Machine learning: lecture 21

Tommi S. Jaakkola
MIT AI Lab
tommi@ai.mit.edu

Topics

- Graphical models and decision theory
  - utilities, expected utility framework
  - influence diagrams
  - examples
- Exact inference in graphical models
  - basics of probabilistic inference
  - chains and clustering
  - belief propagation and messages

Background: utility theory

- Utilities provide a numerical scale at which to measure preferences about alternative outcomes
- For us to be able to cast our preferences as utilities, our preferences need to satisfy a few (reasonable) conditions:
  - "transitivity": preferring B over A and C over B means that you also prefer C over A
  - "continuity": for any such A, B, and C, there's a probability p such that a lottery where we get A with probability p and C with probability $(1-p)$ is equally preferable to B

(a few other perhaps more self-evident assumptions are needed)

Background: expected utility framework

- Once we have (subjective) utilities for each available action and each possible outcome/situation, we can decide which action we should take

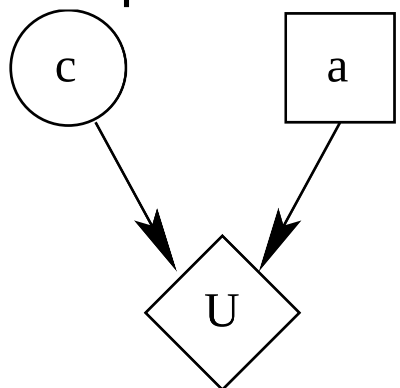
- Expected utility framework:
  Given a probability distribution $P(c)$ over all the uncertain quantities c, we ought to select the action that maximizes the *expected utility*:

  $$a^* = \arg\max_{a \in A} E_{c \sim P} \left\{ U(c, a) \right\} = \arg\max_{a \in A} \sum_c P(c)\, U(c, a)$$

  where A is the set of available actions and $U(c, a)$ is the utility of choosing action a in context c.

Influence diagrams: basics

- Influence diagrams are graphical models that combine beliefs about uncertain quantities, available actions, and associated utilities into a common representation



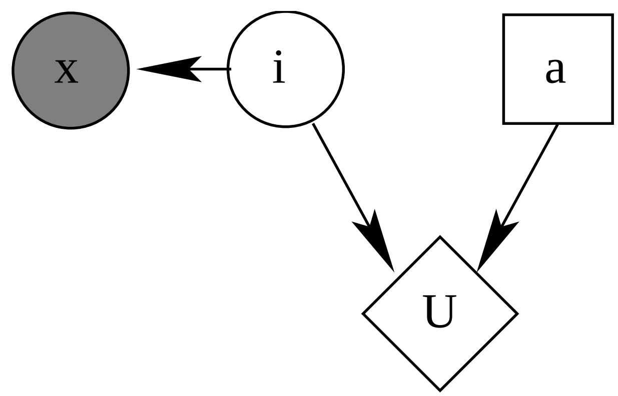
$$a^* = \arg\max_{a \in A} E_{c \sim P} \left\{ U(c, a) \right\}$$

- Basic components:
  - decision node(s) (squares) specify the actions we can take
  - chance nodes (circles) specify our belief about the values of variables relevant for decisions
  - utility node (diamond) specifies the utility of any decision in a contex c

Influence diagrams: example

- Minimum probability of error classification:

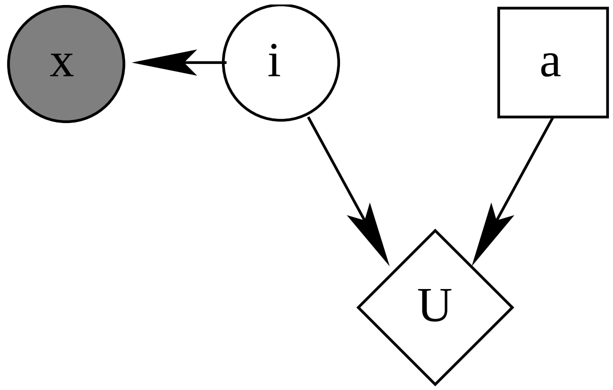
  Suppose we know all the class conditional densities $p(\mathbf{x}|i)$, $i = 1, \ldots, m$, and the prior class probabilities $P(i)$.

