

These lecture notes are heavily based on notes originally written by Nikhil Sharma.

## Utilities

效用函数是一种数学函数，它将状态或决策映射到相关的效用值，以便智能代理能够根据这些值来选择最佳的行动。

Throughout our discussion of rational agents, the concept of utility came up repeatedly. In games, for example, Utility values are generally hard-wired into the game, and agents use these utility values to select an action. We'll now discuss what's necessary in order to generate a viable utility function.

Rational agents must follow the **principle of maximum utility** - they must always select the action that maximizes their expected utility. However, obeying this principle only benefits agents that have **rational preferences**. To construct an example of **irrational preferences**, say there exist 3 objects,  $A$ ,  $B$ , and  $C$ , and our agent is currently in possession of  $A$ . Say our agent has the following set of irrational preferences:

- Our agent prefers  $B$  to  $A$  plus \$1
- Our agent prefers  $C$  to  $B$  plus \$1
- Our agent prefers  $A$  to  $C$  plus \$1

A malicious agent in possession of  $B$  and  $C$  can trade our agent  $B$  for  $A$  plus a dollar, then  $C$  for  $B$  plus a dollar, then  $A$  again for  $C$  plus a dollar. Our agent has just lost \$3 for nothing! In this way, our agent can be forced to give up all of its money in an endless and nightmarish cycle.

Let's now properly define the mathematical language of preferences:

- If an agent prefers receiving a prize  $A$  to receiving a prize  $B$ , this is written  $A \succ B$
- If an agent is indifferent between receiving  $A$  or  $B$ , this is written as  $A \sim B$
- A **lottery** is a situation with different prizes resulting with different probabilities. To denote lottery where  $A$  is received with probability  $p$  and  $B$  is received with probability  $(1 - p)$ , we write

$$L = [p, A; (1 - p), B]$$

In order for a set of preferences to be rational, they must follow the five **Axioms of Rationality**:

- **Orderability**:  $(A \succ B) \vee (B \succ A) \vee (A \sim B)$   
A rational agent must either prefer one of  $A$  or  $B$ , or be indifferent between the two.
- **Transitivity**:  $(A \succ B) \wedge (B \succ C) \Rightarrow (A \succ C)$   
If a rational agent prefers  $A$  to  $B$  and  $B$  to  $C$ , then it prefers  $A$  to  $C$ .
- **Continuity**:  $A \succ B \succ C \Rightarrow \exists p [p, A; (1 - p), C] \sim B$   
If a rational agent prefers  $A$  to  $B$  but  $B$  to  $C$ , then it's possible to construct a lottery  $L$  between  $A$  and  $C$  such that the agent is indifferent between  $L$  and  $B$  with appropriate selection of  $p$ .

- **Substitutability:**  $A \sim B \Rightarrow [p, A; (1-p), C] \sim [p, B; (1-p), C]$   
A rational agent indifferent between two prizes  $A$  and  $B$  is also indifferent between any two lotteries which only differ in substitutions of  $A$  for  $B$  or  $B$  for  $A$ .
- **Monotonicity:**  $A \succ B \Rightarrow (p \geq q \Leftrightarrow [p, A; (1-p), B] \succeq [q, A; (1-q), B])$   
If a rational agent prefers  $A$  over  $B$ , then given a choice between lotteries involving only  $A$  and  $B$ , the agent prefers the lottery assigning the highest probability to  $A$ .

If all five axioms are satisfied by an agent, then it's guaranteed that the agent's behavior is describable as a **maximization of expected utility**. More specifically, this implies that there exists a real-valued **utility function**  $U$  that when implemented will assign greater utilities to preferred prizes, and also that the utility of a lottery is the **expected value of the utility of the prize** resulting from the lottery. These two statements can be summarized in two concise mathematical equivalences:

$$U(A) \geq U(B) \Leftrightarrow A \succeq B \quad (1)$$

$$U([p_1, S_1; \dots; p_n, S_n]) = \sum_i p_i U(S_i) \quad (2)$$

If these constraints are met and an appropriate choice of algorithm is made, the agent implementing such a utility function is **guaranteed to behave optimally**. Let's discuss utility functions in greater detail with a concrete example. Consider the following lottery:

$$L = [0.5, \$0; 0.5, \$1000]$$

This represents a lottery where you receive \$1000 with probability 0.5 and \$0 with probability 0.5. Now consider three agents  $A_1, A_2$ , and  $A_3$  which have utility functions  $U_1(\$x) = x$ ,  $U_2(\$x) = \sqrt{x}$ , and  $U_3(\$x) = x^2$  respectively. If each of the three agents were faced with a choice between participating in the lottery and receiving a flat payment of \$500, which would they choose? The respective utilities for each agent of participating in the lottery and accepting the flat payment are listed in the following table:

Agent	Lottery	Flat Payment
1	500	500
2	15.81	22.36
3	500000	250000

These utility values for the lotteries were calculated as follows, making use of equation (2) above:

$$U_1(L) = U_1([0.5, \$0; 0.5, \$1000]) = 0.5 \cdot U_1(\$1000) + 0.5 \cdot U_1(\$0) = 0.5 \cdot 1000 + 0.5 \cdot 0 = \boxed{500}$$

$$U_2(L) = U_2([0.5, \$0; 0.5, \$1000]) = 0.5 \cdot U_2(\$1000) + 0.5 \cdot U_2(\$0) = 0.5 \cdot \sqrt{1000} + 0.5 \cdot \sqrt{0} = \boxed{15.81}$$

$$U_3(L) = U_3([0.5, \$0; 0.5, \$1000]) = 0.5 \cdot U_3(\$1000) + 0.5 \cdot U_3(\$0) = 0.5 \cdot 1000^2 + 0.5 \cdot 0^2 = \boxed{500000}$$

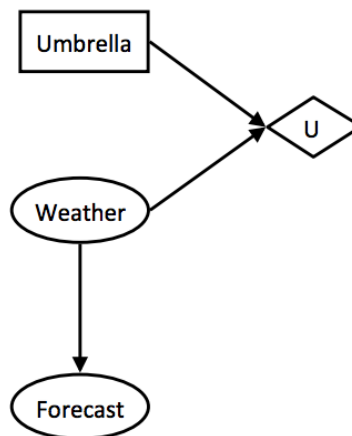
With these results, we can see that agent  $A_1$  is indifferent between participating in the lottery and receiving the flat payment (the utilities for both cases are identical). Such an agent is known as **risk-neutral**. Similarly, agent  $A_2$  prefers the flat payment to the lottery and is known as **risk-averse** and agent  $A_3$  prefers the lottery to the flat payment and is known as **risk-seeking**.

# Decision Networks

In the third note, we learned about game trees and algorithms such as minimax and expectimax which we used to determine optimal actions that maximized our expected utility. Then in the fifth note, we discussed Bayes' nets and how we can use evidence we know to run probabilistic inference to make predictions. Now we'll discuss <sup>建模</sup> a combination of both Bayes' nets and expectimax known as a **decision network** that we can use to **model** the effect of various actions on utilities based on an overarching graphical probabilistic model. Let's dive right in with the anatomy of a decision network:

- **Chance nodes** - Chance nodes in a decision network behave identically to Bayes' nets. Each outcome in a chance node has an associated probability, which can be determined by running inference on the underlying Bayes' net it belongs to. We'll represent these with ovals.
- **Action nodes** - Action nodes are nodes that we have complete control over; they're nodes representing a choice between any of a number of actions which we have the power to choose from. We'll represent action nodes with rectangles.
- **Utility nodes** - Utility nodes are children of some combination of action and chance nodes. They output a utility based on the values taken on by their parents, and are represented as diamonds in our decision networks.

