Decision Trees - Influence Diagrams
Descriptive Decision Theory tries to simulate human behavior in finding the right or best decision for a given problem

Example:
- Company can chose one of two places for a new store
- Option 1: 125,000 EUR profit per year
- Option 2: 150,000 EUR profit per year

Company should take Option 2, because it maximized the profit.
In real world not everything is known, so there are uncertainties in the model

Example:
- There are plans for restructure the local traffic, which changes the predicted profit
- Option 1: 125,000 EUR profit per year
- Option 2: 80,000 EUR profit per year

With modification Option 1 is the better one and without modification Option 2 is the better one

To model these variations in the environment we use so called Decision Tables

<table>
<thead>
<tr>
<th></th>
<th>$z_1$ (no modification)</th>
<th>$z_2$ (restructure)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$ (Option 1)</td>
<td>125,000 = $e_{11}$</td>
<td>125,000 = $e_{12}$</td>
</tr>
<tr>
<td>$a_2$ (Option 2)</td>
<td>150,000 = $e_{21}$</td>
<td>80,000 = $e_{22}$</td>
</tr>
</tbody>
</table>
An alternative $a_1$ dominates $a_2$ iff the value of $a_1$ is always greater or equal to the value of $a_2$
That means, for all $j$: $e_{1j} \geq e_{2j}$

Example:

<table>
<thead>
<tr>
<th></th>
<th>$z_1$</th>
<th>$z_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>150.000 = $e_{11}$</td>
<td>90.000 = $e_{12}$</td>
</tr>
<tr>
<td>$a_2$</td>
<td>125.000 = $e_{21}$</td>
<td>80.000 = $e_{22}$</td>
</tr>
</tbody>
</table>

Alternative $a_2$ could be dropped
Some more alternatives:

<table>
<thead>
<tr>
<th></th>
<th>$z_1$</th>
<th>$z_2$</th>
<th>$z_3$</th>
<th>$z_4$</th>
<th>$z_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>0</td>
<td>20</td>
<td>10</td>
<td>60</td>
<td>25</td>
</tr>
<tr>
<td>$a_2$</td>
<td>-20</td>
<td>80</td>
<td>10</td>
<td>10</td>
<td>60</td>
</tr>
<tr>
<td>$a_3$</td>
<td>20</td>
<td>60</td>
<td>20</td>
<td>60</td>
<td>50</td>
</tr>
<tr>
<td>$a_4$</td>
<td>55</td>
<td>40</td>
<td>60</td>
<td>10</td>
<td>40</td>
</tr>
<tr>
<td>$a_5$</td>
<td>50</td>
<td>10</td>
<td>30</td>
<td>5</td>
<td>20</td>
</tr>
</tbody>
</table>

- $a_3$ dominated $a_1$
- $a_4$ dominated $a_5$

Alternatives $a_1$ and $a_5$ could be dropped
Various Decision Rules are available

- Maximin - Rule
- Maximax - Rule
- Hurwicz - Rule
- Laplace - Rule
Maximin - Rule

Choose the one with the highest minimum

Contrare To pessimistic, only focus on one column

Example

<table>
<thead>
<tr>
<th></th>
<th>$z_1$</th>
<th>$z_2$</th>
<th>$z_3$</th>
<th>$z_4$</th>
<th>Minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>60</td>
<td>30</td>
<td>50</td>
<td>60</td>
<td>30</td>
</tr>
<tr>
<td>$a_2$</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>140</td>
<td>10</td>
</tr>
<tr>
<td>$a_3$</td>
<td>-30</td>
<td>100</td>
<td>120</td>
<td>130</td>
<td>-30</td>
</tr>
</tbody>
</table>
Maximax - Rule

<table>
<thead>
<tr>
<th></th>
<th>$z_1$</th>
<th>$z_2$</th>
<th>$z_3$</th>
<th>$z_4$</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>60</td>
<td>30</td>
<td>50</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>$a_2$</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>140</td>
<td>140</td>
</tr>
<tr>
<td>$a_3$</td>
<td>-30</td>
<td>100</td>
<td>120</td>
<td>130</td>
<td>130</td>
</tr>
</tbody>
</table>

Choose the one with the highest maximum

**Contra** To optimistic, only focus on one column

Example

<table>
<thead>
<tr>
<th></th>
<th>$z_1$</th>
<th>$z_2$</th>
<th>$z_3$</th>
<th>$z_4$</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>1,000,000</td>
<td>1,000,000</td>
<td>1,000,000</td>
<td>1,000,000</td>
<td>1,000,000</td>
</tr>
<tr>
<td>$a_2$</td>
<td>1,000,001</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1,000,001</td>
</tr>
</tbody>
</table>
Combination of Maximin and Maximax - Rule
\[ \Phi(a) = \lambda \cdot \max(e_i) + (1 - \lambda) \cdot \min(e_i) \]
\( \lambda \) represents readiness to assume risk

**Contra** Only focus on two column

Example (\( \min(a_1) < \min(a_2) \), \( \max(a_1) < \max(a_2) \) ⇒ chose \( a_2 \))
Choose the one with the highest mean value

