Instructors: Stuart Russell & Dawn Song. Lecturer: Regina Wang

[Adapted from slides created by Dan Klein, Pieter Abbeel, and Anca. http://ai.berkeley.edu.]
Rational Preferences

The Axioms of Rationality

?
Rational Preferences

The Axioms of Rationality

Orderability:
\((A > B) \lor (B > A) \lor (A \sim B)\)

Transitivity:
\((A > B) \land (B > C) \Rightarrow (A > C)\)

Continuity:
\((A > B > C) \Rightarrow \exists \rho [\rho, A; 1-p, C] \sim B\)

Substitutability:
\((A \sim B) \Rightarrow [\rho, A; 1-p, C] \sim [\rho, B; 1-p, C]\)

Monotonicity:
\((A > B) \Rightarrow (p \geq q) \iff [\rho, A; 1-p, B] \geq [q, A; 1-q, B]\)

What’s the implication?
Theorem [Ramsey, 1931; von Neumann & Morgenstern, 1944]

Given any preferences satisfying the previous constraints, there exists a real-valued function $U$ such that:

$$\sum_i p_i U(s_i)$$
Theorem [Ramsey, 1931; von Neumann & Morgenstern, 1944]
- Given any preferences satisfying the previous constraints, there exists a real-valued function $U$ such that:

$$U(A) > U(B) \iff A \succ B \quad \text{and} \quad U(A) \sim U(B) \iff A \sim B$$

$$U([p_1,S_1; \ldots; p_n,S_n]) = p_1U(S_1) + \ldots + p_nU(S_n)$$

- Maximum expected utility (MEU) principle:
  - A rational agent chooses the action that maximizes expected utility
Decision Networks
Decision Networks

- Decision network = Bayes net + Actions + Utilities
  - **Chance nodes** (just like BNs)
  - **Action nodes** (rectangles, cannot have parents, will have value fixed by algorithm)
  - **Utility nodes** (diamond, depends on action and chance nodes)

- Decision algorithm:
  - Fix evidence $e$
  - For each possible action $a$
    - Fix action node to $a$
    - Compute posterior $P(W|e)$ for parents $W$ of $U$
    - Compute expected utility $\sum_w P(w|e) U(a, w)$
  - Return the action with highest expected utility
Maximum Expected Utility

Umbrella = leave

\[ \text{EU}(\text{leave}) = \sum_w P(w)U(\text{leave, } w) \]
\[ = 0.7 \cdot 100 + 0.3 \cdot 0 = 70 \]

Umbrella = take

\[ \text{EU}(\text{take}) = \sum_w P(w)U(\text{take, } w) \]
\[ = 0.7 \cdot 20 + 0.3 \cdot 70 = 35 \]

Optimal decision = leave

\[ \text{MEU}(\phi) = \max_a \text{EU}(a) = 70 \]
Decisions as Outcome Trees

- Almost exactly like expectimax!
- What’s changed?
Example: Take an umbrella?

Umbrella

<table>
<thead>
<tr>
<th>A</th>
<th>W</th>
<th>U(A,W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>leave</td>
<td>sun</td>
<td>100</td>
</tr>
<tr>
<td>leave</td>
<td>rain</td>
<td>0</td>
</tr>
<tr>
<td>take</td>
<td>sun</td>
<td>20</td>
</tr>
<tr>
<td>take</td>
<td>rain</td>
<td>70</td>
</tr>
</tbody>
</table>

Weather

<table>
<thead>
<tr>
<th>W</th>
<th>P(W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Forecast = bad

| W       | P(F=bad|W) |
|---------|--------|
| sun     | 0.17   |
| rain    | 0.77   |
Example: Take an umbrella?

- Decision algorithm:
  - Fix evidence $e$
  - For each possible action $a$
    - Fix action node to $a$
    - Compute posterior $P(W|e)$ for parents $W$ of $U$
    - Compute expected utility of action $a$: $\sum_w P(w|e) U(a,w)$
  - Return the action with highest expected utility

- Umbrella = leave

EU(leave|F=bad) = $\sum_w P(w|F=bad) U(leave,w)$

We have: $P(W) P(F|W) P(W|F) = \frac{P(W,F)}{\sum_w P(w,F)} = \frac{P(F|W)P(W)}{\sum_w P(F|w)P(w)}$
Example: Take an umbrella?