Consider a situation when you're deciding whether or not to take an umbrella when you're leaving for class in the morning, and you know there's a forecasted 30% chance of rain. Should you take the umbrella? If there was a 80% chance of rain, would your answer change? This situation is ideal for modeling with a decision network, and we do it as follows:



As we've done throughout this course with the various modeling techniques and algorithms we've discussed, our goal with decision networks is again to **select the action which yields the maximum expected utility (MEU)**. This can be done with a fairly straightforward and intuitive procedure:

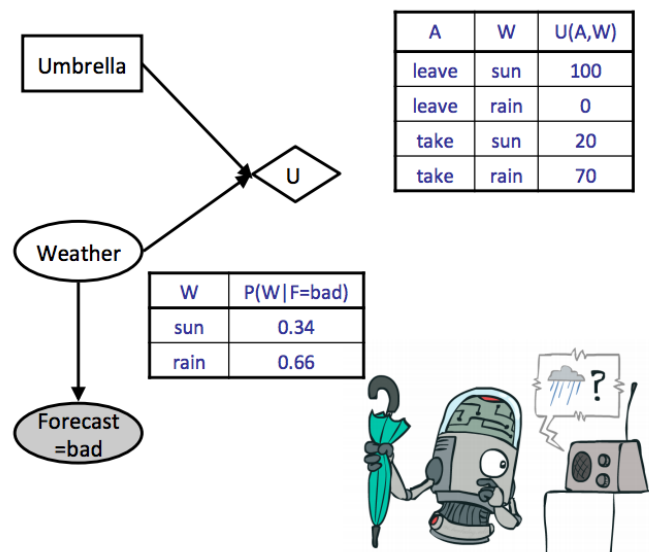
- Start by instantiating all evidence that's known, and run inference to calculate the **posterior probabilities** of all chance node parents of the utility node into which the action node feeds. 后验概率
- Go through each possible action and compute the expected utility of taking that action given the posterior probabilities computed in the previous step. The expected utility of taking an action  $a$  given evidence  $e$  and  $n$  chance nodes is computed with the following formula:

$$EU(a|e) = \sum_{x_1, \dots, x_n} P(x_1, \dots, x_n|e) U(a, x_1, \dots, x_n)$$

where each  $x_i$  represents a value that the  $i^{th}$  chance node can take on. We simply take a weighted sum over the utilities of each outcome under our given action with weights corresponding to the probabilities of each outcome.

- Finally, select the action which yielded the highest utility to get the MEU.

Let's see how this actually looks by calculating the optimal action (should we *leave* or *take* our umbrella) for our weather example, using both the conditional probability table for weather given a bad weather forecast (forecast is our evidence variable) and the utility table given our action and the weather:



Note that we **have omitted the inference computation for the posterior probabilities  $P(W|F = \text{bad})$** , but we could compute these using **any of the inference algorithms we discussed for Bayes Nets**. Instead, here we simply assume the above table of posterior probabilities for  $P(W|F = \text{bad})$  as given. Going through both our actions and computing expected utilities yields:

$$\begin{aligned}
 EU(\text{leave}|\text{bad}) &= \sum_w P(w|\text{bad})U(\text{leave}, w) \\
 &= 0.34 \cdot 100 + 0.66 \cdot 0 = \boxed{34}
 \end{aligned}$$

$$\begin{aligned}
 EU(\text{take}|\text{bad}) &= \sum_w P(w|\text{bad})U(\text{take}, w) \\
 &= 0.34 \cdot 20 + 0.66 \cdot 70 = \boxed{53}
 \end{aligned}$$

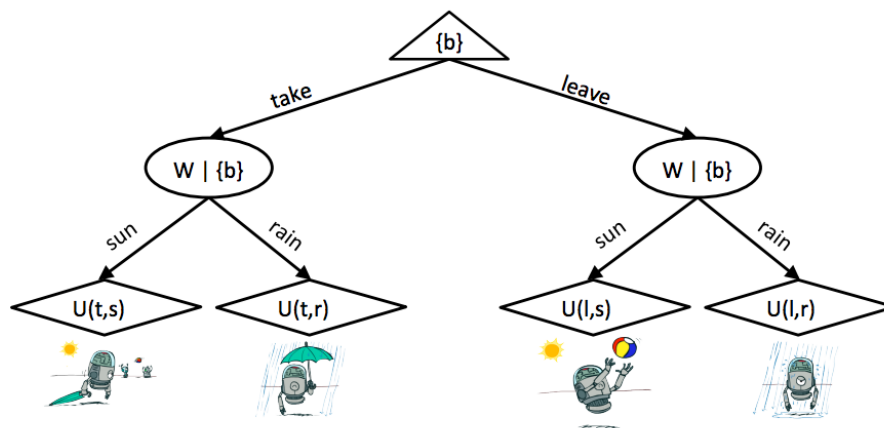
All that's left to do is take the maximum over these computed utilities to determine the MEU:

$$MEU(F = \text{bad}) = \max_a EU(a|\text{bad}) = \boxed{53}$$

The action that yields the maximum expected utility is *take*, and so this is the action recommended to us by the decision network. More formally, the action that yields the MEU can be determined by taking the **argmax** over expected utilities.

## Outcome Trees

We mentioned at the start of this note that decision networks involved some expectimax-esque elements, so let's discuss what exactly that means. We can unravel the selection of an action corresponding to the one that maximizes expected utility in a decision network as an **outcome tree**. Our weather forecast example from above unravels into the following outcome tree:



The root node at the top is a maximizer node, just like in expectimax, and **is controlled by us**. We select an action, which takes us to the **next level** in the tree, **controlled by chance nodes**. At this level, chance nodes resolve to different utility nodes at the final level with probabilities corresponding to the posterior probabilities derived from probabilistic inference run on the underlying Bayes' net. What exactly makes this different from vanilla expectimax? The only real difference is that for outcome trees **we annotate our nodes with what we know at any given moment** (inside the curly braces).

# The Value of Perfect Information

In everything we've covered up to this point, we've generally always assumed that our agent has all the information it needs for a particular problem and/or has no way to acquire new information. In practice, this is hardly the case, and one of the most important parts of decision making is knowing **whether or not it's worth gathering more evidence to help decide** which action to take. Observing new evidence almost always has some cost, whether it be in terms of time, money, or some other medium. In this section, we'll talk about a very important concept - **the value of perfect information (VPI) - which mathematically quantifies the amount an agent's maximum expected utility is expected to increase if it observes some new evidence**. We can **compare the VPI** of learning some new information **with the cost** associated with observing that information to make decisions about whether or not it's worthwhile to observe.

这里比较的是 Utility 增量的预期和观测的成本

## General Formula

Rather than simply presenting the formula for computing the value of perfect information for new evidence, let's walk through an intuitive derivation. We know from our above definition that the value of perfect information is the amount our maximum expected utility is expected to increase if we decide to observe new evidence. We know our current maximum utility given our current evidence  $e$ :

$$MEU(e) = \max_a \sum_s P(s|e)U(s, a)$$

Additionally, we know that if we observed some new evidence  $e'$  before acting, the maximum expected utility of our action at that point would become

$$MEU(e, e') = \max_a \sum_s P(s|e, e')U(s, a)$$

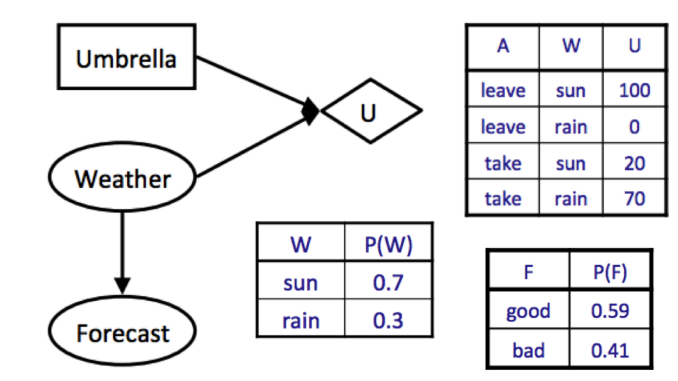
However, note that *we don't know what new evidence we'll get*. For example, if we didn't know the weather forecast beforehand and chose to observe it, the forecast we observe might be either *good* or *bad*. Because we don't know what what new evidence  $e'$  we'll get, we must **represent it as a random variable  $E'$** . How do we represent the new MEU we'll get if we choose to observe a new variable if we don't know what the evidence gained from observation will tell us? The answer is to compute the **expected value of the maximum expected utility** which, while being a mouthful, is the natural way to go:

$$MEU(e, E') = \sum_{e'} P(e'|e)MEU(e, e')$$

Observing a new evidence variable yields a different MEU with probabilities corresponding to the probabilities of observing each value for the evidence variable, and so by computing  $MEU(e, E')$  as above, we compute what we expect our new MEU will be if we choose to observe new evidence. We're just about done now - returning to our definition for VPI, we want to find the amount our MEU is expected to increase if we choose to observe new evidence. We know our current MEU and the expected value of the new MEU if we choose to observe, so the expected MEU increase is simply the difference of these two terms! Indeed,

$$VPI(E'|e) = MEU(e, E') - MEU(e)$$

where we can read  $VPI(E'|e)$  as "the value of observing new evidence  $E'$  given our current evidence  $e$ ". Let's work our way through an example by revisiting our weather scenario one last time:



If we don't observe any evidence, then our maximum expected utility can be computed as follows:

$$\begin{aligned}
 MEU(\emptyset) &= \max_a EU(a) \\
 &= \max_a \sum_w P(w)U(a, w) \\
 &= \max\{0.7 \cdot 100 + 0.3 \cdot 0, 0.7 \cdot 20 + 0.3 \cdot 70\} \\
 &= \max\{70, 35\} \\
 &= 70
 \end{aligned}$$

Note that the convention when we have no evidence is to write  $MEU(\emptyset)$ , denoting that our evidence is the empty set. Now let's say that we're deciding whether or not to observe the weather forecast. We've already computed that  $MEU(F = \text{bad}) = 53$ , and let's assume that running an identical computation for  $F = \text{good}$  yields  $MEU(F = \text{good}) = 95$ . We're now ready to compute  $MEU(e, E')$ :

$$\begin{aligned}
 MEU(e, E') &= MEU(F) \\
 &= \sum_{e'} P(e'|e)MEU(e, e') \\
 &= \sum_f P(F = f)MEU(F = f) \\
 &= P(F = \text{good})MEU(F = \text{good}) + P(F = \text{bad})MEU(F = \text{bad}) \\
 &= 0.59 \cdot 95 + 0.41 \cdot 53 \\
 &= 77.78
 \end{aligned}$$

Hence we conclude  $VPI(F) = MEU(F) - MEU(\emptyset) = 77.78 - 70 = \boxed{7.78}$ .

# Properties of VPI

The value of perfect information has several very important properties, namely:

- **Nonnegativity.**  $\forall E', e \text{ } VPI(E'|e) \geq 0$

Observing new information always allows you to **make a more informed decision, and so your maximum expected utility can only increase** (or stay the same if the information is irrelevant for the decision you must make).

- **Nonadditivity.**  $VPI(E_j, E_k|e) \neq VPI(E_j|e) + VPI(E_k|e)$  in general.

This is probably the trickiest of the three properties to understand intuitively. It's true because generally observing some new evidence  $E_j$  might change how much we care about  $E_k$ ; therefore we can't simply add the VPI of observing  $E_j$  to the VPI of observing  $E_k$  to get the VPI of observing both of them. Rather, the VPI of observing two new evidence variables is equivalent to observing one, incorporating it into our current evidence, then observing the other. This is encapsulated by the order-independence property of VPI, described more below.

- **Order-independence.**  $VPI(E_j, E_k|e) = VPI(E_j|e) + VPI(E_k|e, E_j) = VPI(E_k|e) + VPI(E_j|e, E_k)$

Observing multiple new evidences yields the same gain in maximum expected utility regardless of the order of observation. This should be a fairly straightforward assumption - because we don't actually take any action until after observing any new evidence variables, it doesn't actually matter whether we observe the new evidence variables together or in some arbitrary sequential order.

这个式子表示的是预计增量VPI的之间的关系，从直观上感觉是对的  
但不知如何证明