**Contra**
- Not every condition has the same probability
- Duplication of one condition could change the result

Most people would also chose \( a_3 \) in this example
Probability-based Decisions

In many cases probabilities could be assigned to each option

**Objective Probabilities** based on mathematic or statistic background

**Subjective Probabilities** based on intuition or estimations

Example:
- The management estimates the probability for the restructure to 30%

The decision can be chosen by analyzing the expected values

<table>
<thead>
<tr>
<th></th>
<th>$z_1$ (no modification)</th>
<th>$z_2$ (restructure)</th>
<th>Expectation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$ (Option 1)</td>
<td>$p_1 = 0.7$</td>
<td>$125.000 = e_{11}$</td>
<td>$125.000 = e_{12}$</td>
</tr>
<tr>
<td>$a_2$ (Option 2)</td>
<td>$p_2 = 0.3$</td>
<td>$150.000 = e_{21}$</td>
<td>$80.000 = e_{22}$</td>
</tr>
</tbody>
</table>

Option 2 has the higher expectation and should be used
Probability Domination

<table>
<thead>
<tr>
<th></th>
<th>$z_1$</th>
<th>$z_2$</th>
<th>$z_3$</th>
<th>$z_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_1$</td>
<td>0.3</td>
<td>0.2</td>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>$a_1$</td>
<td>20</td>
<td>40</td>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td>$a_2$</td>
<td>60</td>
<td>30</td>
<td>50</td>
<td>20</td>
</tr>
</tbody>
</table>

Probability Domination means that the cumulated probability for the payout for $a$ is always higher.

**Algorithm**
- Order payout by value in a decreasing order
- Cumulate probabilities

**Example**
- $a_1 : 50(0.1) \ 40(0.2) \ 20(0.3) \ 10(0.4)$
- $a_2 : 60(0.3) \ 50(0.4) \ 30(0.2) \ 20(0.1)$
Example

\[ a_1 : \ 50(0.1) \ 40(0.2) \ 20(0.3) \ 10(0.4) \]
\[ a_2 : \ 60(0.3) \ 50(0.4) \ 30(0.2) \ 20(0.1) \]

\[ a_2 \text{ dominates } a_1 \]
Multi Criteria Decisions - Example

<table>
<thead>
<tr>
<th></th>
<th>Sales $e_1$</th>
<th>Profit $e_2$</th>
<th>Environment Pollution $e_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>800</td>
<td>7000</td>
<td>-4</td>
</tr>
<tr>
<td>$a_2$</td>
<td>600</td>
<td>7000</td>
<td>-2</td>
</tr>
<tr>
<td>$a_3$</td>
<td>400</td>
<td>6000</td>
<td>0</td>
</tr>
<tr>
<td>$a_4$</td>
<td>200</td>
<td>4000</td>
<td>0</td>
</tr>
</tbody>
</table>

**Efficient Alternatives**

- Only focus on alternatives which are not dominated by others
- Example: Drop $a_4$

**Finding a Decision**

- If multiple alternatives are effective we need an algorithm to choose the preferred one
Goal find a function $U(e_1, e_2, \ldots, e_n)$ as a combination of all targets/criteria, which could be optimized.

**Linear combination**
- Simplest variant: Linear combination of all targets using *weights* for criteria.
- $U(e_1, e_2, \ldots, e_i) = \sum_{i=1}^{n} \omega_i \cdot e_i$

**Example**
- $\omega_1 = 10$, $\omega_2 = 1$, $\omega_3 = 500$

<table>
<thead>
<tr>
<th></th>
<th>Sales $e_1$</th>
<th>Profit $e_2$</th>
<th>Environment Pollution $e_3$</th>
<th>$U(e_1, e_2, e_3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>800</td>
<td>7000</td>
<td>-4</td>
<td>13000</td>
</tr>
<tr>
<td>$a_2$</td>
<td>600</td>
<td>7000</td>
<td>-2</td>
<td>12000</td>
</tr>
<tr>
<td>$a_3$</td>
<td>400</td>
<td>6000</td>
<td>0</td>
<td>10000</td>
</tr>
</tbody>
</table>

Optimum
Preference Orderings

Note that transitivity is not always given in decision making.

Consider the following set of dice (so called Efron-Dice):

- Dice A has sides: 2, 2, 4, 4, 9, 9
- Dice B has sides: 1, 1, 6, 6, 8, 8
- Dice C has sides: 3, 3, 5, 5, 7, 7

The probability that A rolls a higher number than B, the probability that B rolls higher than C, and the probability that C rolls higher than A are all $\frac{5}{9}$, so this set of dice lead to nontransitive decisions. In fact, it has the property that, for each dice in the set, there is another dice that rolls a higher number than in more than half the time.

Standard economic theory assumes that preferences are transitive.

In most real applications there are good arguments for imposing such “rationality requirements”, e.g. the money pump argument, or the Dutch Book argument.
A preference ordering \( \leq \) is a ranking of all possible states of affairs (worlds) \( S \)

- these could be outcomes of actions, truth assignments, states in a search problem, etc.

