CSE 417T
Introduction to Machine Learning

Lecture 23

Instructor: Chien-Ju (CJ) Ho

Logistics

- Homework 5 is due Apr 19 (Tuesday)

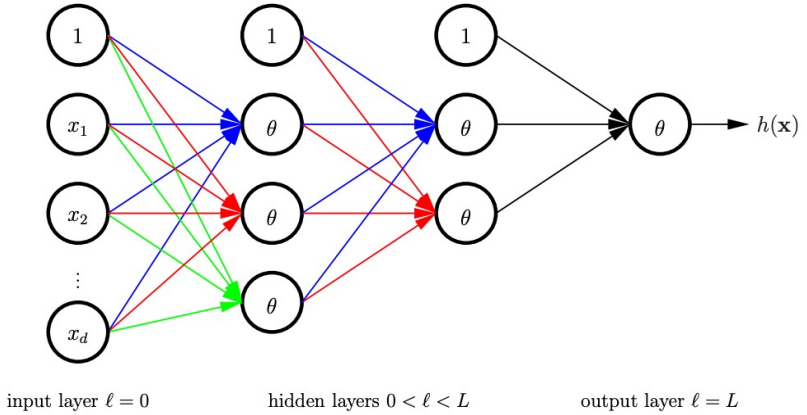
- Exam 2 will be on April 28 (Thursday)
  - Will focus on the topics in the second half of the semester
  - Format / logistics will be similar with what we have in Exam 1
    - Timed exam (75 min) during lecture time in the classroom
    - Closed-book exam with 2 letter-size cheat sheets allowed (4 pages in total)
      - No format limitations (it can be typed, written, or a combination)
  - April 26 (Tuesday) will be a review lecture

Recap

Neural Network

- Evaluate $h(\vec{x})$ given h (characterized by $\left\{w_{i,j}^{(\ell)}\right\}$)
  - Forward propagation

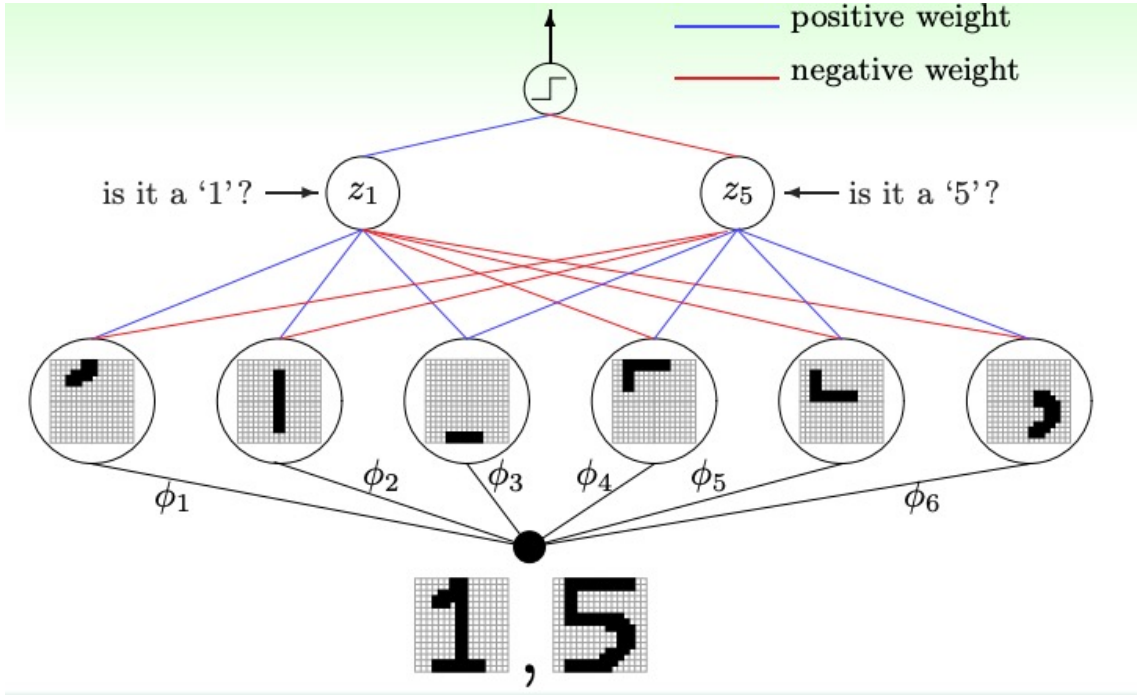
$$\mathbf{x} = \mathbf{x}^{(0)} \xrightarrow{\mathrm{W}^{(1)}} \mathbf{s}^{(1)} \xrightarrow{\theta} \mathbf{x}^{(1)} \xrightarrow{\mathrm{W}^{(2)}} \mathbf{s}^{(2)} \xrightarrow{\theta} \mathbf{x}^{(2)} \cdots \xrightarrow{\mathrm{W}^{(L)}} \mathbf{s}^{(L)} \xrightarrow{\theta} \mathbf{x}^{(L)} = h(\mathbf{x}).$$



- Given D, learn the weights $W = \left\{w_{i,j}^{(\ell)}\right\}$
  - Backpropagation
  - Initialize $w_{i,j}^{(\ell)}$ randomly
  - For $t = 1$ to T
    - Randomly pick a point from D (for stochastic gradient descent)
    - Forward propagation: Calculate all $x_i^{(\ell)}$ and $s_i^{(\ell)}$
    - Backward propagation: Calculate all $\delta_j^{(\ell)}$
    - Update the weights $w_{i,j}^{(\ell)} \leftarrow w_{i,j}^{(\ell)} - \eta \delta_j^{(\ell)} x_i^{(\ell-1)}$
  - Return the weights

Deep Neural Network

- "Shallow" neural network is powerful (universal approximation theorem holds with a single hidden layer). Why "deep" neural networks?



Each layer captures features of the previous layers.

We can use "raw data" (e.g., pixels of an image) as input. The hidden layer are extracting the features.

Design different network architectures to incorporate domain knowledge.