  The utility $U(i, a) = 1$ if $i = a$ and zero otherwise (utility is defined as one minus error).

- We do not have access to class i directly but the uncertainty about the class label is captured by the posterior $P(i|\mathbf{x})$

Example cont'd



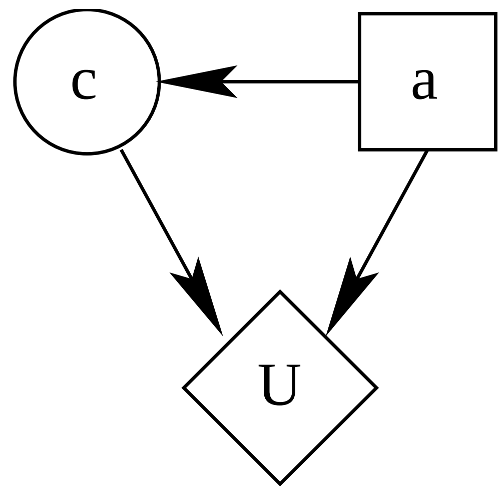
We choose the action/class with the highest expected utility

$$
\begin{aligned}
a^* &= \arg\max_a E_{i \sim P(i|\mathbf{x})} \left\{ U(i,a) \right\} \\
&= \arg\max_a \sum_{i=1}^{m} P(i|\mathbf{x}) U(i,a) \\
&= \arg\max_a P(a|\mathbf{x})
\end{aligned}
$$

which is the action that we have previously seen to minimize the probability of error

Influence diagrams cont'd

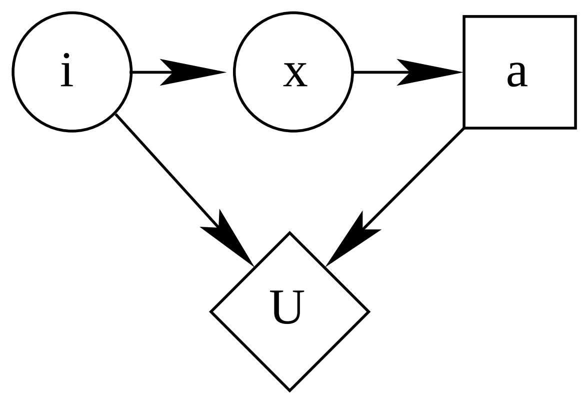
- Our decisions may affect the random quantities (outcomes)



(For example, the course you choose to take can influence the grade you are likely to get)

Influence diagrams cont'd

- We may expect to know the values of some variables at the decision time; our "decisions" become responsive strategies (functions of known quantities)

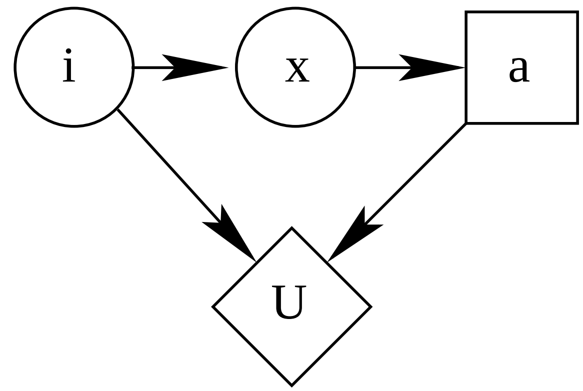


(arrows into the decision node(s) denote the information available at the decision time)

$$a^*(\cdot) \quad = \quad \arg\max_{a(\cdot)} E_{(i,\mathbf{x})\sim P(i)p(\mathbf{x}|i)} \left\{ U(i, a(\mathbf{x})) \right\}$$

where $a(\cdot)$ is a strategy providing an action $a(\mathbf{x})$ for each $\mathbf{x}$.