- \( s \leq t \): means that state \( t \) is at least as good as \( s \)

- \( s > t \): means that state \( s \) is strictly preferred to \( t \)

We insist that \( \leq \) is

- reflexive: i.e., \( s \leq s \) for all states \( s \)

- transitive: i.e., if \( s \leq t \) and \( t \leq w \), then \( s \leq w \)

- connected: for all states \( s,t \), either \( s \leq t \) or \( t \leq s \)
Rather than just ranking outcomes, we are often able to quantify our degree of preference

A *utility function* $U : S \rightarrow \mathbb{R}$ associates a real-valued *utility* with each outcome.

- $U(s)$ measures the *degree* of preference for $s$

Note: $U$ induces a preference ordering $\leq_U$ over $S$ defined as: $s \leq_U t$ iff $U(s) \leq U(t)$

- $\leq_U$ is reflexive, transitive, and connected
Under conditions of uncertainty, each decision \( d \) induces a distribution \( P_d \) over possible outcomes

- \( P_d(s) \) is probability of outcome \( s \) under decision \( d \)

The **expected utility** of decision \( d \) is defined by

\[
EU(d) = \sum_{s \in S} P_d(s) U(s)
\]

The **principle of maximum expected utility (MEU)** states that the optimal decision under conditions of uncertainty is that with the greatest expected utility.
Decision Trees
in Machine Learning
Assignment of a drug to a patient:

Decision Trees (in Machine Learning)

Blood pressure

Drug A

Drug B

Age

Drug A

Drug B

≤ 40

> 40

high

normal

low
Recursive Descent:

Start at the root node.

If the current node is an leaf node:
  ◦ Return the class assigned to the node.

If the current node is an inner node:
  ◦ Test the attribute associated with the node.
  ◦ Follow the branch labeled with the outcome of the test.
  ◦ Apply the algorithm recursively.

Intuitively: Follow the path corresponding to the case to be decided.
Example: Assignment of a drug to a patient

Blood pressure

- high: Drug A
- normal: Age
  - ≤ 40: Drug A
  - > 40: Drug B
- low: Drug B
Assignment of a drug to a patient:

Decision Trees in Machine Learning

Blood pressure

- high: Drug A
- normal: Age
  - ≤ 40: Drug A
  - > 40: Drug B
- low: Drug B
Assignment of a drug to a patient:

- **Blood pressure**
  - **high** → Drug A
  - **normal**
  - **low** → Drug B

- **Age**
  - ≤ 40 → Drug A
  - > 40 → Drug B
Learning of Decision Trees

Top-down approach

◦ Build the decision tree from top to bottom (from the root to the leaves).

Greedy Selection of a Test Attribute

◦ Compute an evaluation measure for all attributes.
◦ Select the attribute with the best evaluation.

Divide and Conquer / Recursive Descent

◦ Divide the example cases according to the values of the test attribute.
◦ Apply the procedure recursively to the subsets.
◦ Terminate the recursion if
  ◦ all cases belong to the same class
  ◦ no more test attributes are available
Induction of a Decision Tree: Example

Patient database

- 12 example cases
- 3 descriptive attributes
- 1 class attribute

Assignment of drug
(without patient attributes)
always drug A or always drug B:
50% correct (in 6 of 12 cases)

<table>
<thead>
<tr>
<th>No</th>
<th>Sex</th>
<th>Age</th>
<th>Blood pr.</th>
<th>Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>male</td>
<td>20</td>
<td>normal</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>female</td>
<td>73</td>
<td>normal</td>
<td>B</td>
</tr>
<tr>
<td>3</td>
<td>female</td>
<td>37</td>
<td>high</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>male</td>
<td>33</td>
<td>low</td>
<td>B</td>
</tr>
<tr>
<td>5</td>
<td>female</td>
<td>48</td>
<td>high</td>
<td>A</td>
</tr>
<tr>
<td>6</td>
<td>male</td>
<td>29</td>
<td>normal</td>
<td>A</td>
</tr>
<tr>
<td>7</td>
<td>female</td>
<td>52</td>
<td>normal</td>
<td>B</td>
</tr>
<tr>
<td>8</td>
<td>male</td>
<td>42</td>
<td>low</td>
<td>B</td>
</tr>
<tr>
<td>9</td>
<td>male</td>
<td>61</td>
<td>normal</td>
<td>B</td>
</tr>
<tr>
<td>10</td>
<td>female</td>
<td>30</td>
<td>normal</td>
<td>A</td>
</tr>
<tr>
<td>11</td>
<td>female</td>
<td>26</td>
<td>low</td>
<td>B</td>
</tr>
<tr>
<td>12</td>
<td>male</td>
<td>54</td>
<td>high</td>
<td>A</td>
</tr>
</tbody>
</table>
Induction of a Decision Tree: Example

Sex of the patient
Division w.r.t. male/female.