Some Techniques in Improving Deep Learning

- Regularization to mitigate overfitting
  - Weight-based, early stopping, dropout, etc
- Incorporating domain knowledges
  - Network architectures (e.g., Convolutional Neural Nets)
- Improving computation with huge amount of data
  - Hardware architecture to improve parallel computation
- Improving gradient-based optimization
  - Choosing better initialization points

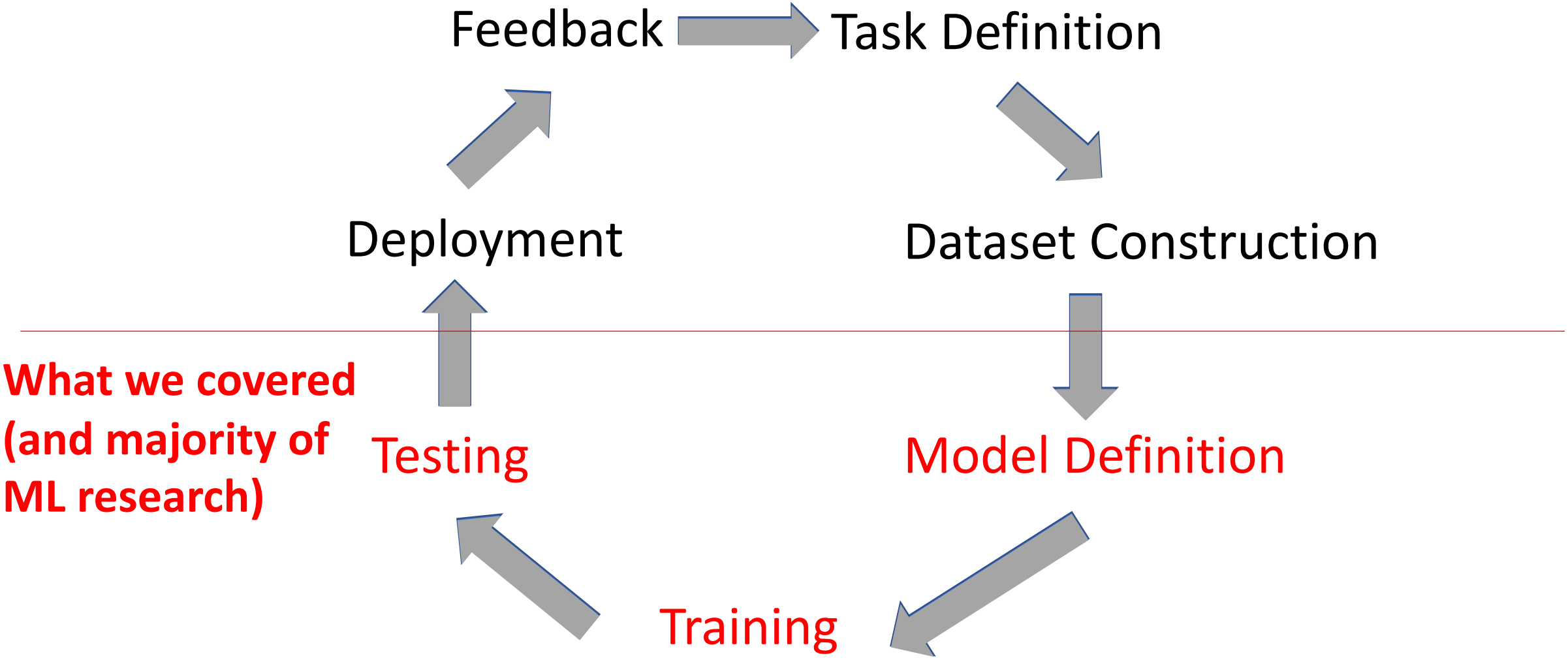
Radial Basis Functions (RBF)

- $h(\vec{x}) = \sum_{k=1}^{K} w_k \, \phi\left(\frac{\|\vec{x}-\vec{\mu}_k\|}{r}\right)$

- Connection to linear models
  - Parametric RBF is essentially linear model with nonlinear transformation
- Connection to nearest neighbor
  - RBF is based on the similarity to a set of points
- Connection to SVM with RBF Kernel
  - Using K representative points vs. using support vectors
- Connection to Neural Networks
  - RBF can be graphically represented as a one-hidden layer network

Machine Learning Lifecycle

For ML to have "positive" impacts, we need to be careful in every stage



What we covered (and majority of ML research)

Today’s Lecture

Types of Learning

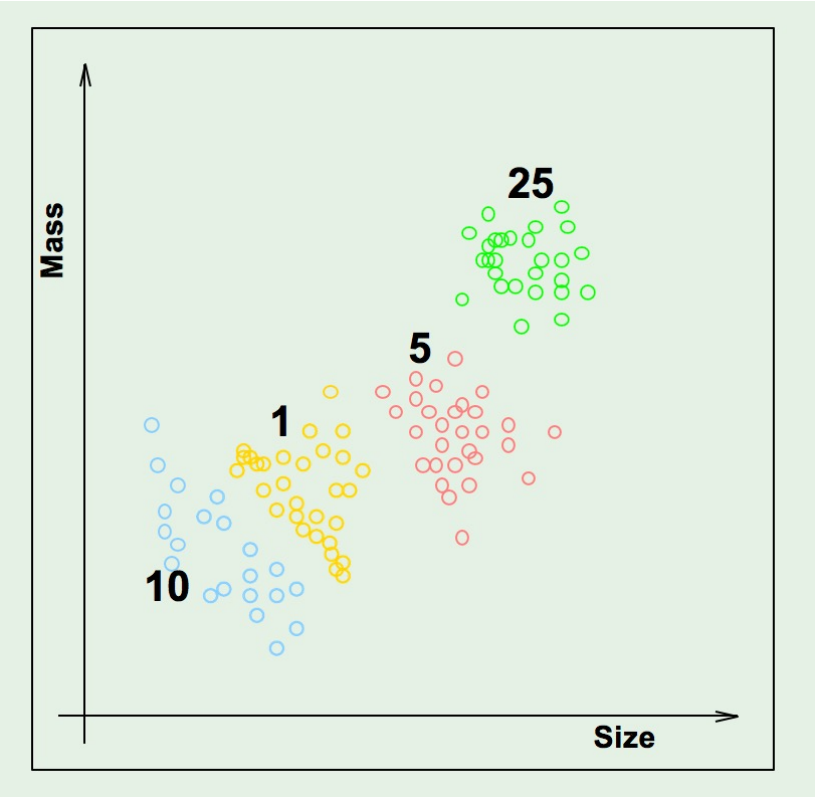
- Supervised learning (the focus of this course)
  - Given training data (input, correct output)
  - Try to predict the output for data not seen before