Influence diagrams cont'd



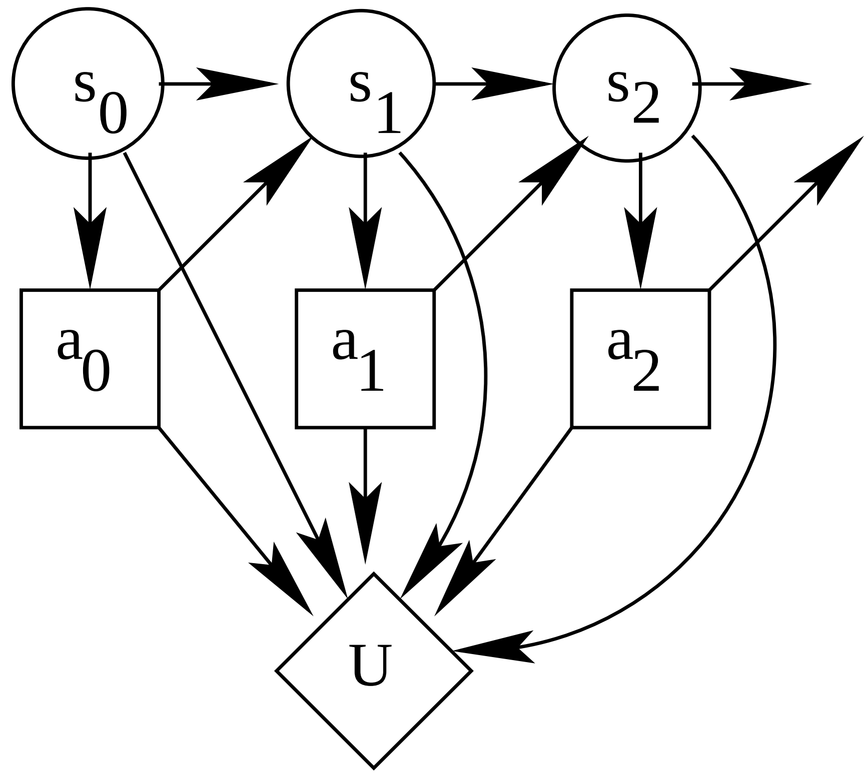
- The resulting strategy will again yield the minimum probability of error provided that $U(i,a) = 1$ when $i = a$ and zero otherwise.

$$
\begin{aligned}
a^*(\cdot) &= \arg\max_{a(\cdot)} \sum_i \int P(i)p(\mathbf{x}|i)\, U(i, a(\mathbf{x}))d\mathbf{x} \\
&= \arg\max_{a(\cdot)} \sum_i \int p(\mathbf{x})P(i|\mathbf{x})\, U(i, a(\mathbf{x}))\, d\mathbf{x} \\
&= \arg\max_{a(\cdot)} \int p(\mathbf{x})\, P(a(\mathbf{x})|\mathbf{x})\, d\mathbf{x}
\end{aligned}
$$

where the maximum is attained when $a^*(\mathbf{x}) = \arg\max_a P(a|\mathbf{x})$.

Influence diagrams cont'd

- A bit more complex example: controlled Markov chain



$$U(s_0, a_0, s_1, a_1, \ldots) = \sum_{t=0}^{n} r(s_t, a_t)$$

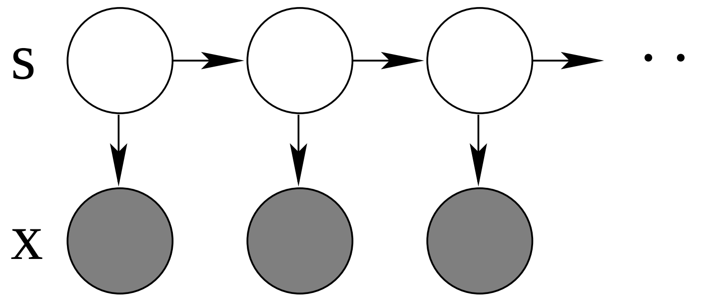
Topics

- Exact inference in graphical models
  - basics of probabilistic inference
  - chains and clustering
  - belief propagation and messages

Nature of probabilistic inference

- Example: a hidden Markov model

$$P(s_0, x_0, \ldots, s_n, x_n) = P_0(s_0)\, P_o(x_0|s_0)\, P_1(s_1|s_0)\, \cdots$$