Assignment of drug
male: 50% correct (in 3 of 6 cases)
female: 50% correct (in 3 of 6 cases)

<table>
<thead>
<tr>
<th>No</th>
<th>Sex</th>
<th>Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>male</td>
<td>A</td>
</tr>
<tr>
<td>6</td>
<td>male</td>
<td>A</td>
</tr>
<tr>
<td>12</td>
<td>male</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>male</td>
<td>B</td>
</tr>
<tr>
<td>8</td>
<td>male</td>
<td>B</td>
</tr>
<tr>
<td>9</td>
<td>male</td>
<td>B</td>
</tr>
<tr>
<td>3</td>
<td>female</td>
<td>A</td>
</tr>
<tr>
<td>5</td>
<td>female</td>
<td>A</td>
</tr>
<tr>
<td>10</td>
<td>female</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>female</td>
<td>B</td>
</tr>
<tr>
<td>7</td>
<td>female</td>
<td>B</td>
</tr>
<tr>
<td>11</td>
<td>female</td>
<td>B</td>
</tr>
</tbody>
</table>

total: 50% correct (in 6 of 12 cases)
Induction of a Decision Tree: Example

Age of the patient

- Sort according to age.
- Find best age split.
- Here: ca. 40 years

Assignment of drug

<table>
<thead>
<tr>
<th>No</th>
<th>Age</th>
<th>Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>A</td>
</tr>
<tr>
<td>11</td>
<td>26</td>
<td>B</td>
</tr>
<tr>
<td>6</td>
<td>29</td>
<td>A</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>33</td>
<td>B</td>
</tr>
<tr>
<td>3</td>
<td>37</td>
<td>A</td>
</tr>
<tr>
<td>8</td>
<td>42</td>
<td>B</td>
</tr>
<tr>
<td>5</td>
<td>48</td>
<td>A</td>
</tr>
<tr>
<td>7</td>
<td>52</td>
<td>B</td>
</tr>
<tr>
<td>12</td>
<td>54</td>
<td>A</td>
</tr>
<tr>
<td>9</td>
<td>61</td>
<td>B</td>
</tr>
<tr>
<td>2</td>
<td>73</td>
<td>B</td>
</tr>
</tbody>
</table>

- $\leq 40$: A. 67% correct (in 4 of 6 cases)
- $> 40$: B. 67% correct (in 4 of 6 cases)

Total: 67% correct (in 8 of 12 cases)
Induction of a Decision Tree: Example

Blood pressure of the patient
Division w.r.t. high/normal/low.

Assignment of drug

<table>
<thead>
<tr>
<th>No</th>
<th>Blood pr.</th>
<th>Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>high</td>
<td>A</td>
</tr>
<tr>
<td>5</td>
<td>high</td>
<td>A</td>
</tr>
<tr>
<td>12</td>
<td>high</td>
<td>A</td>
</tr>
<tr>
<td>1</td>
<td>normal</td>
<td>A</td>
</tr>
<tr>
<td>6</td>
<td>normal</td>
<td>A</td>
</tr>
<tr>
<td>10</td>
<td>normal</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>normal</td>
<td>B</td>
</tr>
<tr>
<td>7</td>
<td>normal</td>
<td>B</td>
</tr>
<tr>
<td>9</td>
<td>normal</td>
<td>B</td>
</tr>
<tr>
<td>4</td>
<td>low</td>
<td>B</td>
</tr>
<tr>
<td>8</td>
<td>low</td>
<td>B</td>
</tr>
<tr>
<td>11</td>
<td>low</td>
<td>B</td>
</tr>
</tbody>
</table>

high: A 100% correct (in 3 of 3 cases)

normal: 50% correct (in 3 of 6 cases)

low: B 100% correct (in 3 of 3 cases)

total: 75% correct (in 9 of 12 cases)
Induction of a Decision Tree: Example

Current Decision Tree:

- Blood pressure
  - high: Drug A
  - normal: ?
  - low: Drug B
Induction of a Decision Tree: Example

Blood pressure and sex

Only patients with normal blood pressure.
Division w.r.t. male/female.

Assignment of drug

<table>
<thead>
<tr>
<th>No</th>
<th>Blood pr.</th>
<th>Sex</th>
<th>Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>normal</td>
<td>male</td>
<td>A</td>
</tr>
<tr>
<td>6</td>
<td>normal</td>
<td>male</td>
<td>A</td>
</tr>
<tr>
<td>9</td>
<td>normal</td>
<td>male</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>normal</td>
<td>female</td>
<td>B</td>
</tr>
<tr>
<td>7</td>
<td>normal</td>
<td>female</td>
<td>B</td>
</tr>
<tr>
<td>10</td>
<td>normal</td>
<td>female</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>low</td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>8</td>
<td>low</td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>11</td>
<td>low</td>
<td></td>
<td>B</td>
</tr>
</tbody>
</table>

male: A 67% correct (2 of 3)
female: B 67% correct (2 of 3)
total: 67% correct (4 of 6)
Induction of a Decision Tree: Example

Blood pressure and age

Only patients with normal blood pressure.
Sort according to age.
Find best age split.
here: ca. 40 years

Assignment of drug

≤ 40:  A. 100% correct (3 of 3)
> 40:  B. 100% correct (3 of 3)