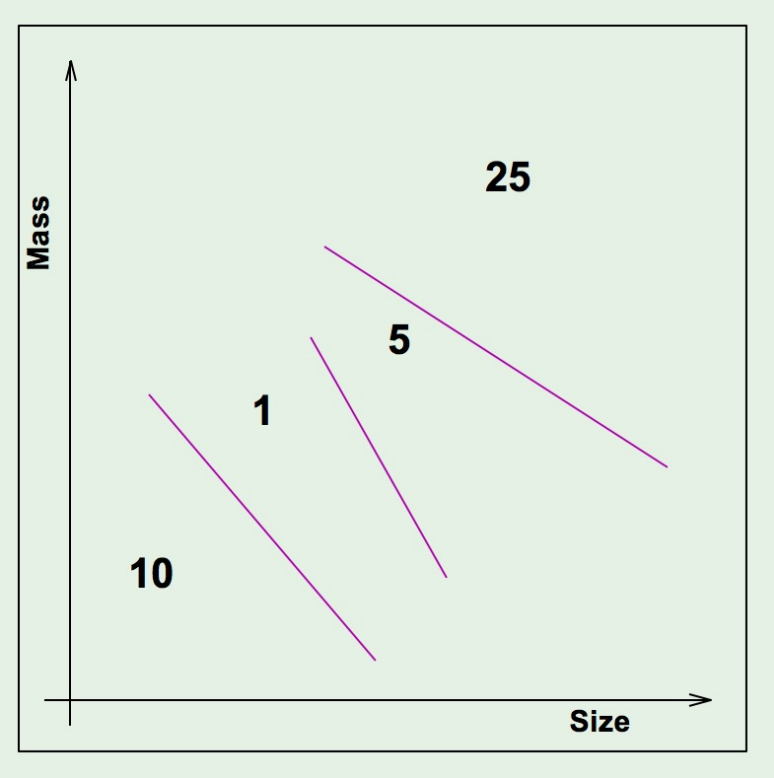
- Unsupervised learning
  - Given data in the form of (input)
  - Find patterns in data

- Reinforcement learning
  - Learn how to act, based on rewards for actions

Supervised Learning

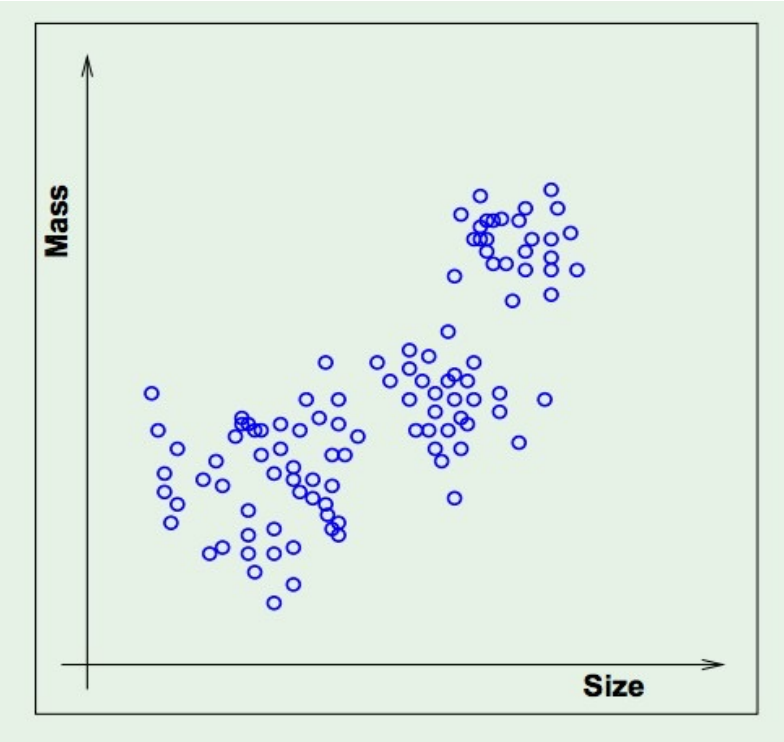
- Given data with (input, correct output), learn a pattern that can predict previously unseen data