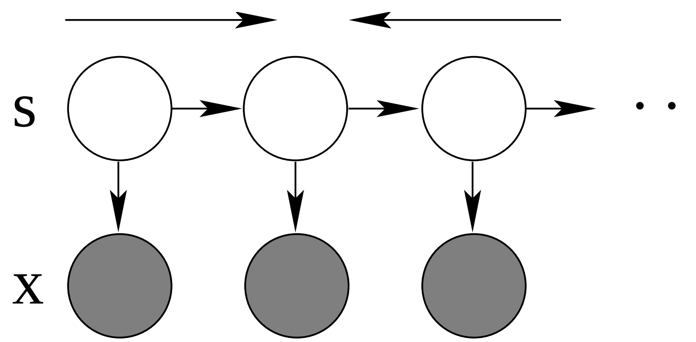
- Given the observation sequence $x_0^*, \ldots, x_n^*$, all the information about the associated hidden states is already contained in the joint probability distribution

$$P(s_0, x_0^*, \ldots, s_n, x_n^*)$$

- What's left to do?

Nature of probabilistic inference

- We have to *explicate* the relevant information; this involves propagating information across the graph model
- Forward-backward algorithm:



Forward step: information from the past about the current state

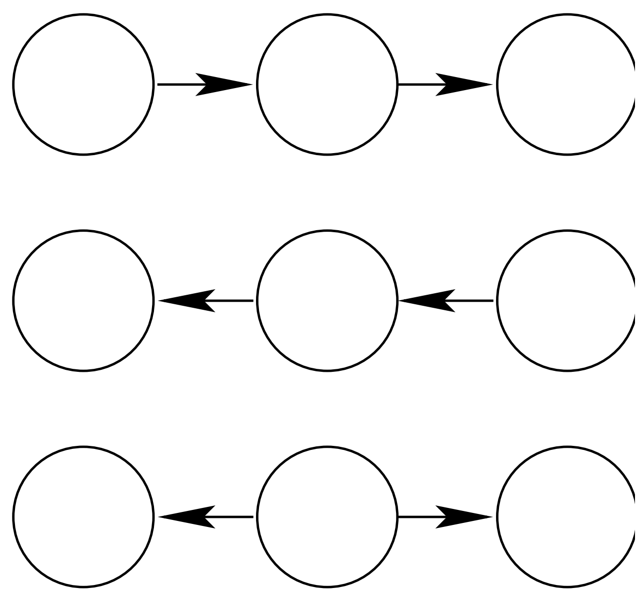
Backward step: information from future observations about the current state

- We want analogous computations for more general graph models

Equivalence of graphs

- We want the inference algorithm to exploit the graph structure (independencies implied by the graph structure)
- Many distinct graphs are equivalent in terms of conditional independencies and therefore can be treated the same in the inference algorithm

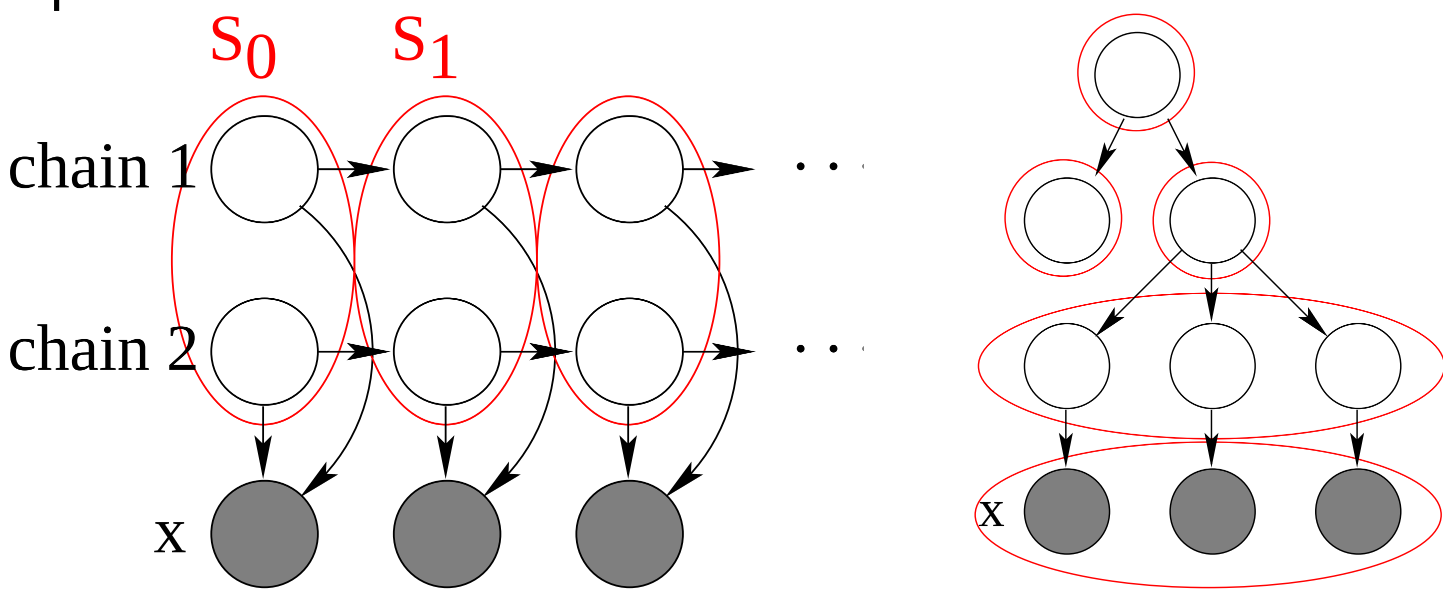
  Example:



Probabilistic inference: example

- It is always possible to cluster the variables (nodes) into larger sets and deal with it as before, just on the level of the sets of variables

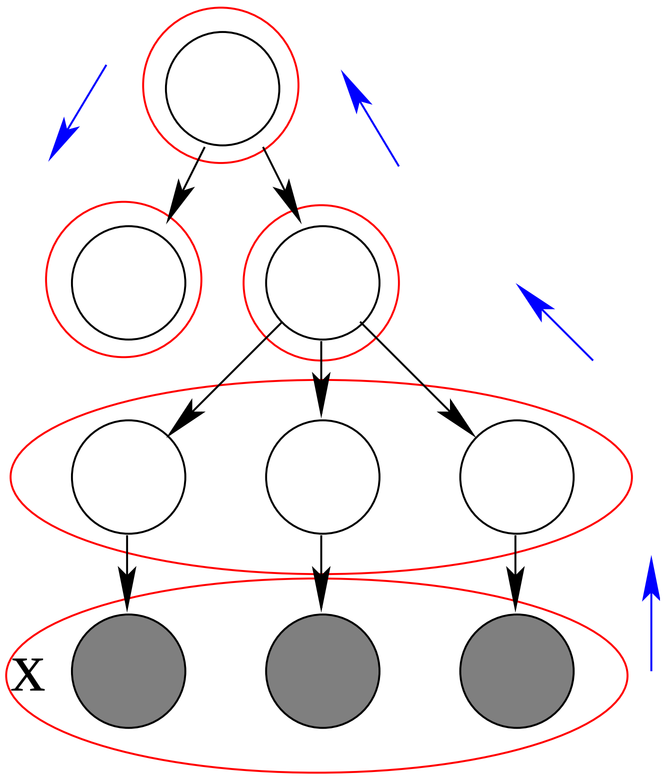
  Examples:



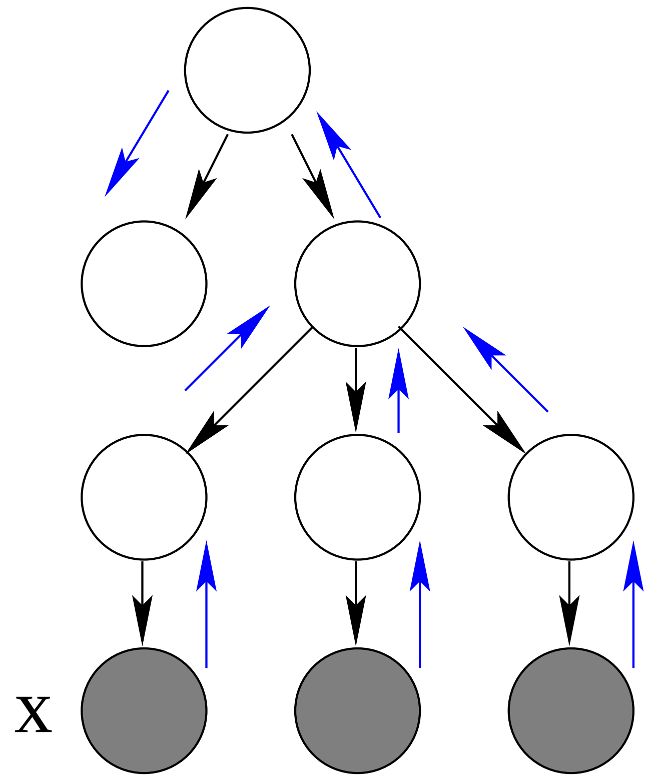
- A chain is not a very efficient structure in this sense