<table>
<thead>
<tr>
<th>No</th>
<th>Blood pr.</th>
<th>Age</th>
<th>Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>high</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>5</td>
<td>high</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>12</td>
<td>high</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>1</td>
<td>normal</td>
<td>20</td>
<td>A</td>
</tr>
<tr>
<td>6</td>
<td>normal</td>
<td>29</td>
<td>A</td>
</tr>
<tr>
<td>10</td>
<td>normal</td>
<td>30</td>
<td>A</td>
</tr>
<tr>
<td>7</td>
<td>normal</td>
<td>52</td>
<td>B</td>
</tr>
<tr>
<td>9</td>
<td>normal</td>
<td>61</td>
<td>B</td>
</tr>
<tr>
<td>2</td>
<td>normal</td>
<td>73</td>
<td>B</td>
</tr>
<tr>
<td>11</td>
<td>low</td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>4</td>
<td>low</td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>8</td>
<td>low</td>
<td></td>
<td>B</td>
</tr>
</tbody>
</table>

total: 100% correct (6 of 6)
Assignment of a drug to a patient:

- **Blood pressure**: high → Drug A, normal → Drug A, low → Drug B
- **Age**: ≤ 40 → Drug A, > 40 → Drug B
Evaluation measure used in the above example: 
Rate of correctly classified example cases.

- Advantage: simple to compute, easy to understand.
- Disadvantage: works well only for two classes.

If there are more than two classes, the rate of misclassified example cases 
Neglects a lot of the available information.

- Only the majority class—that is, the class occurring most often in (a subset of) the example cases—is really considered.
- The distribution of the other classes has no influence. However, a good choice here can be important for deeper levels of the decision tree.

Therefore: Several other evaluation measures are studied, e.g.

Information gain and its various normalizations.
An Information-theoretic Evaluation Measure

Information Gain  
(Kullback and Leibler 1951, Quinlan 1986)

Based on Shannon Entropy  
\[ H = - \sum_{i=1}^{n} p_i \log_2 p_i \]  
(Shannon 1948)

\[
I_{gain}(C, A) = H(C) - H(C|A)
\]

\[
= - \sum_{i=1}^{n_C} p_{i.} \log_2 p_{i.} - \sum_{j=1}^{n_A} p_{.j} \left( - \sum_{i=1}^{n_C} p_{i|j} \log_2 p_{i|j} \right)
\]

- \( H(C) \)  
  Entropy of the class distribution  
  \( (C: \text{class attribute}) \)

- \( H(C|A) \)  
  Expected entropy of the class distribution  
  if the value of the attribute \( A \) becomes known

- \( H(C) - H(C|A) \)  
  Expected entropy reduction or information gain
Inducing the Decision Tree by Information Gain

Information gain for drug and sex:

\[
H(\text{Drug}) = -\left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2}\right) = 1
\]

\[
H(\text{Drug} | \text{Sex}) = \frac{1}{2} \left(\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2}\right) + \frac{1}{2} \left(\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2}\right) = 1
\]

\[
I_{gain}(\text{Drug, Sex}) = 1 - 1 = 0
\]

No gain at all since the initial the uniform distribution of drug is splitted into two (still) uniform distributions.
Inducing the Decision Tree by Information Gain

Information gain for drug and age:

\[
H(\text{Drug}) = -\left( \frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2} \right) = 1
\]

\[
H(\text{Drug} \mid \text{Age}) = \frac{1}{2} \left( -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} \right) + \frac{1}{2} \left( -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} \right) \approx 0.9183
\]

\[
I_{\text{gain}}(\text{Drug}, \text{Age}) = 1 - 0.9183 = 0.0817
\]

Splitting w.r.t. age can reduce the overall entropy.
Inducing the Decision Tree by Information Gain

Information gain for drug and blood pressure:

\[ H(\text{Drug}) = -\left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2}\right) = 1 \]

\[ H(\text{Drug} \mid \text{Blood}_\text{pr}) = \frac{1}{4} \cdot 0 + \frac{1}{2} \left( \frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} \right) + \frac{1}{4} \cdot 0 = 0.5 \]

\[ I_{\text{gain}}(\text{Drug}, \text{Blood}_\text{pr}) = 1 - 0.5 = 0.5 \]

Largest information gain, so we first split w. r. t. blood pressure (as in the example with misclassification rate).
Inducing the Decision Tree by Information Gain

Next level: Subtree blood pressure is normal.

Information gain for drug and sex:

\[
H(\text{Drug}) = - \left( \frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2} \right) = 1
\]

\[
H(\text{Drug} \mid \text{Sex}) = \frac{1}{2} \left( \frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} \right) + \frac{1}{2} \left( \frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} \right) = 0.9183
\]

\[
I_{\text{gain}}(\text{Drug}, \text{Sex}) = 0.0817
\]

Entropy can be decreased.
Next level: Subtree blood pressure is normal.

Information gain for drug and age:

\[
H(\text{Drug}) = -\left( \frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2} \right) = 1
\]

\[
H(\text{Drug} \mid \text{Age}) = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 0 = 0
\]

\[
I_{\text{gain}}(\text{Drug}, \text{Age}) = 1
\]

Maximal information gain, that is we result in a perfect classification (again, as in the case of using misclassification rate).
Decision Trees
in Economics
Decision Trees (in Economy, with Uncertainty)

Should I buy NASDIP or leave my $1000 in bank?

