KDD CUP 2019: Humanity RL Track Solution by Alpha

Speaker: Suiqian Luo

Guazi

August 6, 2019

Overview

1. Introduction
2. Environment Analysis
   - Feedback phase
   - Check phase
   - Verification phase
3. Solution
4. About Guazi

Overview

1. Introduction
2. Environment Analysis
   - Feedback phase
   - Check phase
   - Verification phase
3. Solution
4. About Guazi

Background

- Policy learning for malaria control in Sub Saharan Africa
- Sequential decision making task
- Applying machine learning tools to determine novel solutions

Evaluation

- Final score
  - Median of reward scores from 10 instantiations
- Instantiation
  - 21 episodes in an instantiation
  - Scores in previous episodes can be used
  - Maximum score from all episodes
- Episode
  - 5 sequential actions in an episode
  - Action represented by two real numbers between 0 and 1
  - Denoted by $(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), (x_5, y_5)$

Overview

1 Introduction

2 Environment Analysis
- Feedback phase
- Check phase
- Verification phase

3 Solution

4 About Guazi

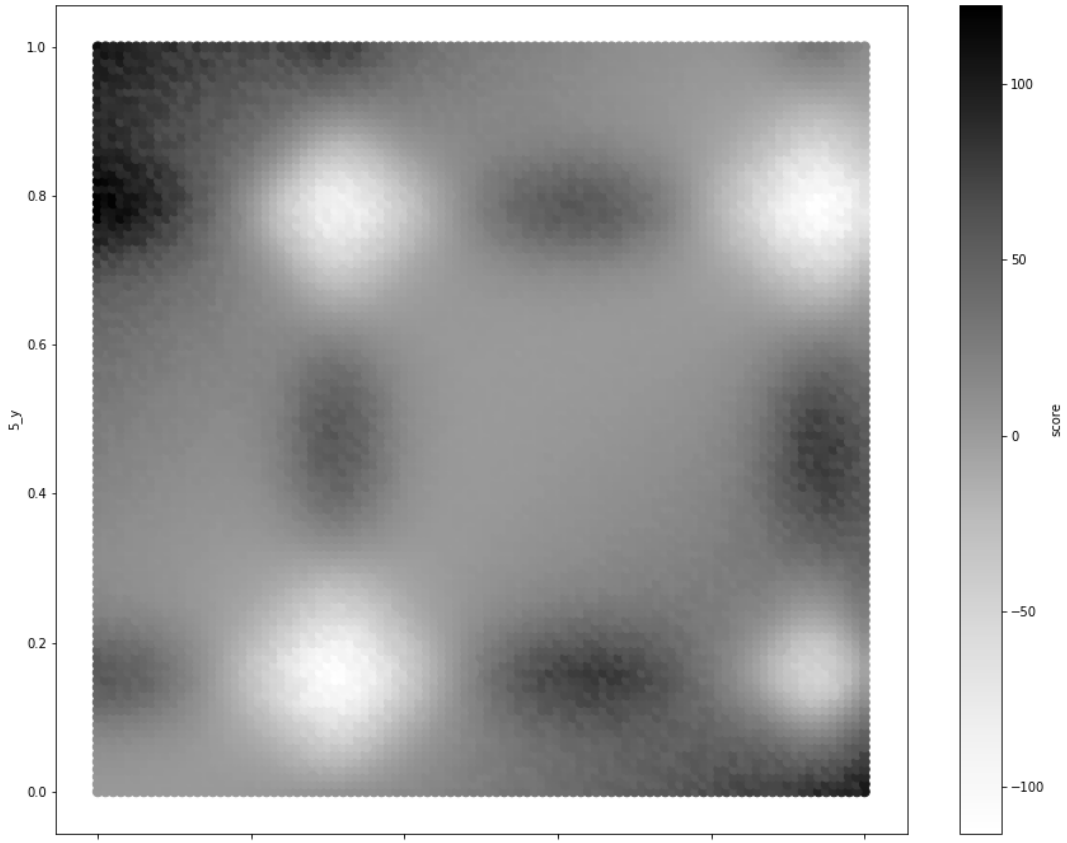
Environment analysis

- Three interesting environments in total
- The understanding of environments is critical
- No background material about environments