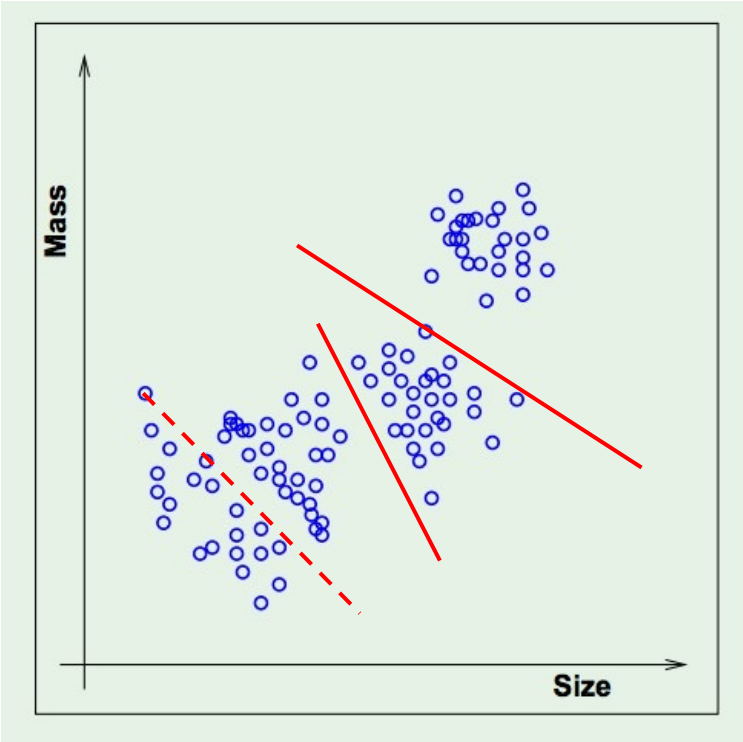




Unsupervised Learning

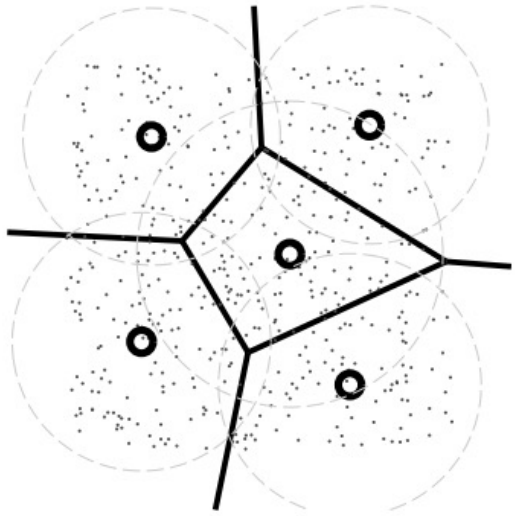
- Suppose you only have the feature vectors but no labels. Still want to describe the data in some useful way.

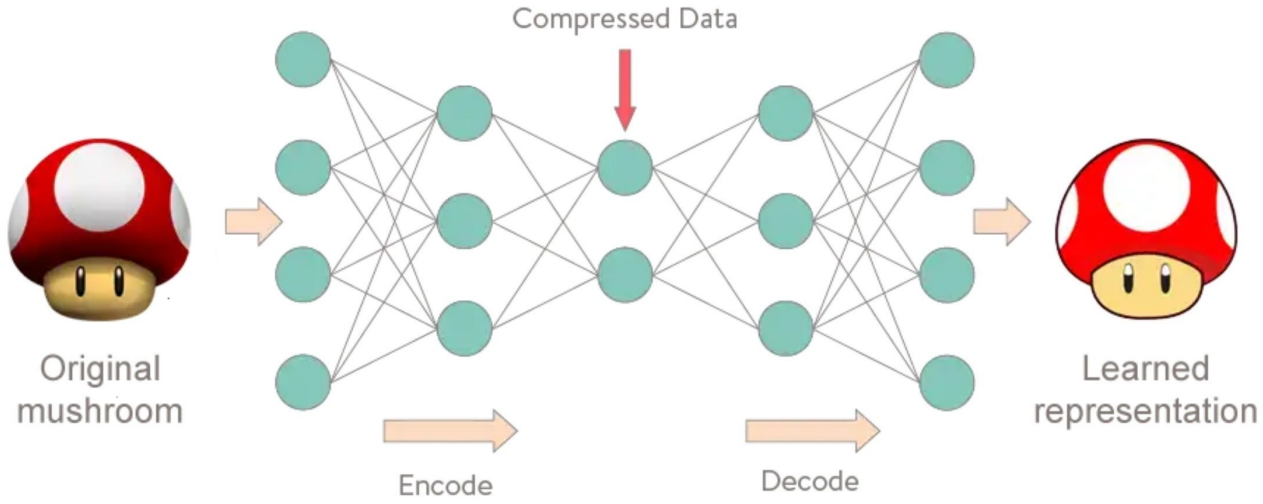




Unsupervised Learning

- Suppose you only have the feature vectors but no labels. Still want to describe the data in some useful way.





A Short Intro of Reinforcement Learning

[The entire section of reinforcement learning is **safe to skip** for the exam]

Illustrating Applications

- Consider the following tasks
  - **News site**: picks a news header to show to each user
    Goal: maximize #clicks.
  - **Dynamic pricing**: Pick a price for your product.
    Goal: maximize total profit.
  - **Personalized health advice**: Pick a recommendation of activity.
    Goal: maximize #adopted recommendations
  - **…**

- Overall goal:
  - Maximizing total rewards from sequential actions
  - Learning by doing

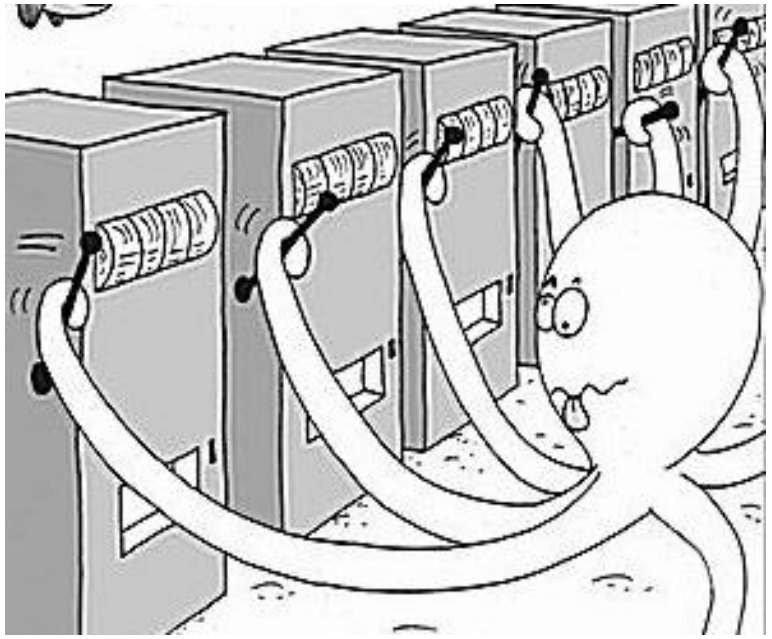
[Safe to Skip for the Exam]

Multi-Armed Bandit

One classical formulation for reinforcement learning

One-Armed Bandit = Slot Machine

Multi-Armed Bandit



- Choose a slot machine to pull at each time
- IID Rewards: each arm reward is iid drawn from unknown distribution
- Bandit feedback: Observe only the reward of your choice
- Goal:
  - Maximize the total rewards you obtain
  - Equivalently, minimize **regret**