Probabilities:
- Buy NASDIP: 0.25
- Leave $1000 in bank: 0.5

Outcomes:
- NASDIP: $500, $1000, $2000
- Leave $1000 in bank: $1005
Decision Trees

Should I buy NASDIP or leave my $1000 in bank?

Maximal Expected Value

Buy NASDIP

$1375

.25

$500

.25

$1000

.5

$2000

$1005

Leave $1000 in bank

$1005
Expected Utility

Where do utilities come from?
  ◦ underlying foundations of utility theory tightly couple utility with action/choice
  ◦ a utility function can be determined by asking someone about their preferences for actions in specific scenarios (or “lotteries” over outcomes)

Utility functions needn’t be unique
  ◦ if I multiply $U$ by a positive constant, all decisions have same relative utility
  ◦ if I add a constant to $U$, same thing
  ◦ $U$ is unique up to positive affine transformation
A decision problem under uncertainty is:

- a set of decisions $D$
- a set of outcomes or states $S$
- an outcome function $P : D \rightarrow \Delta(S)$
  - $\Delta(S)$ is the set of distributions over $S$ (e.g., $P_d$)
  - a utility function $U$ over $S$

A solution to a decision problem under uncertainty is any $d^* \in D$ such that $EU(d) \leq EU(d^*)$ for all $d \in D$
Influence Diagrams
Influence Diagrams vs Decision Trees

Outcome space is large
- like all of our problems, states spaces can be huge
- don’t want to spell out distributions like $P_d$ explicitly
- Solution: Extend Bayes Networks with decision nodes and utility nodes
- Use **Influence Diagrams**

Should I buy NASDIP or leave my money in the bank?

$P(\text{NASDIP} = \$5) = .25$
$P(\text{NASDIP} = \$10) = .25$
$P(\text{NASDIP} = \$20) = .5$

$d1 = \text{Buy NASDIP}$
$d2 = \text{Leave \$1000 in bank}$

$U(d1, \$5) = \$500$
$U(d1, \$10) = \$1000$
$U(d1, \$20) = \$2000$
$U(d2, n) = \$1005$
Decision space is large

- usually our decisions are not one-shot actions
- rather they involve sequential choices (like plans)
- if we treat each plan as a distinct decision, decision space is too large to handle directly

Solution: use dynamic programming methods to *construct* optimal plans (actually generalizations of plans, called policies . . . like in game trees)
Sam has a non-small-cell carcinoma of the lung.
The preferred treatment in this situation is a thoracotomy.
The alternative treatment is radiation.
We don’t know if mediastinal metastasis is present.
If mediastinal metastasis is present, a thoracotomy would be contraindicated because it subjects the patient to a risk of death with no health benefit.
If mediastinal metastases are absent, a thoracotomy offers a substantial survival advantage as long as the primary tumor has not metastasized to distant organs.
We have two tests available for assessing the involvement of the mediastinum.
They are CT scan and mediastinoscopy.
This problem involves three decisions:

1. Should Sam undergo a CT scan?
2. Second, given this decision and any CT results, should he undergo mediastinoscopy?
3. Third, given these decisions and any test results, should the patient undergo a thoracotomy?
Influence diagrams provide a way of representing sequential decision problems

- basic idea: represent the variables in the problem as you would in a BN
- add decision variables – variables that you “control”
- add utility variables – how good different states are
Simple Example: Influence Diagrams

- **C** = Do CT
  - **c1** = Do CT
  - **c2** = Do not do

- **T** = Do thoracotomy
  - **t1** = Do thoracotomy
  - **t2** = Do radiation

- **E** = Do mediastinoscopy
  - **e1** = Do mediastinoscopy
  - **e2** = Do not do

- **U**
  - **U(t1,present,tlive,mLive) = 1.8 yrs**
  - **U(t1,absent,tlive,mLive) = 4.45 yrs**
  - **U(t2,present,tlive,mLive) = 1.8 yrs**
  - **U(t2,absent,tlive,mLive) = 2.64 yrs**
  - **U(t,m,tdie,d) = 0**
  - **U(t,m,d,mdie) = 0**

- **P(present) = .46**
- **P(cpos|present,c1) = .82**
- **P(cpos|absent,c1) = .19**
- **P(notrun|m,c2) = 1**
- **P(mpos|present,e1) = .82**
- **P(mpos|absent,e1) = .005**
- **P(notrun|m,e2) = 1**
- **P(tdie|t1) = .037**
- **P(tdie|t2) = .002**
- **P(mdie|e1) = .005**
- **P(mdie|e2) = 0**

Rudolf Kruse
Bayesian Networks
Chance Nodes

- **Chance nodes**
  - random variables, denoted by circles
  - as in a BN, probabilistic dependence on parents

![Diagram with probabilities]

