Plan

Recap

Reminders

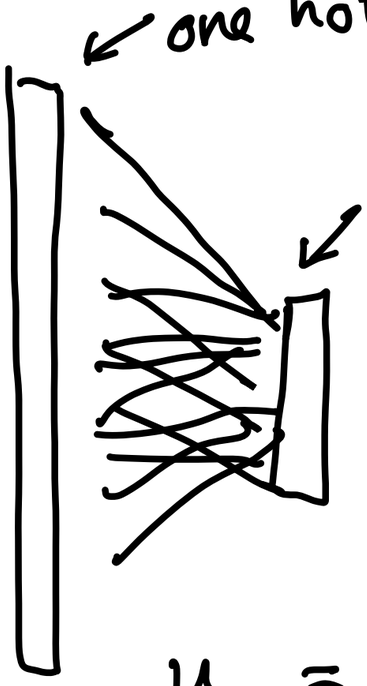
Reinforcement Learning

Policy Gradients

Recap

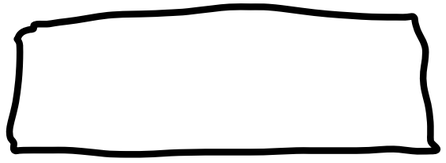
How do we embed text?

Word2Vec with Contrastive Learning



$f(x) \approx f(x')$

if x, x' are close

$u =$ 

$f(x) = ux$

GloVe

highly heuristic

parameters

$X \in \mathbb{R}^{n \times n}$

$U, V \in \mathbb{R}^{\text{embed} \times n}$

$b, c \in \mathbb{R}^n$

heuristic scalar

$$\mathcal{L}(U, V, b, c) = \frac{1}{2} f(X_{ij}) \underbrace{\left(u_i^T v_j + b_i + c_j - \log(X_{ij})\right)^2}_{\text{cost}}$$

$$\frac{\partial \mathcal{L}}{\partial u_i} = f(X_{ij}) \cdot \text{cost} \cdot v_j$$

$$\frac{\partial \mathcal{L}}{\partial v_j} = f(X_{ij}) \cdot \text{cost} \cdot u_i$$

update

$$u_i \leftarrow u_i - \eta \frac{\partial \mathcal{L}}{\partial u_i}$$

$$\frac{\partial \mathcal{L}}{\partial u_i} = f(X_{ij}) \cdot \text{cost}$$

Logistics

Academic Integrity

↳ work together ☺

↳ write solutions by yourself

☺ white board

error debugging

☹ show solution

copy prose/code

Lots of violations

↳ no learning

↳ headache

Self grade 5pm

1) ☺? 2) ~?

3) Did you share/copy on this problem?
Plan to prevent.

Forms: 22/26

Extra Credit: Thanks!

↳ good lectures, helpful recap

↳ slower demos

↳ more direction on HW

- easier last problem
- more prompt

Machine Learning

Supervised (labelled data)

Unsupervised (embedding, generative)

But real world is online:

↳ game

↳ movement robot

↳ self driving car

↳ financial trading with bots

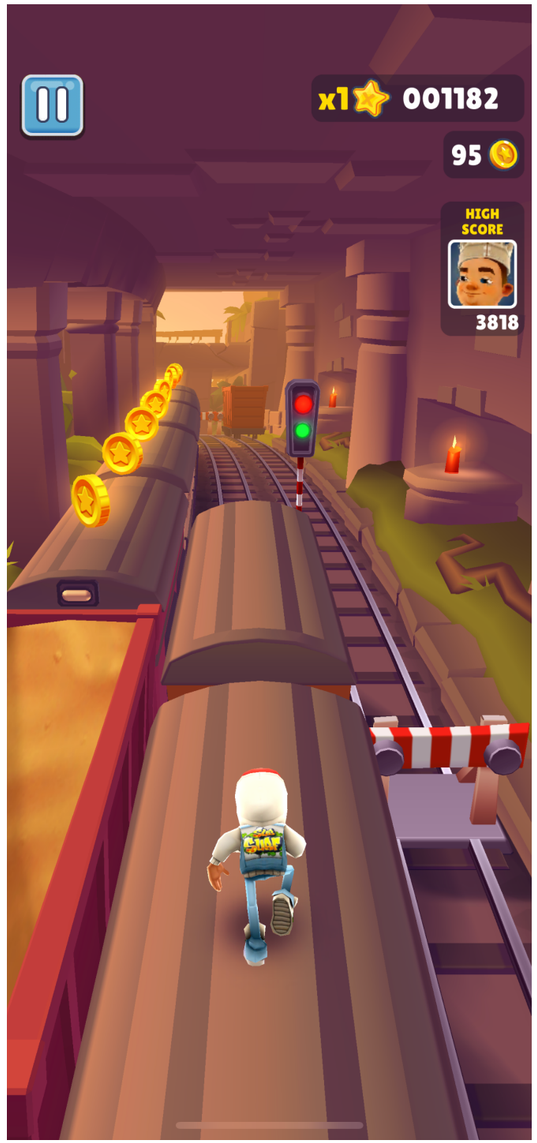
Reinforcement Learning

↳ Alpha Go

↳ DOTA2

Intuition: Humans learn through experimentation and observation

Subway Surfer



State $\xrightarrow{\text{action}}$ state $\xrightarrow{\text{action}}$ state

Environment
- height
- lane
- obstacle

Action
- left
- right
- up
- down
- stay

Reward
- coins
- not being caught
- letters

Proposal due 5pm today!