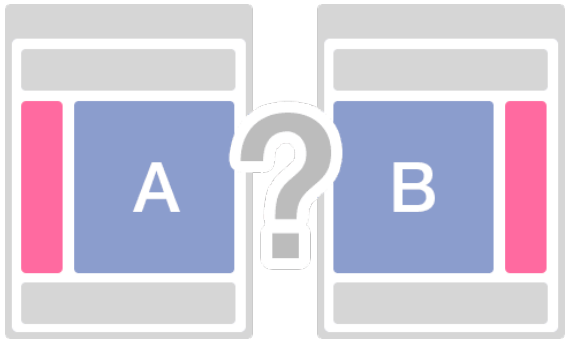
Regret:
Utility(OPT) - Utility(ALG)

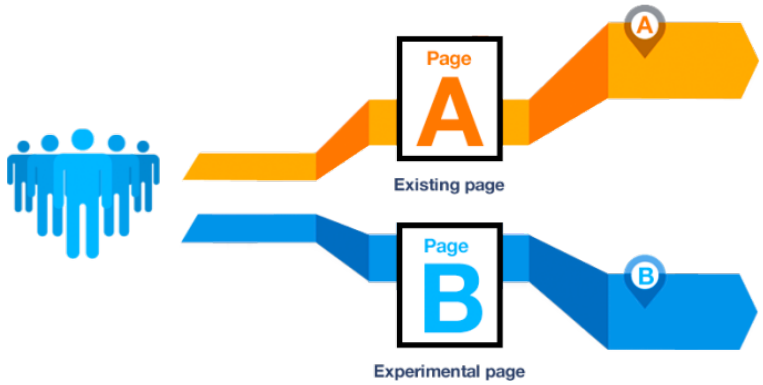
A classical framework for studying exploration-exploitation tradeoff

[Safe to Skip for the Exam]

Example

- Which interface design should I use for my website?



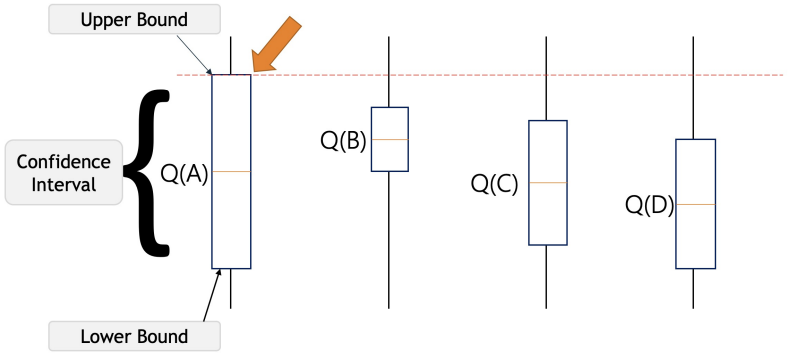


- Epsilon greedy
  - Exploration: For an initial **epsilon** population, randomly split them into seeing A or B
  - Exploitation: Choose the one with better empirical “performance” afterwards
- How much should we explore? Can we adaptively explore?

[Safe to Skip for the Exam]

Adaptive Algorithm: Upper Confidence Bound (UCB)

- An index-based method for stochastic bandits
  - Maintain an index for each arm *k* at every time *t*
  - Select the arm with the largest index



$$I_k(t) = \bar{X}_k(t) + \sqrt{\frac{L \log t}{n_k(t)}}, \forall k.$$

Empirical mean: exploitation

Confidence interval: exploration

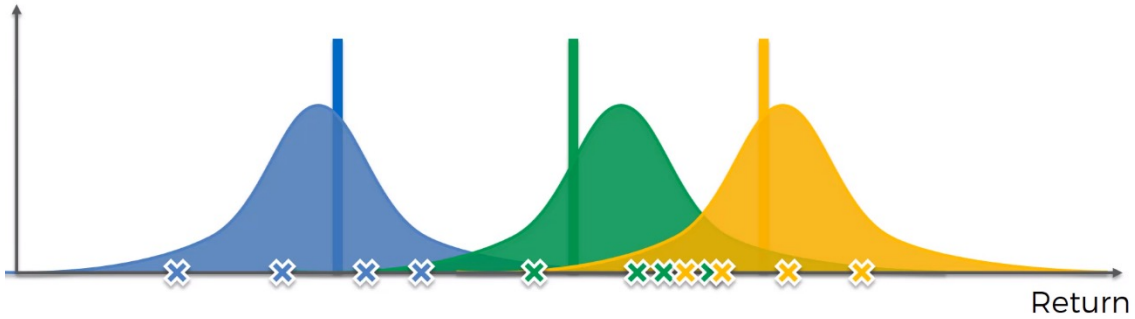
- UCB achieves regret bound *O(log T)* in stochastic settings

Common goal: Sublinear regret / no-regret:
Regret / T -> 0; converges to optimal action

[Safe to Skip for the Exam]

Adaptive Algorithm: Thompson Sampling

- Thompson Sampling Algorithm
  - Maintain reward distributions of arms
  - Sample an “index” for each arm based on the distribution



  - Update the distribution based on observations

- Properties:
  - Theoretically, it also achieves no regrets
  - Empirically, it usually has superior performance

[Safe to Skip for the Exam]

Many Variations of Multi-Armed Bandits

- We might be able to observe rewards for the actions we didn't take
  - Full-information feedback vs bandit feedback
- What if rewards are not generated stochastically
  - Adversarial bandits
- We might have additional contextual information about the rewards
  - Contextual bandits
- The rewards might depend on the "state" of the environment
  - Using Markov decision process (MDP) formulation
  - The "reinforcement learning" problem people usually refer to
  - Will talk about this next
- And more…

[Safe to Skip for the Exam]

[Safe to Skip for the Exam]

- In some ways, this actually happens everyday.