Pr(f|flu) = .5
Pr(f|mal) = .3
Pr(f|none) = .05

Pr(pos|flu, bt) = .2
Pr(neg|flu, bt) = .8
Pr(null|flu, bt) = 0
Pr(pos|mal, bt) = .9
Pr(neg|mal, bt) = .1
Pr(null|mal, bt) = 0
Pr(pos|no, bt) = .1
Pr(neg|no, bt) = .9
Pr(null|no, bt) = 0
Pr(pos|D, ~bt) = 0
Pr(neg|D, ~bt) = 0
Pr(null|D, ~bt) = 1
Decision Nodes

Decision nodes
- variables decision maker sets, denoted by squares
- parents reflect *information available* at time decision is to be made

In example decision node: the actual values of Chills and Fever will be observed before the decision to take test must be made
- agent can make different decisions for each instantiation of parents (i.e., policies)

\[
\text{BloodTst} \quad \text{BT} \in \{bt, \sim bt\}
\]

Chills
Fever
Utility Nodes

Utility node
- specifies utility of a state, denoted by a diamond
- utility depends *only on state of parents* of value node
- generally: only one value node in a decision network

Utility depends only on disease and drug

$U(\text{fludrug, flu}) = 20$
$U(\text{fludrug, mal}) = -300$
$U(\text{fludrug, none}) = -5$
$U(\text{maldrug, flu}) = -30$
$U(\text{maldrug, mal}) = 10$
$U(\text{maldrug, none}) = -20$
$U(\text{no drug, flu}) = -10$
$U(\text{no drug, mal}) = -285$
$U(\text{no drug, none}) = 30$
In **Standard Influence Diagrams** two assumptions are made:

**Decision nodes are totally ordered**
- decision variables $D_1, D_2, \ldots, D_n$
- decisions are made in sequence

**No-forgetting property**
- any information available when decision $D_i$ is made is available when decision $D_j$ is made (for $i < j$)

Thus all parents of $D_i$ are parents of $D_j$

Example: BloodTest is done before Drug Assignment, and at the time of the Drug Assignment the decision maker is aware of the result of the Blood Test.

In **Non-Standard Influence Diagrams** other assumptions may hold

Example: The solution of a **Limited Memory Influence Diagram** (LIMID, used in HUGIN) is a strategy consisting of one policy for each decision. The policy is a function from the known variables to the states of the decision. It is not a function of all past observations as the decision maker is assumed only to know the most recent observation. This is different from the traditional influence diagram where the policy would be a function from all past observations and decisions as the decision maker is assumed to be non-forgetting. **There need not be a total order on the decisions.**
Let $Par(D_i)$ be the parents of decision node $D_i$

- $Dom(Par(D_i))$ is the set of assignments to parents

A policy $\delta$ is a set of mappings $\delta_i$, one for each decision node $D_i$

- $\delta_i : Dom(Par(D_i)) \rightarrow (D_i)$

- $\delta_i$ associates a decision with each parent assignment for $D_i$

For example, a policy for BT might be:

\[
\begin{align*}
\delta_{BT}(c, f) &= bt \\
\delta_{BT}(c, \sim f) &= \sim bt \\
\delta_{BT}(\sim c, f) &= bt \\
\delta_{BT}(\sim c, \sim f) &= \sim bt
\end{align*}
\]
Value of a policy $\delta$ is the expected utility given that decision nodes are executed according to $\delta$

Given associates $\mathbf{x}$ to the set $\mathbf{X}$ of all chance variables, let $\delta(\mathbf{x})$ denote the assignment to decision variables dictated by $\delta$

- e.g., assigned to $D_1$ determined by it’s parents’ assignment in $\mathbf{x}$

- e.g., assigned to $D_2$ determined by it’s parents’ assignment in $\mathbf{x}$ along with whatever was assigned to $D1$

- etc.

Value of $\delta$:

$$EU(\delta) = \sum_{\mathbf{X}} P(\mathbf{X}, \delta(\mathbf{X}))U(\mathbf{X}, \delta(\mathbf{X}))$$

An optimal policy is a policy $\delta^*$ such that $EU(\delta^*) \geq EU(\delta)$ for all policies $\delta$
Computing the Best Policy

We can work backwards as follows

First compute optimal policy for Drug (last decision)

- for each assignment to parents (C,F,BT,TR) and for each decision value (D = md, fd, none), compute the expected value of choosing that value of D

- set policy choice for each value of parents to be the value of D that has max value

- eg: $\delta_D(c, f, bt, pos) = md$
Next compute policy for BT given policy $\delta_D(C, \ F, \ BT, \ TR)$ just determined for Drug

- since $\delta_D(C, \ F, \ BT, \ TR)$ is fixed, we can treat Drug as a normal random variable with deterministic probabilities

- i.e., for any instantiation of parents, value of Drug is fixed by policy $\delta_D$

- this means we can solve for optimal policy for BT just as before

- only uninstantiated variables are random variables (once we fix its parents)
You want to buy a used car, but there’s a good chance it is a “lemon” (i.e., prone to breakdown). Before deciding to buy it, you can take it to a mechanic for inspection. S/he will give you a report on the car, labelling it either “good” or “bad”. A good report is positively correlated with the car being sound, while a bad report is positively correlated with the car being a lemon.

The report costs $50 however. So you could risk it, and buy the car without the report.

