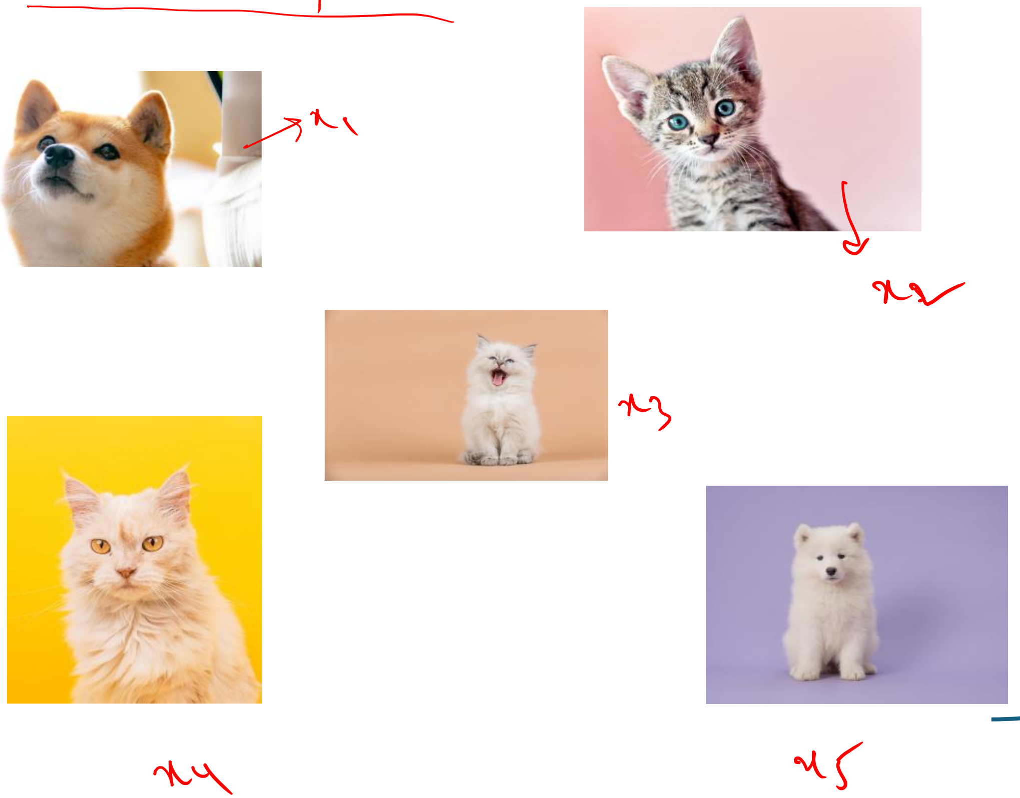
$f'(x)$

Cat | Dog = ??

Unsupervised Learning

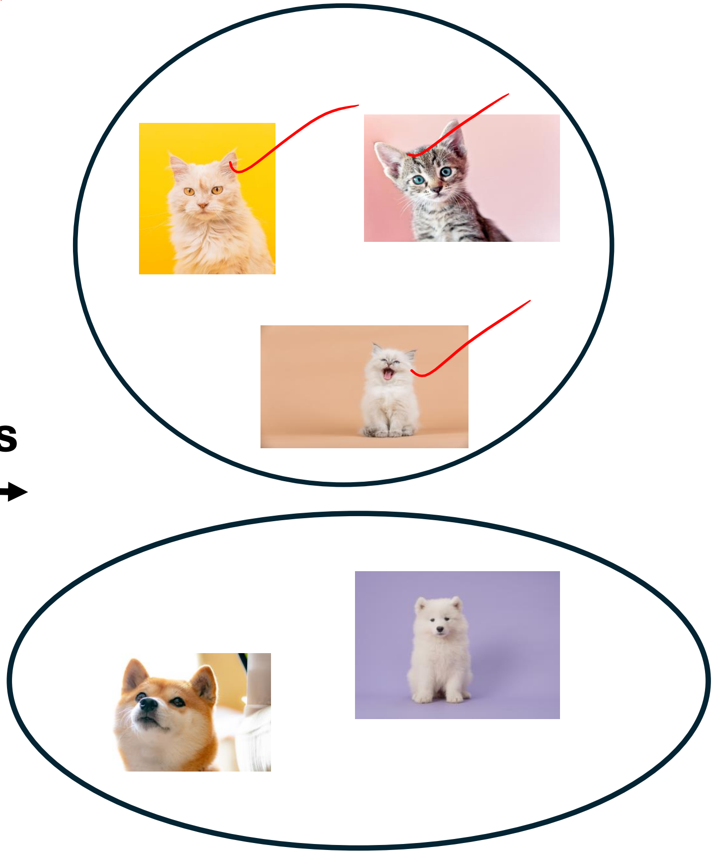
(x_1, y_1)
$x_1 \to$ i/p
$y_1 \to$ label