Probabilistic inference

- We can generalize forward-backward to operate on a tree structure rather than a chain



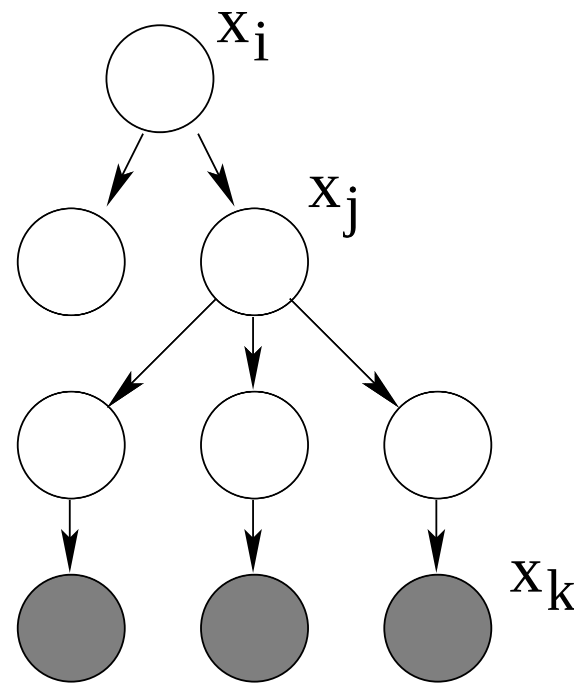
tree as a chain



tree propagation

Belief propagation

- Let's first adopt a common notation for the underlying probability model and the observed data



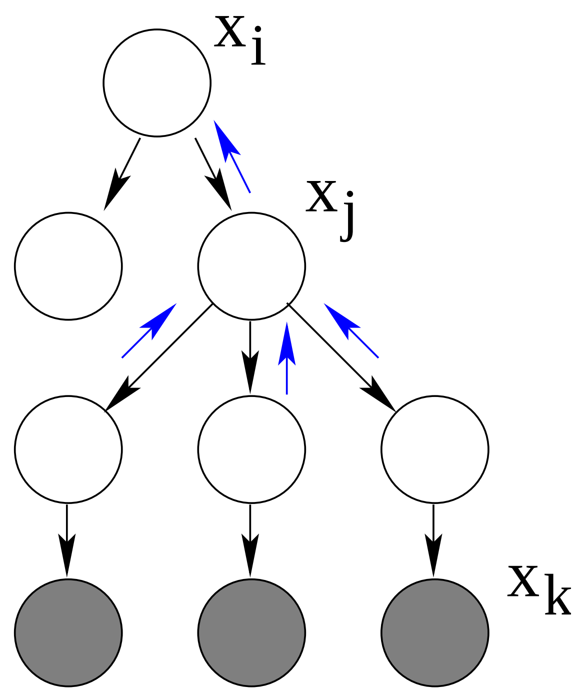
$$\begin{aligned}\psi_{ij}(x_i, x_j) &= P(x_j|x_i) \\ \psi_i(x_i) &= P(x_i) \\ \psi_j(x_j) &= 1 \\ \psi_k(x_k) &= \delta(x_k, \hat{x}_k)\end{aligned}$$

These "potential functions" are chosen to ensure that

$$P(x_1, \ldots, x_n, \text{data}) = \prod_i \psi_i(x_i) \cdot \prod_{(i,j)\in\text{edges}} \psi_{ij}(x_i, x_j)$$

Belief propagation cont'd

- We can now define how each node in the graph should communicate with upstream and downstream nodes (neighbors).



A message passing scheme:

1. Initialize messages to 1.
2. Message (information) that j sends to i is

$$m_{j\to i}(x_i) = \sum_{x_j} \psi_{ij}(x_i, x_j)\psi_j(x_j) \prod_{l \neq i} m_{l\to j}(x_j)$$