Feedback phase

- How does the 5^{th} action affect the final reward?
  - Fix the first 4 actions with $(0.0, 0.0)$

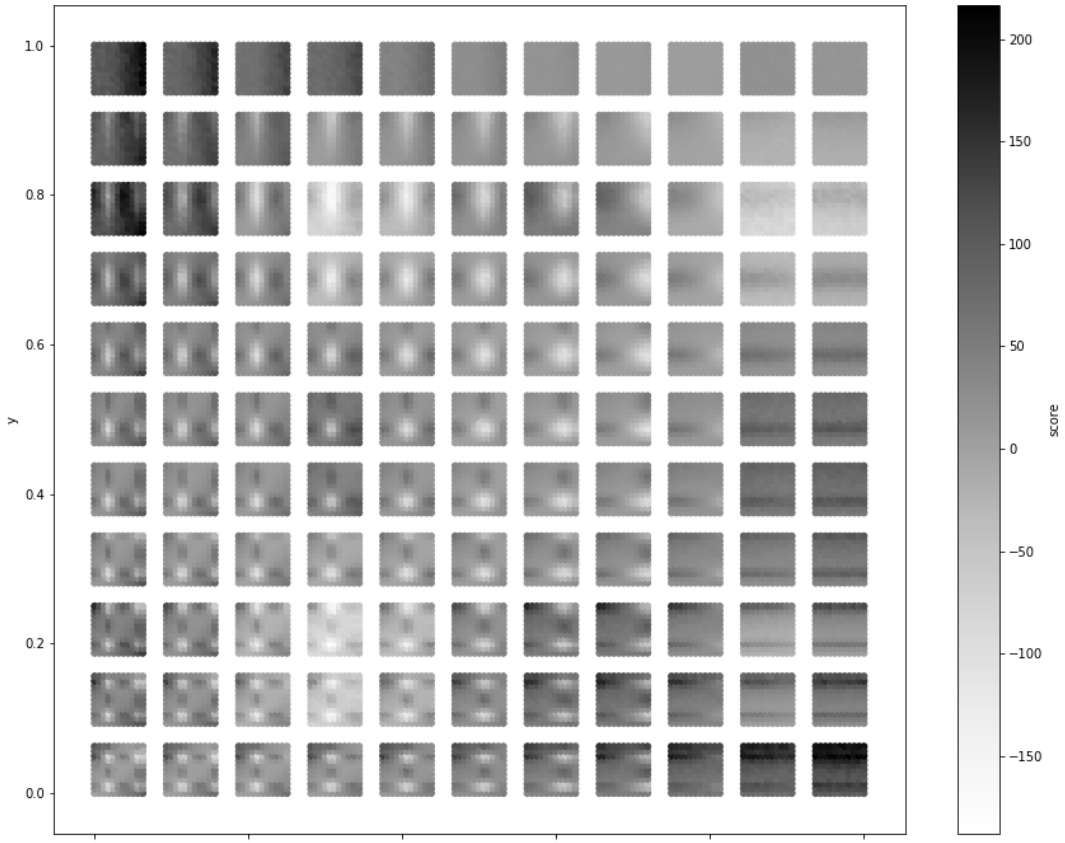
Feedback phase



Feedback phase

- How does the 4^{th} action effect the final reward pattern?
  - Fix the first 3 actions with $(0.0, 0.0)$

Feedback phase



Assumption

- Basic pattern $f(x, y)$
- Reward

$$
\begin{aligned}
score \quad = \quad & f(x_1, y_1) \\
& + f((1 - x_1) \cdot x_2, (1 - y_1) \cdot y_2) \\
& + f((1 - x_2) \cdot x_3, (1 - y_2) \cdot y_3) \\
& + f((1 - x_3) \cdot x_4, (1 - y_3) \cdot y_4) \\
& + f((1 - x_4) \cdot x_5, (1 - y_4) \cdot y_5)
\end{aligned}
$$