Owning a sound car is better than having no car, which is better than owning a lemon.
Car Buyer’s Network

Report: good, bad, none

<table>
<thead>
<tr>
<th></th>
<th>g</th>
<th>b</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>l</td>
<td>0.2</td>
<td>0.8</td>
<td>0</td>
</tr>
<tr>
<td>~l</td>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>l</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>~l</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

~ = not

Utility for Lemon and Buy

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>buy l</td>
<td></td>
<td>-600</td>
</tr>
<tr>
<td>buy ~l</td>
<td></td>
<td>1000</td>
</tr>
<tr>
<td>~buy l</td>
<td></td>
<td>-300</td>
</tr>
<tr>
<td>~buy ~l</td>
<td></td>
<td>-300</td>
</tr>
</tbody>
</table>

-50 if inspect
Evaluate Last Decision: Buy (1)

\[ EU(B|I, R) = \sum_L P(L|I, R, B)U(L, B) \]

\[ I = i, R = g: \]

\[ EU(\text{buy}) = P(l|i, g)U(l, \text{buy}) + P(\sim l|i, g)U(\sim l, \text{buy}) - 50 \]
\[ = 0.18 \cdot (-600) + 0.82 \cdot 1000 - 50 = 662 \]
\[ EU(\sim \text{buy}) = P(l|i, g)U(l, \sim \text{buy}) + P(\sim l|i, g)U(\sim l, \sim \text{buy}) - 50 \]
\[ = -300 - 50 = -350 \]
So optimal \( \delta_{Buy}(i, g) = \text{buy} \)

\[ I = i, R = b: \]

\[ EU(\text{buy}) = P(l|i, b)U(l, \text{buy}) + P(\sim l|i, b)U(\sim l, \text{buy}) - 50 \]
\[ = 0.89 \cdot (-600) + .11 \cdot 1000 - 50 = -474 \]
\[ EU(\sim \text{buy}) = P(l|i, b)U(l, \sim \text{buy}) + P(\sim l|i, b)U(\sim l, \sim \text{buy}) - 50 \]
\[ = -300 - 50 = -350 (\text{-300 indep. of lemon}) \]
So optimal \( \delta_{Buy}(i, b) = \sim \text{buy} \)
Evaluate Last Decision: Buy (2)

\[ I = \sim i, R = n \text{ (note: no inspection cost subtracted):} \]

\[
EU(\text{buy}) = P(l | \sim i, n)U(l, \text{buy}) + P(\sim l | \sim i, n)U(\sim l, \text{buy}) \\
= 0.5 \cdot (-600) + 0.5 \cdot 1000 = 200 \\
EU(\sim \text{buy}) = P(l | \sim i, n)U(l, \sim \text{buy}) + P(\sim l | \sim i, n)U(\sim l, \sim \text{buy}) - 50 \\
= -300 - 50 = -350
\]

So optimal is: \( \delta_{Buy}(\sim i, g) = \text{buy} \)

So optimal policy for Buy is:
- \( \delta_{Buy}(i, g) = \text{buy}; \delta_{Buy}(i, b) = \sim \text{buy}; \delta_{Buy}(\sim i, g) = \text{buy} \)

Note: we don’t bother computing policy for the other cases since these occur with probability 0
Evaluate First Decision: Inspect

\[
EU(I) = \sum_{L,R} P(L, R|I)U(L, \delta_{Buy}(I, R)),
\]
where \( P(R, L|I) = P(R|L, I)P(L|I) \)

\[
EU(i) = 0.1 \cdot (-650) + 0.4 \cdot (-300) + 0.45 \cdot 1000 + 0.05 \cdot (-300) - 50
= 187.5
\]

\[
EU(\sim i) = P(l| \sim i, n)U(l, \text{buy}) + P(\sim l| \sim i, n)U(\sim l, \text{buy})
= .5 \cdot -600 + .5 \cdot 1000 = 200
\]

So optimal \( \delta_{Inspect}(\sim i) = \text{buy} \)

| \( P(R, L|I) \) | \( \delta_{Buy} \) | \( U(L, \delta_{Buy}) \) |
|---|---|---|
| \( g, l \) | 0.1 | \text{buy} | -600 - 50 = -650 |
| \( g, \sim l \) | 0.45 | \text{buy} | 1000 - 50 = 950 |
| \( b, l \) | 0.4 | \( \sim \text{buy} \) | -300 - 50 = -350 |
| \( b, \sim l \) | 0.05 | \( \sim \text{buy} \) | -300 - 50 = -350 |
So optimal policy is: don’t inspect, buy the car
  ◦ $EU = 200$

  ◦ Notice that the EU of inspecting the car, then buying it iff you get a good report, is 237.5 less the cost of the inspection (50). So inspection not worth the improvement in EU.

  ◦ But suppose inspection cost $25: then it would be worth it ($EU = 237.5 - 25 = 212.5 > EU(\sim i)$)

  ◦ The *expected value of information* associated with inspection is 37.5 (it improves expected utility by this amount ignoring cost of inspection). Gives opportunity to change decision ($\sim buy$ if bad).

  ◦ You should be willing to pay up to $37.5 for the report