Unlabeled Data



x

Identify patterns



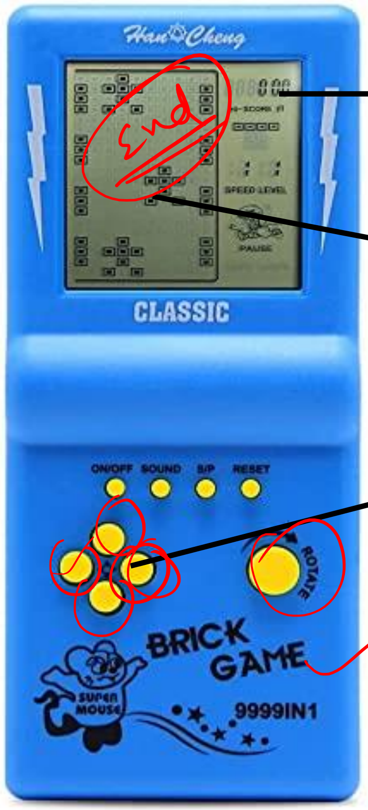
Photos provided by Pexels

Reinforcement Learning

$S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}, R_{t+2}, S_{t+2}, \ldots S_T$

Terminal

Episode



Feedback:
Score,
new display

State:
Display

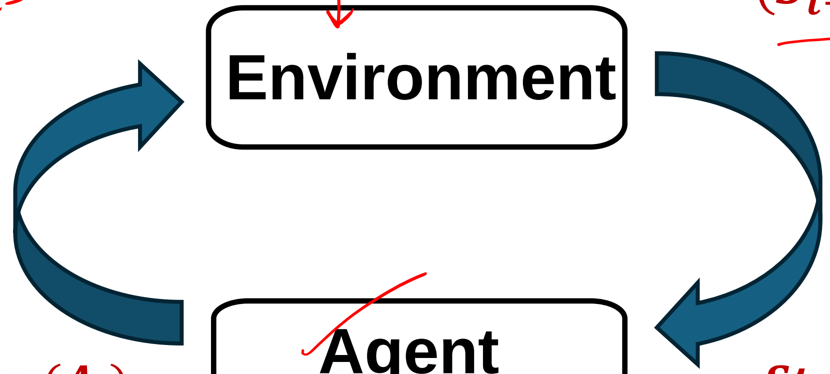
Actions:
UP / Left /
Right / Down

+100

+1
-5

Learn by Trial and Error

program



Next State, Reward
(S_{t+1}, R_{t+1})

$\pi : S \to A$

$\pi(s)$

Action (A_t)

State (S_t)

1. Agent observes the state and takes action
2. Environment puts the agent in a new state &
3. Gives a reward based on the action taken

+1 -1
win / lose

GOAL: Learn policy to maximize the cumulative reward

$\sum_t R\_t$

$\boxed{R_1 + R_2}$

$S_0, A_0, R_1, S_1, A_1, R_2, S_2.$

0 0

Photos provided by Pexels

Paradigms of Machine Learning

(x_i, y_i)

- Supervised Learning
  - Fitting a function for the given labeled data (x, y)
  - $y \approx f(x)$

- Unsupervised Learning
  - Identifying patterns in unlabeled data
  - E.g. Clustering

π $S \rightarrow A$

$(S, ?)$

- Reinforcement Learning
  - Learning sequential tasks through "trial and error"
  - Feedback through reward/penalty