- Confirmed by random trials

Dynamic programming

- Let $H(k, x, y)$ denote the highest score
  - From $(k+1)^{th}$ action to 5^{th} action
  - Given that the k^{th} action is (x, y)
- Transition equation

$$H(k, x, y) = \max_{a,b}\{H(k+1, a, b) + f((1-x)a, (1-y)b)\}$$

  - $H(5, x, y) = 0$

Dynamic programming

- The goal is $H(0, 0, 0)$
- The best score is around these actions

$$(0, 0.79), (1, 0), (0, 0.79), (1, 0), (0, 0.79)$$

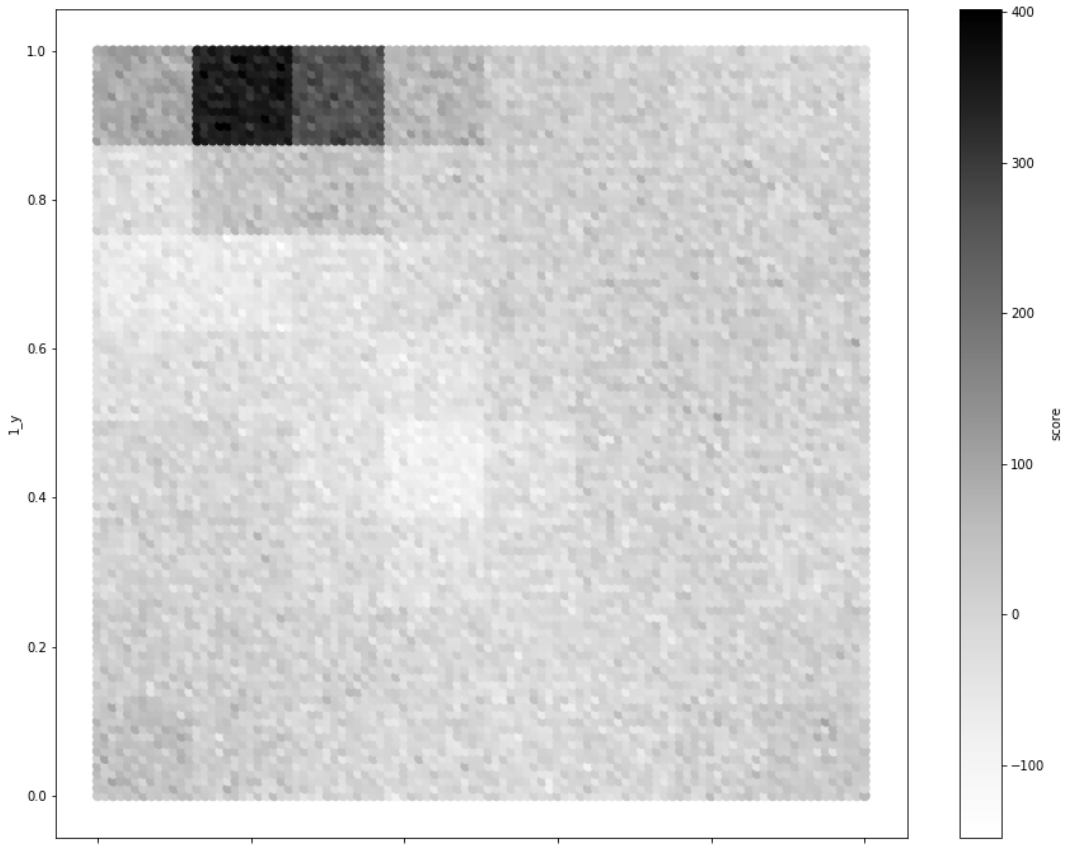
Check phase

- Use the same assumption in previous phase

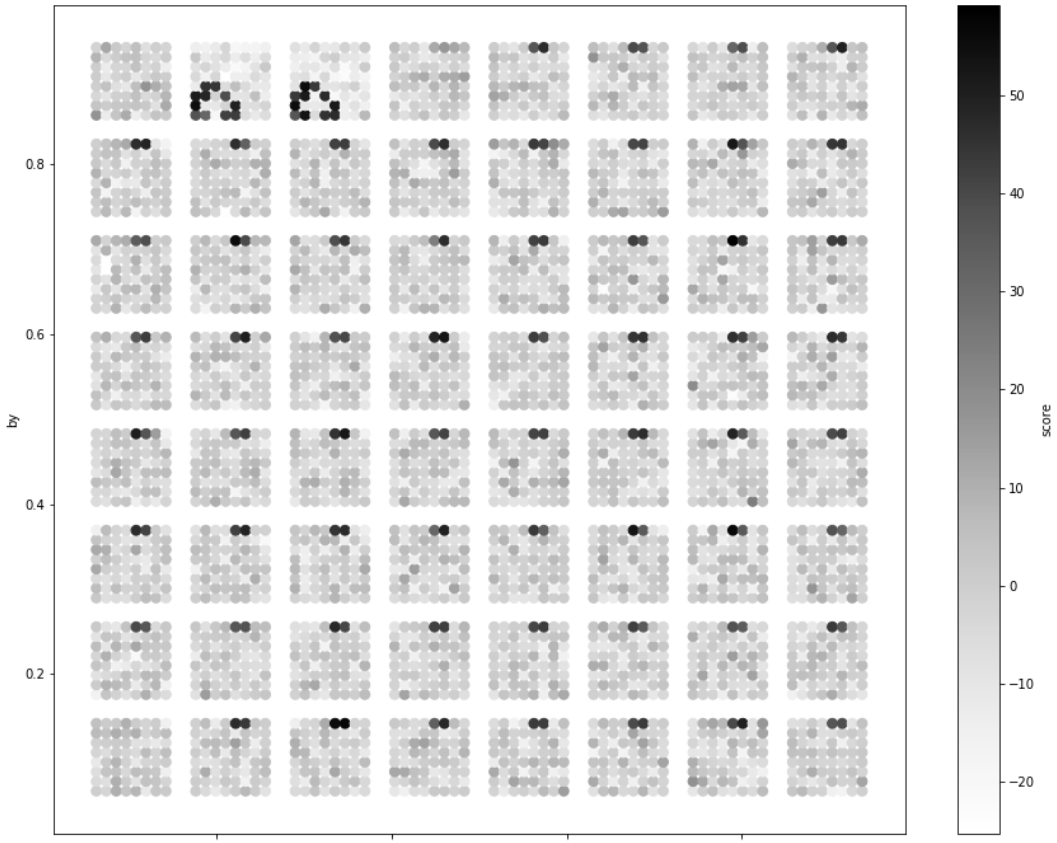
$$\begin{aligned} score &= f(x_1, y_1) \\ &+ f(x_1, x_2, y_1, y_2) \\ &+ f(x_2, x_3, y_2, y_3) \\ &+ f(x_3, x_4, y_3, y_4) \\ &+ f(x_4, x_5, y_4, y_5) \end{aligned}$$