Consider the slot machine as the algorithm

- Human-generated data powers the algorithm
- The algorithm influences human welfare

NETFLIX yelp

amazon facebook

YouTube Quora twitter

[Safe to Skip for the Exam]

My Own Research: Human-in-the-Loop Bandit Learning

- **Biased** human feedback
  - The feedback received for selecting an arm might be biased by humans
- Learning as **incentives**
  - The number of possible actions could increase when humans are involved
- **Ethical** implications
  - Long-term impacts of actions in bandits
  - Privacy-preserving online learning
  - Secure sequential learning

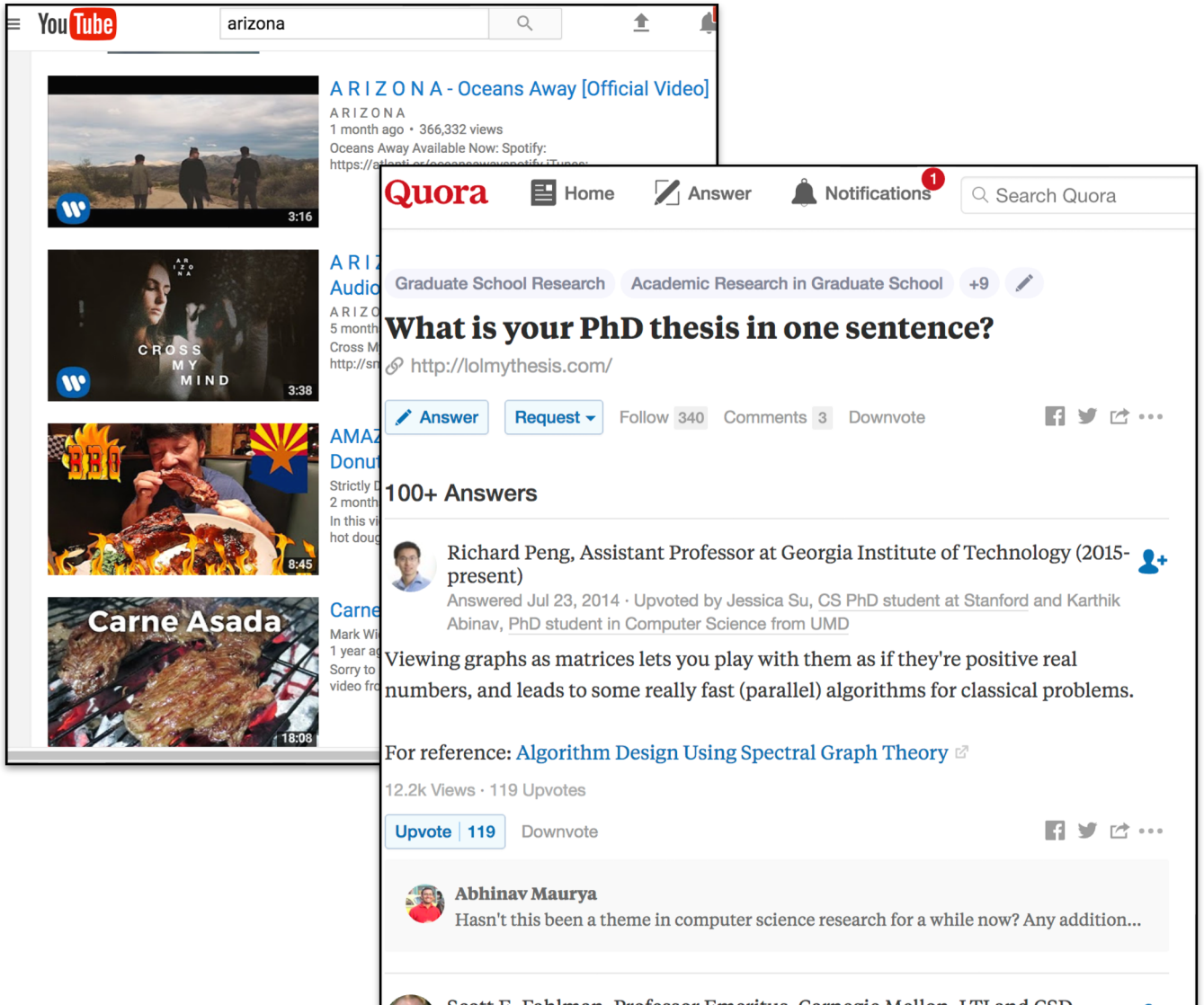
[Safe to Skip for the Exam]

Bandit Learning with Biased Feedback

joint work with Wei Tang

[Safe to Skip for the Exam]

User Generated Content Platforms



1,504,905 views

42K 1K

12.2k Views · 119 Upvotes

A Bandit Formulation

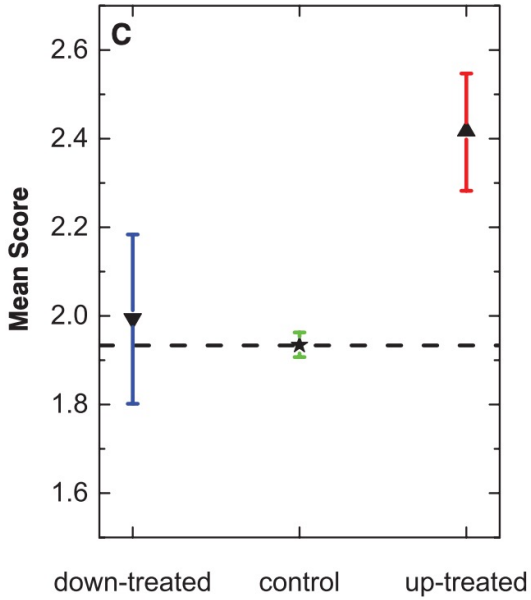
- When each new user arrives
  - Show the user some (set of) content
  - Obtain feedback (upvotes, likes, shares, etc) from the user

- Goal:
  - maximize total user happiness
    - Which may **NOT** be the same as the total number of positive feedback

[Safe to Skip for the Exam]

Users' feedback might be biased

- In a Reddit-like platform, randomly insert an upvote/downvote to some posts right after they are posted.