Photos provided by Pexels

RL Demonstrations

DeepRL

MCTS

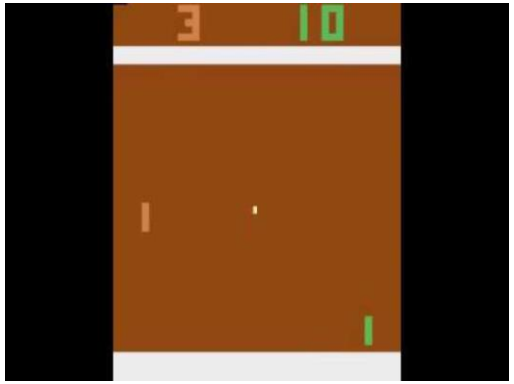
Go

→ Finance
→ Wireless Networks
→ Robotics

2016

"**Autonomous Helicopter**"

Pong game



AlphaGo by DeepMind

"Multi-arm Bandits"

Photos provided by Pexels

One State RL: Multi-arm Bandits

- Simplified version of RL problem: "Multi-arm Bandit" problem.
  - Only one state
  - Multiple actions (a.k.a. arms)
    - $\mathcal{A}$ – Action set

a_1

a_2

a_3

- A reward distribution corresponding to each arm
  - $\mathcal{R}_a$ – Reward distrution for action a
  - $\mu_a = \mathbb{E}[\mathcal{R}_a]$ – Expected reward for action a

- Applications: Recommendation systems, Ad placement, ...

"Cognitive Radios"

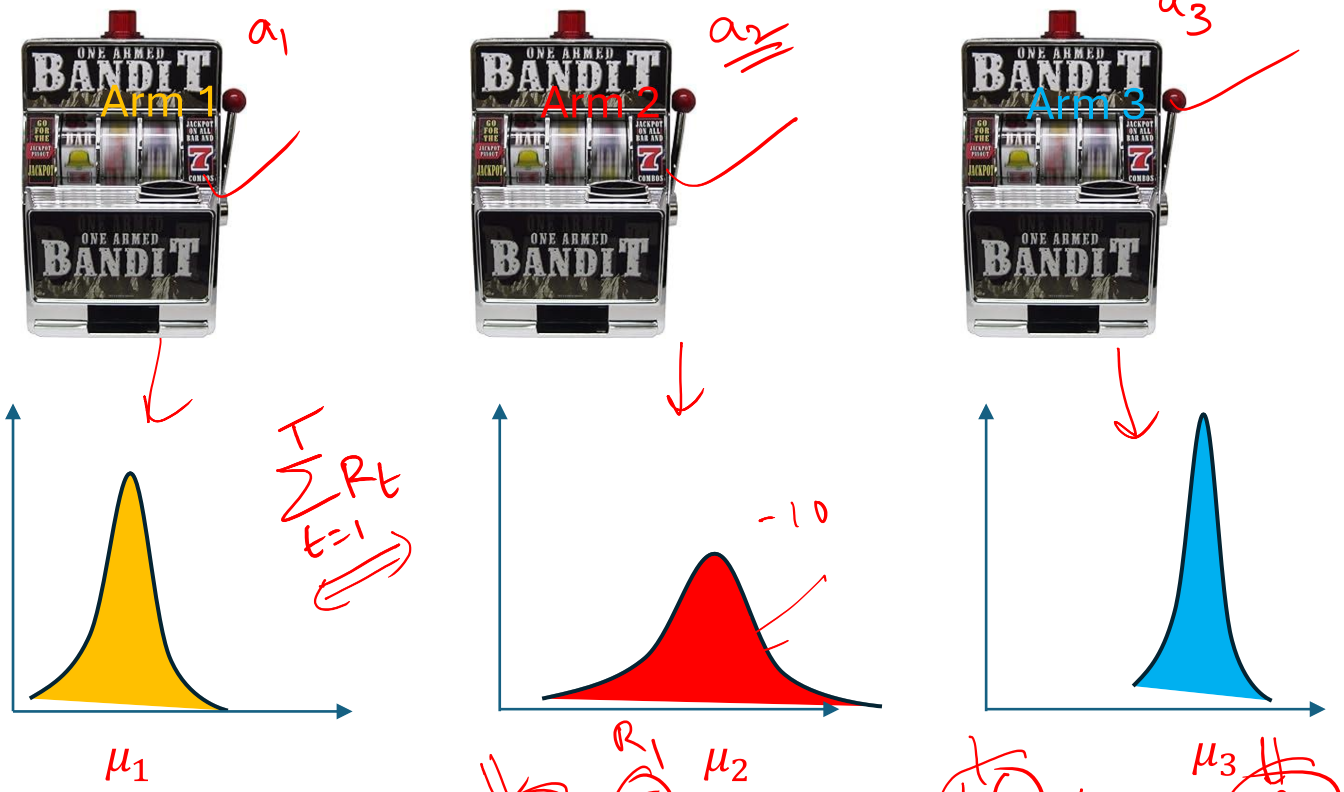
Photos provided by Pexels

Multi-arm Bandits

"thief"