- Decision algorithm:
  - Fix evidence $e$
  - For each possible action $a$
    - Fix action node to $a$
    - Compute posterior $P(W|e)$ for parents $W$ of $U$
    - Compute expected utility of action $a$: $\sum_w P(w|e) U(a,w)$
  - Return the action with highest expected utility

Umbrella = leave

\[
EU(\text{leave}|F=\text{bad}) = \sum_w P(w|F=\text{bad}) U(\text{leave},w)
\]

\[
= 0.34 \times 100 + 0.66 \times 0 = 34
\]

Umbrella = take

\[
EU(\text{take}|F=\text{bad}) = \sum_w P(w|F=\text{bad}) U(\text{take},w)
\]

\[
= 0.34 \times 20 + 0.66 \times 70 = 53
\]

Optimal decision = take!

\[
\text{MEU}(F=\text{bad}) = \max_a EU(a|F=\text{bad}) = 53
\]
Decisions as Outcome Trees

Weather Forecast = bad

Umbrella

W | \{b\}

take

sun
U(t,s)

U(t,r)
rain

W | \{b\}

leave

sun
U(l,s)
rain
U(l,r)
Value of (Perfect) Information
How do you tell if you want to take a specific class next semester?
Value of Perfect Information

- **Idea:** compute value of acquiring evidence
  - Can be done directly from decision network

- **Example:** buying oil drilling rights
  - Two blocks A and B, exactly one has oil, worth k
  - You can drill in one location
  - Prior probabilities 0.5 each, & mutually exclusive
  - Drilling in either A or B has EU = k/2, MEU = k/2

- **Question:** what’s the value of information of O?
  - Value of knowing which of A or B has oil
  - Value is expected gain in MEU from new info
  - Survey may say “oil in a” or “oil in b,” prob 0.5 each
  - If we know OilLoc, MEU is k (either way)
  - Gain in MEU from knowing OilLoc?
  - \( VPI(OilLoc) = k - k/2 = k/2 \)
  - Fair price of information: k/2
Before you see the forecast (no evidence)
- MEU(∅) = maxₐEU(a) = 70

What if you look at the forecast?
- If Forecast=bad
  - MEU(F=bad) = maxₐEU(a | F=bad) = 53
- If Forecast=good
  - MEU(F=good) = maxₐEU(a | F=good) = 89

But, we don’t know what the forecast will be ahead of time!

So we need a distribution of P(F)
- Expected utility given forecast
  - = 0.35 x 53 + 0.65 x 89 = 76.4
- Value of information = 76.4 - 70 = 6.4
Value of Information

- Assume we have evidence $E=\text{e}$. Value if we act now:
  \[ MEU(e) = \max_a \sum_s P(s|\text{e}) \ U(s, a) \]

- Assume we see that $E' = \text{e}'$. Value if we act then:
  \[ MEU(\text{e}, \text{e}') = \max_a \sum_s P(s|\text{e}, \text{e}') \ U(s, a) \]

But $E'$ is a random variable whose value is unknown, so we don't know what $\text{e}'$ will be.

- Expected value if $E'$ is revealed and then we act:
  \[ MEU(\text{e}, E') = \sum_{\text{e}'} P(\text{e}'|\text{e}) \ MEU(\text{e}, \text{e}') \]

- Value of information: how much MEU goes up by revealing $E'$ first then acting, over acting now:
  \[ VPI(\text{E}'|\text{e}) = MEU(\text{e}, E') - MEU(\text{e}) \]
VPI Properties

VPI is non-negative! $\text{VPI}(E_i \mid e) \geq 0$

VPI is not (usually) additive: $\text{VPI}(E_i, E_j \mid e) \neq \text{VPI}(E_i \mid e) + \text{VPI}(E_j \mid e)$

VPI is order-independent: $\text{VPI}(E_i, E_j \mid e) = \text{VPI}(E_j, E_i \mid e)$
Quick VPI Questions

- The soup of the day is either clam chowder or split pea, but you wouldn’t order either one. What’s the value of knowing which it is?

- There are two kinds of plastic forks at a picnic. One kind is slightly sturdier. What’s the value of knowing which?

- You’re playing the lottery. The prize will be $0 or $100. You can play any number between 1 and 100 (chance of winning is 1%). What is the value of knowing the winning number?
Value of Imperfect Information?

- No such thing
- Information corresponds to the observation of a node in the decision network
- If data is “noisy” that just means we don’t observe the original variable, but another variable which is a noisy version of the original one
Decisions with unknown preferences

- In reality the assumption that we can write down our exact preferences for the machine to optimize is false
- A machine optimizing the wrong preferences causes problems
Decisions with unknown preferences

- In reality the assumption that we can write down our exact preferences for the machine to optimize is false.
- A machine optimizing the wrong preferences causes problems.
- A machine that is explicitly uncertain about the human’s preferences will defer to the human (e.g., allow itself to be switched off).
Off-switch problem (example)

EU(\text{act}) = +10

EU(\text{wait}) = (0.4 \times 0) + (0.6 \times 30) = +18
Off-switch problem (general proof)