π maps states to action

s a $f(s,a) = s', r$

↑ ↑ ↑ ↑

state action new state reward

Goal: learn a policy π for choosing actions in states

Learn by trajectory

$s_0, a_0, s_1, a_1, s_2, a_2,$

Stochastic

↳ environment

↳ reward

↳ policy

Supervised	Unsupervised

RL

generate action which becomes part of training

Policy Gradients

minimize θ — expected reward on trajectory following π_θ

(θ: parameters of π)

$$\min_\theta - \mathop{\mathbb{E}}_{\tau\sim\pi}[R(\tau)]$$

$$= \min_\theta - \sum_\tau R(\tau)\cdot\pi(\tau)$$

$$R(\tau) = \sum_{i=0}^{|\tau|} r(s_i, a_i)$$

$$\frac{\partial}{\partial\theta}\mathop{\mathbb{E}}_{\pi}[R(\tau)] = \frac{\partial}{\partial\theta}\left(\sum_\tau R(\tau)\cdot\pi(\tau)\right)$$

$$= \sum_\tau \frac{\partial}{\partial\theta}\left(R(\tau)\cdot\pi(\tau)\right)$$

$$= \sum_\tau R(\tau)\cdot\frac{\partial\pi(\tau)}{\partial\theta}$$

$$= \sum_\tau R(\tau)\cdot\frac{\partial\log(\pi(\tau))}{\partial\theta}\pi(\tau)$$

$$= \mathbb{E}\left[R(\tau)\cdot\frac{\partial\log\pi(\tau)}{\partial\theta}\right]$$

$$\frac{\partial\log\pi(\tau)}{\partial\theta} = \frac{1}{\pi(\tau)}\cdot\frac{\partial\pi(\tau)}{\partial\theta}$$

$$\Rightarrow \frac{\partial\pi(\tau)}{\partial\theta} = \pi(\tau)\frac{\partial\log(\pi(\tau))}{\partial\theta}$$

$$\frac{\partial}{\partial\theta}\mathop{\mathbb{E}}_{\pi}[R(\tau)] = \mathop{\mathbb{E}}_{\pi}\left[R(\tau)\cdot\frac{\partial \log \pi(\tau)}{\partial\theta}\right]$$

$$\mathcal{L}(\theta) = -\mathop{\mathbb{E}}_{\tau:\pi}[R(\tau)]$$

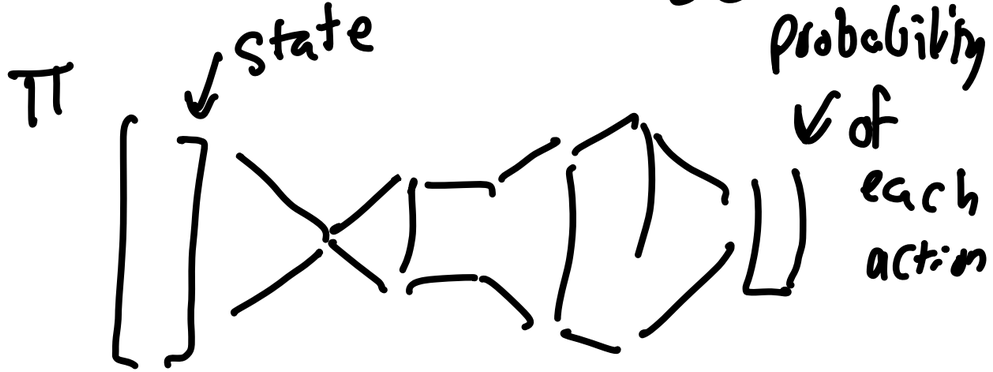
$$\theta \leftarrow \theta - \eta\frac{\partial\mathcal{L}(\theta)}{\partial\theta}$$

$$\Leftrightarrow$$

$$\theta \leftarrow \theta + \eta\,\mathbb{E}\left[R(\tau)\cdot\frac{\partial \log \pi(\tau)}{\partial\theta}\right]$$

REINFORCE

1. Sample trajectory according to π
   $\tau = (s_0, a_0, s_1, a_1, \ldots)$
2. Compute $R(\tau) = \sum_{t=0}^{|\tau|} r(s_t, a_t)$
3. $\theta \leftarrow \theta + \eta\, R(\tau)\cdot\frac{\partial \log \pi(\tau)}{\partial\theta}$



$$\frac{\partial\mathcal{L}(\theta)}{\partial\theta} = -\frac{\partial\mathbb{E}[R(\tau)]}{\partial\theta} = -\mathbb{E}\left[R(\tau)\cdot\frac{\partial \log \pi(\tau)}{\partial\theta}\right]$$