Computational Cognitive Science

Lecture 14: Active learning 2

Chris Lucas

School of Informatics

University of Edinburgh

1 November, 2024

Active learning

Last time we focused on active learning for gaining information. This is useful if

- We might have many goals
- We don't know what our goals will be

Exploration and exploitation

Learning (“epistemic”) goals can conflict with near-term rewards or goals.

Minimizing loss vs learning

Informative actions can come at a cost:

- “What happens if I just don’t sleep?”
- “How fast can this car go?”
- “Doctors *say* you should take antibiotics if you get the black plague, but who has tested that claim *lately*?”

Exploitation and exploration

One could argue that learning is a means to an end, rather than its own goal.

This leads to a trade-off:

- **Exploitation** maximizes (expected) reward in the near term
- **Exploration** aka learning how to exploit more rewarding choices in the future

A rational agent balances these to maximize long-term reward.

Exploitation and exploration

Imagine you're on holiday where there are 4 food trucks. You're going to have 20 meals before you leave.

Suppose they cost the same, you don't care about novelty, and you have an open-ended integer rating scale.

Exploitation and exploration

They serve:

- Soylent A
- Soylent B
- Soylent C
- Soylent D

Let's try to maximize our average rating over all 20 tries.

Exploitation and exploration

(Let's try it)

Exploitation and exploration

Truck	rewards															
A	21	12	25	12	14	6	15	18	17	15	15	14	14	12	16	...
B	24	12	20	20	21	20	12	36	31	24	11	14	3	17	24	...
C	17	19	47	11	26	16	19	19	18	15	23	35	15	17	15	...
D	20	4	15	22	8	9	23	20	2	7	9	22	16	5	15	...

- A: Mean reward of 15.6; distribution is $N(15, 5)$
- B: Mean reward of 16.7; distribution is $N(15, 9)$
- C: Mean reward of 19.75; distribution is $10 + exp(10)$
- D: Mean reward of 11.85; distribution is $25 * beta(.8, .8)$

Multi-armed bandits

This is a “multi-armed bandit” task.

- Bandit tasks are simple, but get at important questions
- A minimal reinforcement learning problem
- “Bernoulli bandits” are common: choose a button, 1/0 reward

Multi-armed bandits

How can we understand and model human behavior in bandit tasks?

We will discuss a few approaches.

Approach 1: Optimal policies

For some bandit tasks it is possible to determine what optimal behavior should be, given assumptions about the reward distributions.

How do we maximize reward? A *policy* maps maps states/observations to actions. We want a policy that maximizes total expected reward

This is the policy that maximizes the sum of

1. The expected reward of the next action and
2. the cumulative expected reward of all subsequent actions given an optimal policy going forward

Approach 1: Optimal policies

How can we find an optimal policy?

If we are at our last choice, then future rewards are zero. If we can use our observations to compute expected rewards (e.g., with a conjugate prior), we're set.

If we have computed expected future rewards for all future actions and observations, and can compute probabilities of different action outcomes, then we are set in general.

Approach 1: Optimal policies

1. For all possible final choices and outcomes, compute the max expected reward given the data
2. Next, step back by one action. We know the probabilities of our state transitions and all expected rewards.
3. Step back again, repeating Step 2 until we get to the first action.

Approach 1: Optimal policies

Pros:

- A rational/computational-level model
- Conceptually simple; clear predictions
- Generally applicable – doesn't assume any particular reward structure

Approach 1: Optimal policies

Cons:

- Demands simulating/tracking enormous spaces of possibilities
- Hard to compute exactly for interesting problems

Approach 2: Success ratio/greedy

- Pick the option with the best ratio of wins to losses
- Or the best expected reward; can use a beta prior for Bernoulli bandits

Approach 2: Success ratio/greedy

Pros:

- Easy: Low time and memory requirements

Cons:

- For SR, need to define wins and losses
- May never recover from bad early evidence

Approach 3: ϵ-greedy

Flip a coin with an ϵ chance of coming up heads

- If heads, pick *completely at random*.
- If tails, pick an option with highest expected reward

Approach 3: ϵ-greedy

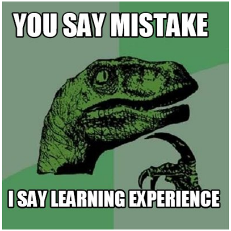
Pros:

- Simple. Low-cost, if you can compute expected reward cheaply
- Surprisingly useful if combined with a good greedy policy

Approach 3: ϵ-greedy

Cons:

- Complete randomness is bad idea if you care about extreme losses, e.g., dying
- There are subtler/safer ways to inject exploration
- Suboptimal unless we take ϵ to zero



Approach 4: Win-stay, lose-switch

- If I win, do the same thing again
- If I lose, do something different

Pros:

- Low time and memory requirements
  - In minimal version, no memory required at all
- Not *too* horrible if the switch policy is sensible (e.g., not completely random)
- Some people do seem to do this, sometimes

Approach 4: Win-stay, lose-switch

Cons:

- “Winning” is not always clear-cut
- Classic version, w/random switch, is unrewarding
    - Extreme, avoidable losses

"Win-stay, lose-switch: For when you just don’t care!"

Approach 5: Thompson sampling

Use your experience to maintain a distribution of possible rewards for each action.

1. Sample from these distributions
2. Pick the best

Approach 5: Thompson sampling

Pros:

- Avoids extreme losses where random choices don't
- Doesn't get "stuck" like success ratio or greedy

Cons:

- More effort than the simplest policies
- Does not avoid extreme negative utilities (e.g., death)

Related: "Upper-confidence bound" (UCB)

Which of these explains human behavior?

Steyvers et al., 2009 link compared success-ratio, win-stay lose-switch (WSLS), optimal (with noise), and random-guessing models using Bayesian methods.

Of these, WSLS fit the largest number of participants, but not a majority.

Other approaches and ideas

- Strategy-switching, e.g., switching from exploration-mode to exploitation-mode
- People may implicitly believe rewards could change
- People may assume that rewards are not independent

Questions

- Would (substantial) rewards change participants' policies?
- Are we sure these models are rich enough to capture human behavior?
- Do people have a repertoire of policies they can choose between?
  - E.g., switching between exploration and exploitation?

References

Steyvers, M., Lee, M. D., & Wagenmakers, E. J. (2009). A Bayesian analysis of human decision-making on bandit problems. Journal of Mathematical Psychology, 53(3), 168-179.