- $EU(\text{act}) = \int_{-\infty}^{+\infty} P(u) \cdot u \, du = \int_{-\infty}^{0} P(u) \cdot u \, du + \int_{0}^{+\infty} P(u) \cdot u \, du$
- $EU(\text{wait}) = \int_{-\infty}^{0} P(u) \cdot 0 \, du + \int_{0}^{+\infty} P(u) \cdot u \, du$
- Obviously $\int_{-\infty}^{0} P(u) \cdot u \, du \leq \int_{-\infty}^{0} P(u) \cdot 0 \, du$
- Hence $EU(\text{act}) \leq EU(\text{wait})$

- “If H doesn’t switch me off, then the action must be good for H, and I’ll get to do it, so that’s good; if H does switch me off, then it’s because the action must be bad for H, so it’s good that I won’t be allowed to do it.”
Sequential decisions under uncertainty

So far, decision problem is one-shot --- concerning only one action

Sequential decision problem: agent’s utility depends on a sequence of actions
Markov Decision Process (MDP)

- Environment history: \([s_0, a_0, s_1, a_1, \ldots, s_t]\)
- “Markov” generally means that given the present state, the future and the past are independent
- For Markov decision processes, “Markov” means action outcomes depend only on the current state

\[
P(S_{t+1} = s' | S_t = s_t, A_t = a_t, S_{t-1} = s_{t-1}, A_{t-1}, \ldots, S_0 = s_0) = P(S_{t+1} = s' | S_t = s_t, A_t = a_t)
\]

- This is just like search, where the successor function could only depend on the current state (not the history)

Andrey Markov (1856-1922)
Markov Decision Process (MDP)

- An MDP is defined by:
  - A set of states \( s \in S \)
  - A set of actions \( a \in A \)
  - A transition model \( T(s, a, s') \)
    - Probability that \( a \) from \( s \) leads to \( s' \), i.e., \( P(s' | s, a) \)
  - A reward function \( R(s, a, s') \) for each transition
  - A start state
  - Possibly a terminal state (or absorbing state)
  - Utility function which is additive (discounted) rewards

- MDPs are fully observable but probabilistic search problems
Example: Grid World

- A maze-like problem
  - The agent lives in a grid
  - Walls block the agent’s path
- Noisy movement: actions do not always go as planned
  - 80% of the time, the action North takes the agent North (if there is no wall there)
  - 10% of the time, North takes the agent West; 10% East
  - If there is a wall in the direction the agent would have been taken, the agent stays put
- The agent receives rewards each time step
  - Small “living” reward $r$ each step (can be negative)
  - Big rewards come at the end (good or bad)
- Goal: maximize sum of rewards
A policy $\pi$ gives an action for each state, $\pi: S \rightarrow A$

In deterministic single-agent search problems, we wanted an optimal plan, or sequence of actions, from start to a goal.

For MDPs, we want an optimal policy $\pi^*: S \rightarrow A$
- An optimal policy maximizes expected utility
- An explicit policy defines a reflex agent
Optimal policy for $r > 0$
Optimal policy for $r > 0$
Figure 17.2 (a) The optimal policies for the stochastic environment with $r = -0.04$ for transitions between nonterminal states. There are two policies because in state (3,1) both *Left* and *Up* are optimal. (b) Optimal policies for four different ranges of $r$. 
Utilities of Sequences
Utilities of Sequences

- What preferences should an agent have over reward sequences?
- More or less?  
  \[ [1, 2, 2] \quad \text{or} \quad [2, 3, 4] \]
- Now or later?  
  \[ [0, 0, 1] \quad \text{or} \quad [1, 0, 0] \]
Theorem: if we assume *stationary preferences*:

\[ [s_0, a_0, s_1, a_1, s_2, \ldots] > [s'_0, a'_0, s'_1, a'_1, s'_2, \ldots], \quad s_0 = s'_0, \quad a_0 = a'_0, \quad \text{and} \quad s_1 = s'_1, \]

\[ \iff [s_1, a_1, s_2, \ldots] > [s'_1, a'_1, s'_2, \ldots] \]

then there is only one way to define utilities:

- **Additive discounted utility**:

\[ U_d([s_0, a_0, s_1, a_1, s_2, \ldots]) = R(s_0, a_0, s_1) + \gamma R(s_1, a_1, s_2) + \gamma^2 R(s_2, a_2, s_3) + \cdots \]

where \( \gamma \in [0,1] \) is the *discount factor*
Discounting

Discounting with conveniently solves the problem of infinite reward streams!

- Geometric series: \( 1 + \gamma + \gamma^2 + \ldots = \frac{1}{1 - \gamma} \)
- Assume rewards bounded by \( \pm R_{\text{max}} \)
- Then \( r_0 + \gamma r_1 + \gamma^2 r_2 + \ldots \) is bounded by \( \pm \frac{R_{\text{max}}}{1 - \gamma} \)

(Another solution: environment contains a terminal state; and agent reaches it with probability 1)