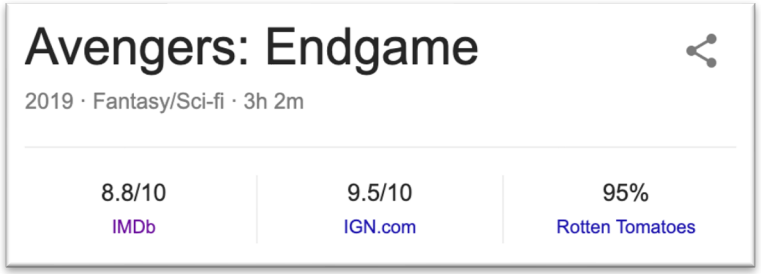


Social Influence Bias: A Randomized Experiment. Muchnik et al. Science 2013.

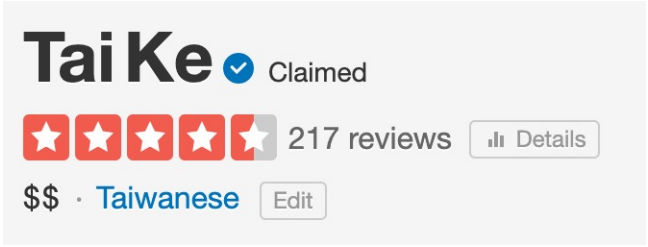
[Safe to Skip for the Exam]

Main Results

- Explore two general set of bias models
- Model 1: feedback is biased by empirical average
  - Designing no-regret bandit learning (i.e., near-optimal learning) is possible



- Model 2: feedback is biased by the whole history
  - Impossible to achieve no-regret learning



[Safe to Skip for the Exam]

Summary

- The models are very stylized, but we show that a small deviation of human behavior could lead to very different outcomes for machine learning.

- Implications: with human biases in data generation, sometimes even with infinitely amount of data, we won’t be able to learn.

[Safe to Skip for the Exam]

New Arm Generation in Bandit Learning:

An application to User Generated Content Platforms

Joint work with Yang Liu

[Safe to Skip for the Exam]

Users are Both Raters and Contributors

- When each new user arrives
  - Show the user some (set of) content
  - Obtain feedback (upvotes, likes, shares, etc) from the user
  - The user decides whether to contribute new content

Assume user feedback is "unbiased".

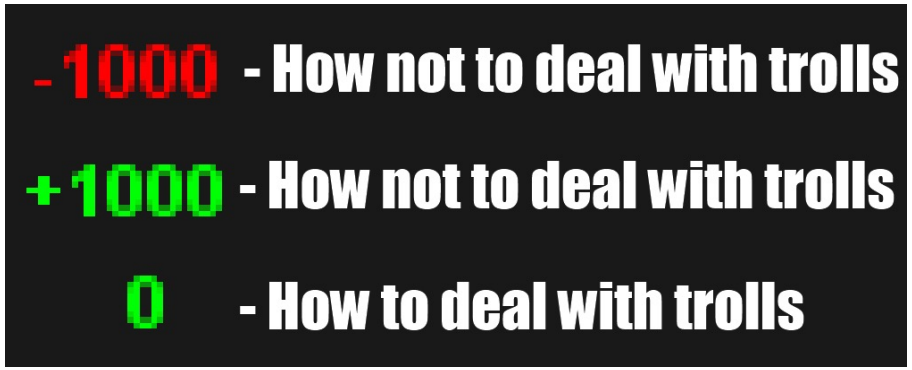
- Goal:
  - maximize total user happiness (total number of positive feedback)

- A standard bandit learning problem
  - arm: the content chosen to show to users

[Safe to Skip for the Exam]

Why Users Contribute?

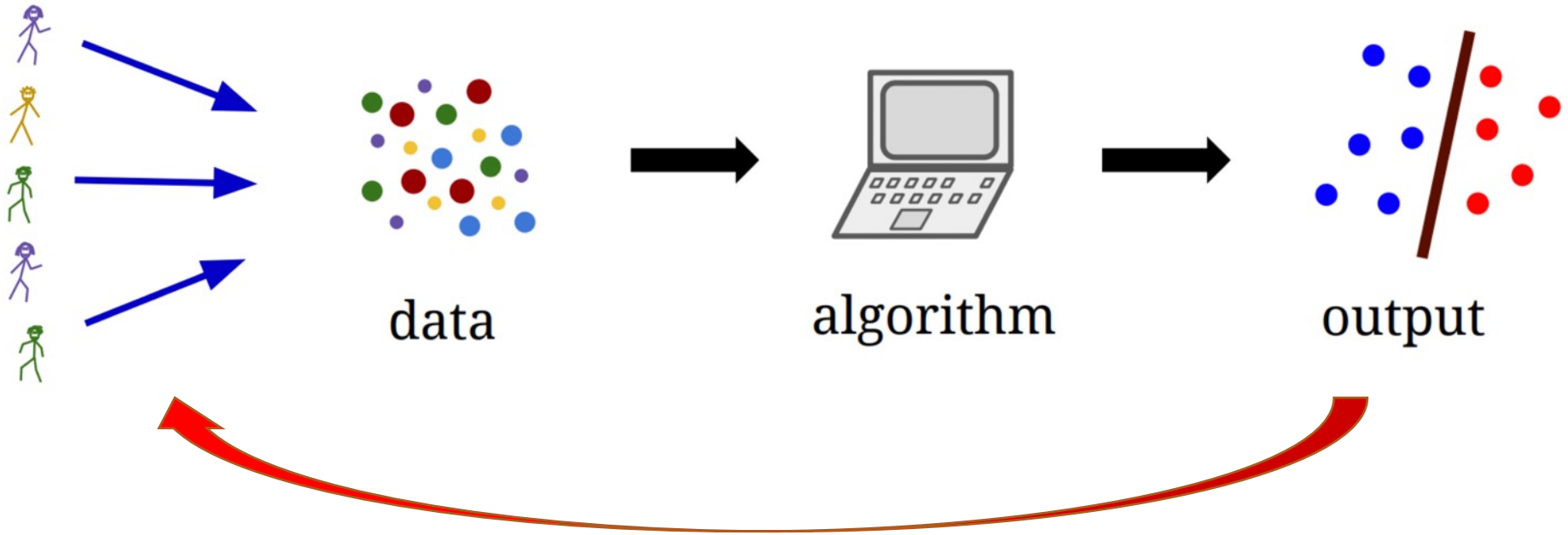
- Model:
  - Users likes attention (e.g., attention => money)



  - Users aim to maximize
    (Total # views of their content) – **(Cost for contributing)**

Need to consider **learning** and **incentives** simultaneously when designing bandit algorithms.