- Try to detect if there is any pattern of score we can get

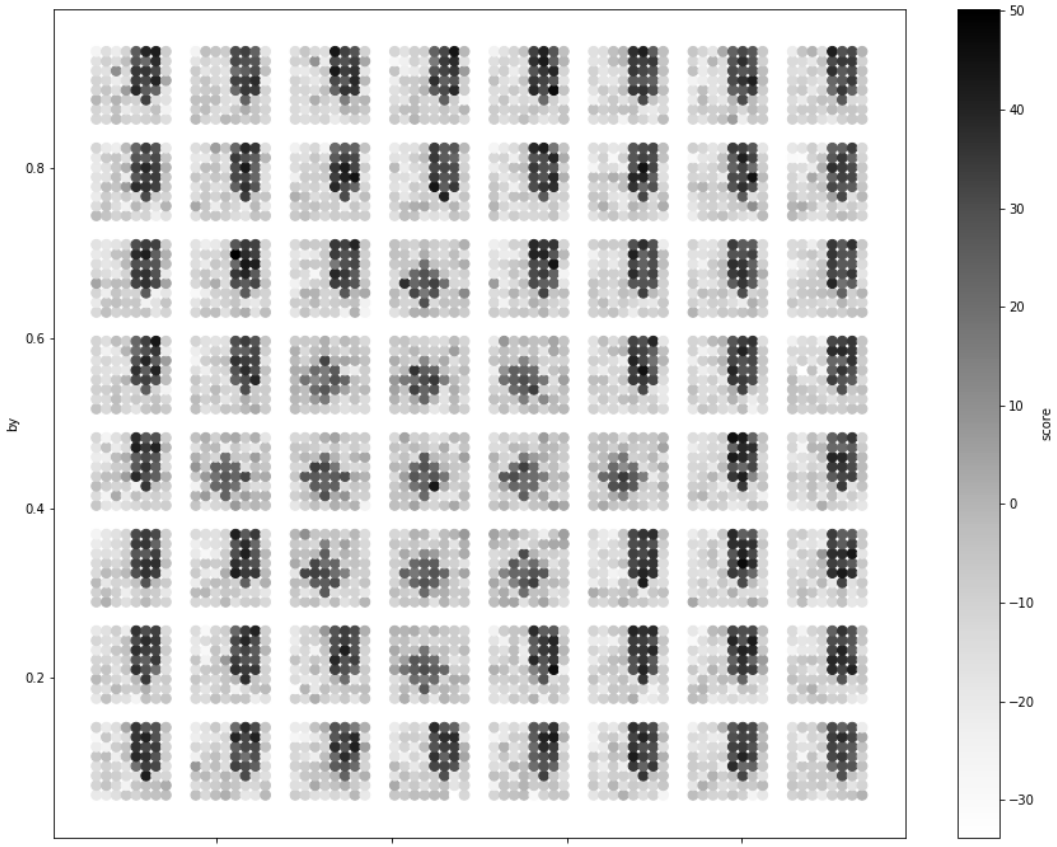
First action



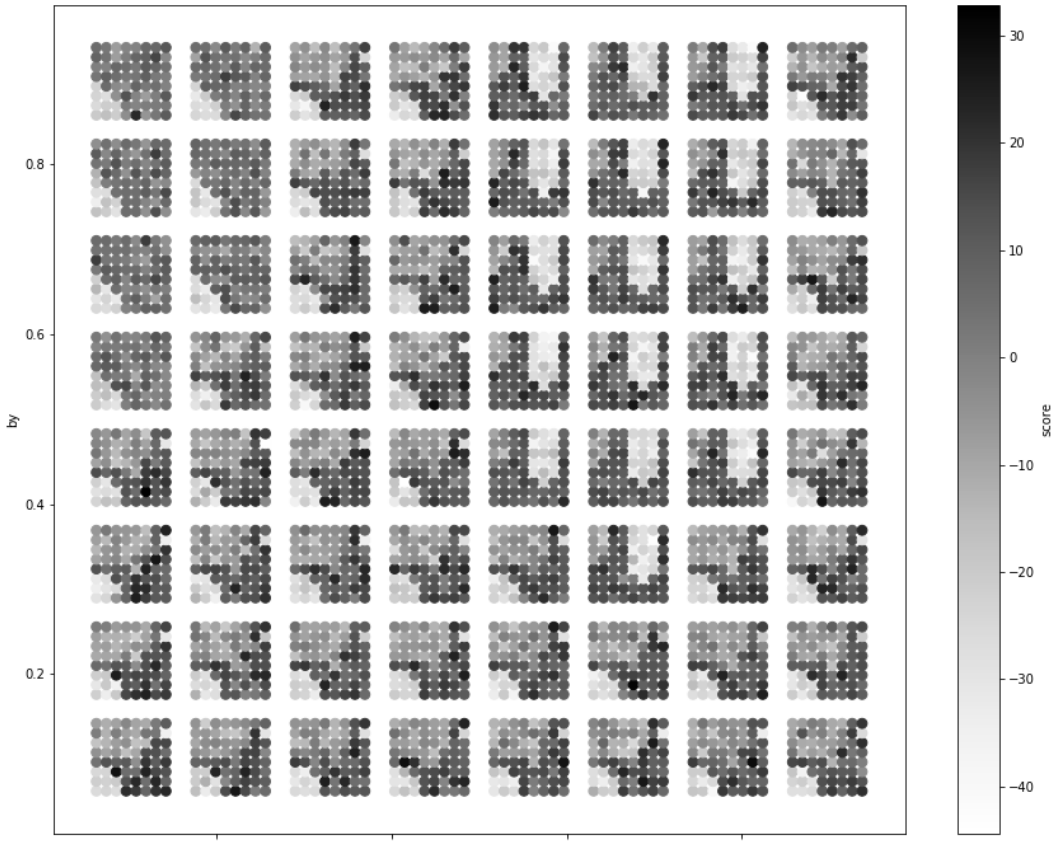
First and second action



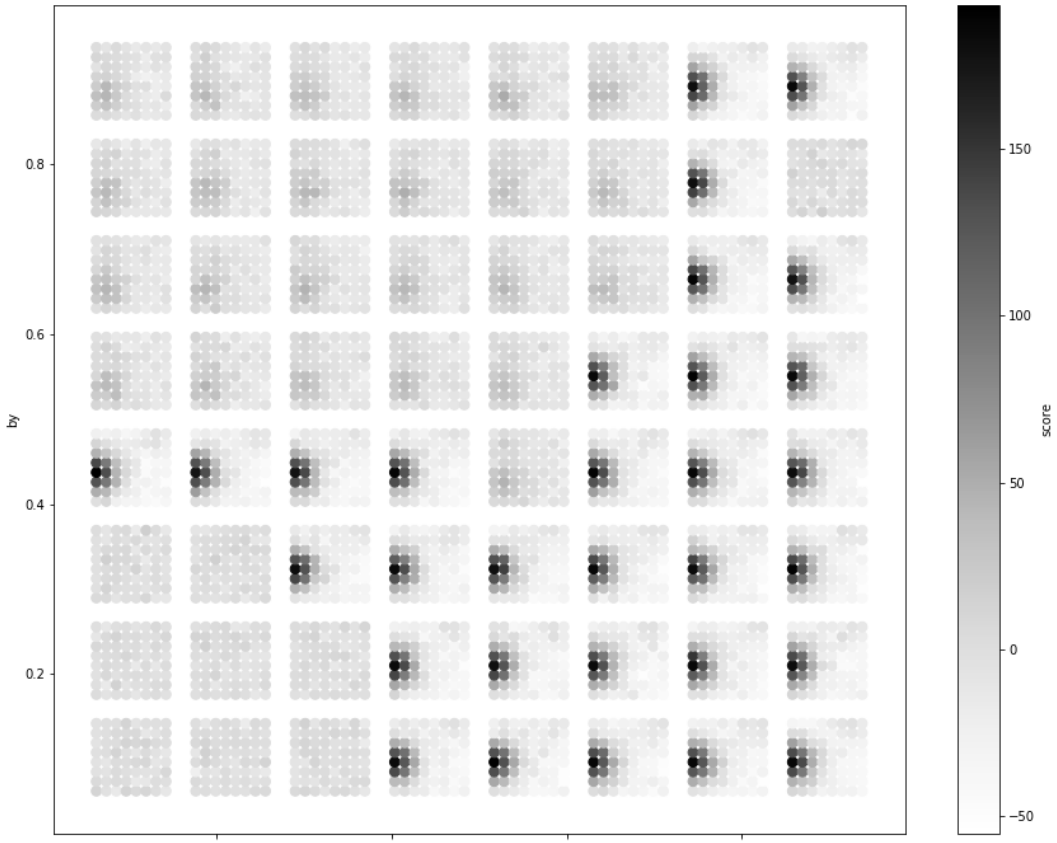
Second and third action



Third and fourth action



Fourth and fifth action



Check phase

- 8×8 grid
- Different scores for different action round
- Very complex basic patterns

Verification phase

- Gradient?
- $N \times N$ grid?
- Not enough chances to detect
- No prior knowledge about patterns and rewards

Overview

1 Introduction

2 Environment Analysis
- Feedback phase
- Check phase
- Verification phase

3 Solution

4 About Guazi

Main process

- Two-stage Random Search[1]
- 21 episodes in total
  - 6 episodes for random choices
    - Pick up the best reward
  - 15 episodes for random adjustments
    - 3 rounds for 5 actions in order
    - If get a higher score, accept the adjustment

[1] https://en.wikipedia.org/wiki/Random_search

Tricks

- Similar actions may be less informative
  - Restrict the number choices in

$$\{0.1, 0.3, 0.5, 0.7, 0.9\}$$

  - Do not choose the same action to explore

Code

- The code is available
  - `https://github.com/luosuiqian/submission`

Overview

1. Introduction
2. Environment Analysis
   - Feedback phase
   - Check phase
   - Verification phase
3. Solution
4. About Guazi