100 rounds

$a_2, a_2, \ldots a_2$ 9 91



Reward Distributions

8.33 R_1 5 10 5 +6 +2

$a_1, a_1, a_1, a_2, a_2, a_2, a_3, a_3, a_3$

$a_1, a_1, \ldots a_1$ 91 times

Problem:

- Reward distributions are **unknown**
- Given **T chances** to pull the arms
- Which arms should be pulled to **maximize the total reward** in those T rounds

Exploration

Vs

Exploitation dilemma

Photos provided by Pexels

ETC (Explore-Then-Commit)

$\mu(a)$ $\mu(b)$ $\mu(c)$ b)

$+5$ $+8$

$\mu^* = \max\{\mu(a), \mu(b), \mu(c)\}$

b

$E[x] = \int_x x P_x(x)\, dx = \mu(a)$

Unknown

1. **Explore:** Play each arm N times

$T-KN$

2. Compute the sample average rewards $\bar{\mu}(a) = \frac{1}{N}\sum_{t=1}^{KN} R_t\, 1\{a_t = a\}$ for each arm $a \in \mathcal{A}$

94

3. **Commit:** Play the arm with the highest sample average for the remaining $T - KN$ rounds

μ^* - Optimal arm's expected reward

a_t - Arm played in round t

K - Number of arms

R_t - Sample reward obtained in round t

T - Total number of rounds

$\mu^* + \mu^* + \mu^* + \cdots \mu^*$ $\quad E[R_1 + R_2 + \cdots R_T]$

Performance (ETC Vs Best possible reward) : $T\mu^* - \sum_{t=1}^{T} \mathbb{E}[R_t]$

How much to Explore? "$N \approx (\frac{T}{K})^{\frac{2}{3}}$"

Photos provided by Pexels

ϵ-Greedy (Explore uniformly)

UCB

a, b, c   ETC

4th   Exp NK   T−NK Exploit

'a' is best arm based on current estimates

a, b, c

0.01   +5   +3   −2

ε   1−ε   t=4

H/T   0.99   "a"

1. Play each arm once

2. In each round t:
   - Toss a coin with bias ϵ.
   - If it lands in head: Explore - Play any arm randomly
   - Else: Exploit - Play the arm with the highest sample average so far

What ϵ to choose? $\epsilon \approx (\frac{K}{T})^{\frac{1}{3}}$

UCB (Upper Confidence Bound)

Optimism under UnCertainty

1. Play each arm once in the first K rounds

2. For $t > K$:

   UCB score

   Exploit Explore

   - Play the arm with the highest $UCB_t(a) = \overline{\mu_{t-1}}(a) + \sqrt{\frac{2 \log T}{n_{t-1}(a)}}$

   - Based on the observed sample reward R_t, update $n_t(a_t)$ and $\overline{\mu_t}(a_t)$
     - $n_t(a_t) = n_{t-1}(a_t) + 1$
     - $\overline{\mu_t}(a_t) = \frac{1}{n_t(a_t)}[(n_t(a_t) - 1)\,\overline{\mu_{t-1}}(a_t) + R_t]$

$a_t = a$

$n_{t-1}(a) = 2$
$n_{t-1}(b) = 2$
$n_{t-1}(c) = 1$

$+5$ $+6$ $+7$ $+1$ $+2$
$a,\ b,\ c,\ a,\ b,$

$t = 7$

$\mu_6(a) = \frac{5+1}{2} = 3$
$\mu_6(b) = 4$
$\mu_6(c) = 7$

$b \to 2$
$\overline{\mu}(b)$

Exploit: High sample reward arms are favoured
Explore: Least played arms are favoured

$\mu^* T - E\left[\sum_{t=1}^{T} R_t\right]$

3 → ETC
→ ε-greedy
→ UCB

Photos provided by Pexels

Contextual Bandits – Multiple states

- News article Recommendation systems

- Articles – arms
- Like / Dislike – Reward
- User – State

Different users have different preferences to articles

Photos provided by Pexels