A Broader Research Agenda: Strategic ML



Goodhart's law:

"If a measure becomes the public's goal, it is no longer a good measure."

[Safe to Skip for the Exam]

History-Dependent Bandits and The Fairness Implications

Joint work with Wei Tang and Yang Liu

[Safe to Skip for the Exam]

Search for “CEO” on search engines

[Kay et al., 2015]

[Safe to Skip for the Exam]

Stereotype Mirroring and Exaggeration

- Is this result mirroring the real statistics or an exaggeration?

- Assume this is mirroring of the real statistics, are there other concerns?
  - Are we reinforcing the stereotypes?
  - Are we being “unfair” to disadvantage groups that are mistreated in the past?

Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. Kay et al. CHI 2015.

In the context of bandit learning

- For every action we take, we need to consider its long term impacts.

- Examples
  - **Loan applications:**
    Give loans to disadvantage groups could improve their social status in the future.
  - **College admission:**
    The long debate about affirmative action.
  - **Labor markets:**
    Workers might get bored if kept assigned the same type of tasks.
  - **Online advertisements:**
    An advertisement might be less effective if displayed too frequently (over-exposure).

[Safe to Skip for the Exam]

Negative results

- Deploying standard bandit algorithms leads to linear regrets.
  - Simply applying standard algorithms won’t work
- Even without considering fairness issues, reinforcing stereotypes is bad for the society from the utility perspective.

[Safe to Skip for the Exam]

What Can We Do?

- Rooney rule:
  - NFL policy - when interviewing head coaches, it is required to include ethnic-minority candidates.
  - We need to give each “arm” a chance since they might be mis-treated in the past
  - A special case of **affirmative action**

- There exists setting that applying Rooney rule achieves better overall utility even when each individual decision is sub-optimal.

- Main results: Designed order-optimal algorithms

[Safe to Skip for the Exam]

Additional Notes: Achieving Fairness is Hard

- There are many possible/reasonable fairness definitions, and they are often not compatible!
  - FAT*18 Tutorial: 21 Definitions of Fairness and Their Politics. Arvind Narayanan.

- Require conversations from people across different disciplines

[Safe to Skip for the Exam]

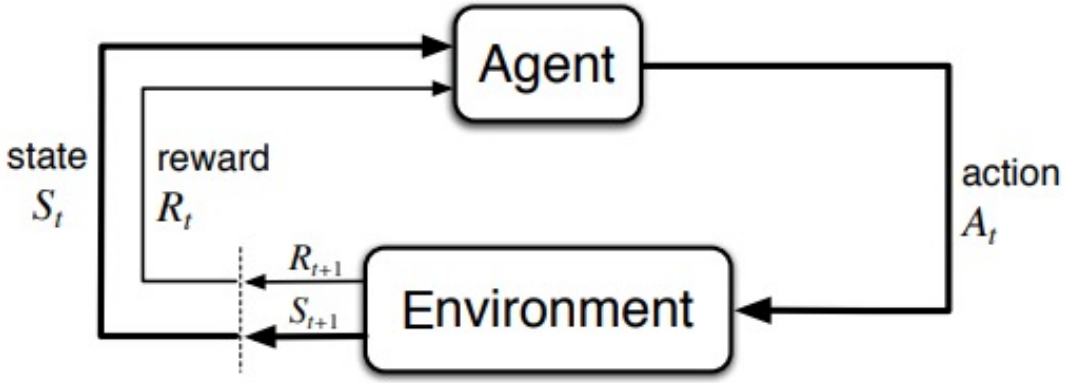
Markov Decision Process (MDP)

The common formulation for reinforcement learning

[Safe to Skip for the Exam]

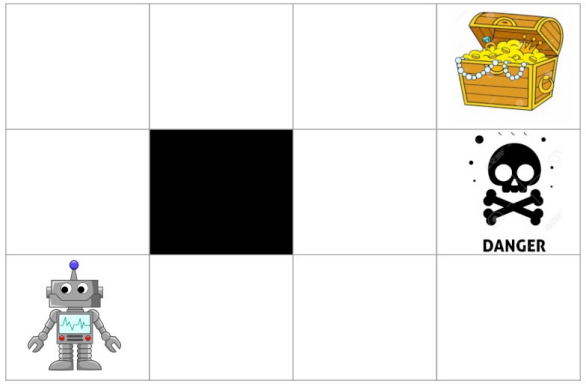
Markov Decision Process

- In multi-armed bandits, the rewards depends on actions only (there are some variations to extend this idea)
- In many practical applications, the rewards depend on both the actions and the state of the environment.



[Safe to Skip for the Exam]

Markov Decision Process



- MDP formulation
  - State space $s \in S$
    - The state of the environment, e.g., your location on the map.
  - Action space $a \in A$
    - The actions you can take, e.g., move up, move down, etc.
  - State transition $P(s', s, a)$
    - The probability of moving to state s' after taking action a in state s
  - Reward function $R(s, a)$
    - The reward obtained by taking action a in state s
- Goal:
  - Find a policy $\pi(a|s)$ (denoting the probability of taking action a in state s) that maximizes total (possibly time-discounted) rewards over time.

[Safe to Skip for the Exam]

Solving MDP

- Classical methods
  - Decompose the rewards as “current rewards” + “expected future rewards”
    - Dynamic programming
    - Value iteration
    - …
- Deep reinforcement learning
  - Use neural networks to approximate important functions; apply deep learning through sampling data
    - Deep Q-learning
    - Policy Gradient
    - …

[Safe to Skip for the Exam]

Rich Space of Studies

[Safe to Skip for the Exam]

Machine Learning Lifecycle