Optimal Policies

Frederik Mallmann-Trenn
6CCS3AIN

How utilities are calculated

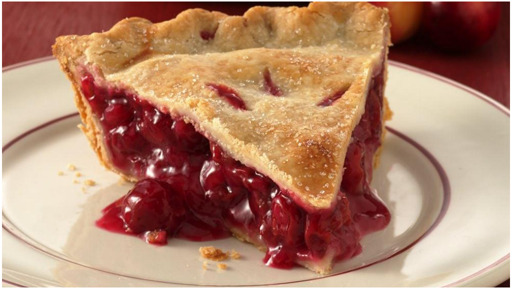
- So far we have assumed that utilities are summed along a run.
  - Not the only way.
- In general we need to compute $U_r([s_0, s_1, \ldots, s_n])$ for general $U_r(\cdot)$. That is, the utility of a run.
- Before $U_r(\cdot)$ was just the sum of rewards in every state.
- Can consider finite and infinite horizons.
  - Is it "game over" at some point?
- Turns out that infinite horizons are mostly easier to deal with.
  - That is what we will use.

How utilities are calculated

- Also have to consider whether **utilities** are stationary or non-stationary.
  - Think of: does the same state always have the same value?
  - E.g., in Pacman when you pick up a fruit, there is a large reward for that tile. That changes after you picked up the fruit.
- Example:
  - Normally we prefer one state to another.
  - Passing the AI module to failing it
  - In this case when the exam is, today or next week, is irrelevant.
- We assume utilities are stationary.

But are they?

- Not clear that utilities are always stationary.



- In truth, I don't always most want to eat cherry pie.
- Despite this, we will assume that utilities are stationary.

How utilities are calculated

- With stationary utilities, there are two ways to establish $U_r([s_0, s_1, \ldots, s_n])$ from $R(s)$.
- Additive rewards:

$$U_r([s_0, s_1, \ldots, s_n]) = R(s_0) + R(s_1) + \ldots + R(s_n)$$

  as above.
- Discounted rewards:

$$U_r([s_0, s_1, \ldots, s_n]) = R(s_0) + \gamma R(s_1) + \ldots + \gamma^n R(s_n)$$

  where the discount factor γ is a number between 0 and 1.
- The discount factor models the preference of the agent for current over future rewards.

How utilities are calculated

- There is an issue with infinite sequences with additive, undiscounted rewards.
  - What will the utility of a policy be?

How utilities are calculated

- There is an issue with infinite sequences with additive, undiscounted rewards.
  - What will the utility of a policy be?
- Unbounded
- ∞ or $-\infty$.
- This is problematic if we want to compare policies.

How utilities are calculated

- Some solutions are (definitions follow):
  - Proper policies
  - Average reward
  - Discounted rewards

How utilities are calculated

- Proper policies always end up in a terminal state eventually.
- Thus they have a finite expected utility.

How utilities are calculated

- We can compute the average reward per time step.
- Even for an infinite policy this will (usually) be finite.

How utilities are calculated

- Assume: $0 \leqslant \gamma < 1$ and rewards are bounded by R_{max}
- With discounted rewards the utility of an infinite sequence is finite:

$$
\begin{aligned}
U_r([s_0, s_1, \ldots, s_n]) &= \sum_{t=0}^{n} \gamma^t R(s_t) \\
&\leqslant \sum_{t=0}^{\infty} \gamma^t R(s_t) \\
&\leqslant \sum_{t=0}^{\infty} \gamma^t R_{max} \\
&\leqslant \frac{R_{max}}{(1-\gamma)}
\end{aligned}
$$

Optimal policies

- With discounted rewards we compare policies by computing their expected values.
- The expected utility of executing π starting in s is given by:

$$U^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R(S_t)\right]$$

where S_t is the state the agent gets to at time t.
- S_t is a random variable and we compute the probability of all its values by looking at all the runs which end up there after t steps.

Optimal policies

- The optimal policy is then:

$$\pi^* = \arg\max_{\pi} U^{\pi}(s)$$

- It turns out that this is independent of the state the agent starts in.

Optimal policies

3	0.812	0.868	0.918	+ 1
2	0.762		0.660	–1
1	0.705	0.655	0.611	0.388
	1	2	3	4

- Here we have the values of states if the agent executes an optimal policy

  $U^{\pi^*}(s)$

Optimal policies

3	0.812	0.868	0.918	$\boxed{+1}$
2	0.762		0.660	$\boxed{-1}$
1	0.705	0.655	0.611	0.388
	1	2	3	4

- Here we have the values of states if the agent executes an optimal policy

  $$U^{\pi^*}(s)$$

- What should the agent do if it is in (3, 1)?

Example

- The answer is *Left*.
- The best action is the one that maximises the expected utility.
- (You have to calculate the expected utility of all the actions to see why Left is the best choice.)