6.867 Machine learning and neural networks

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

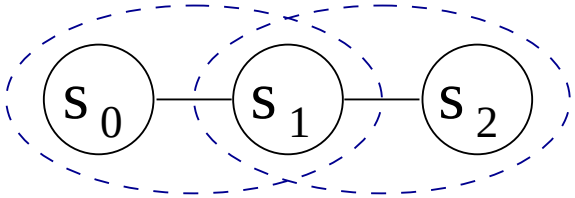
Lecture 24: exact inference cont'd, model selection

Topics

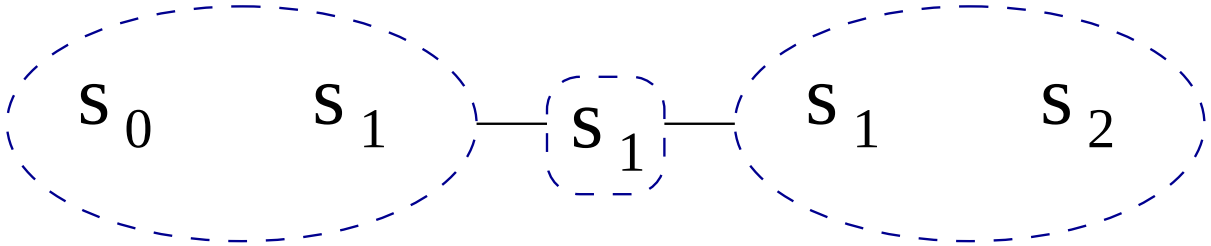
- Exact inference: review
- Model selection
  - basic ideas
  - minimum description length principle

Review: inference with junction trees

- By grouping nodes in the original graph

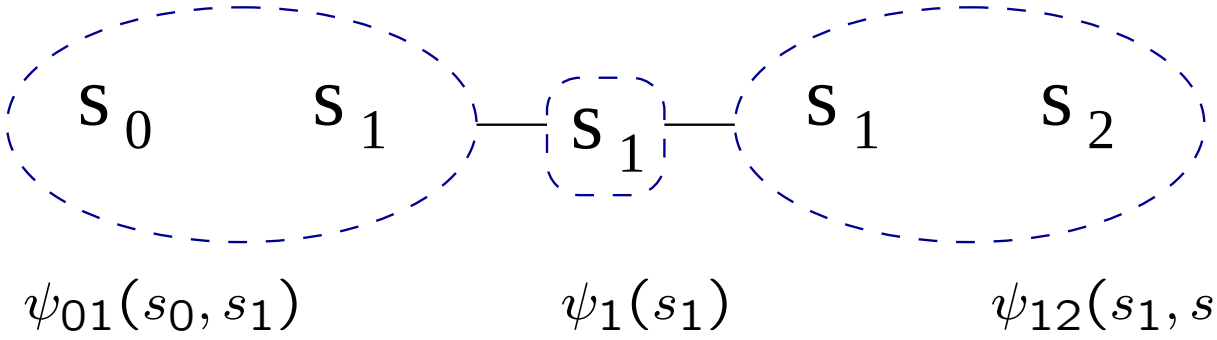


into larger clusters (cliques), we obtain a *junction tree* representation of the associated probability model



- The inference calculations in the junction tree reduce to making sure that the potentials are *consistent*
- Consistency is enforced through the two-stage (collect – distribute) propagation algorithm

Review: consistent junction tree

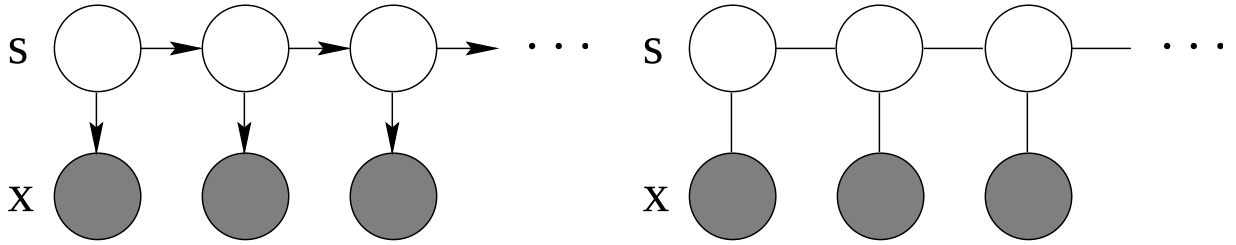


- After both propagation operations, the relevant information is *consistent* and *stored locally*

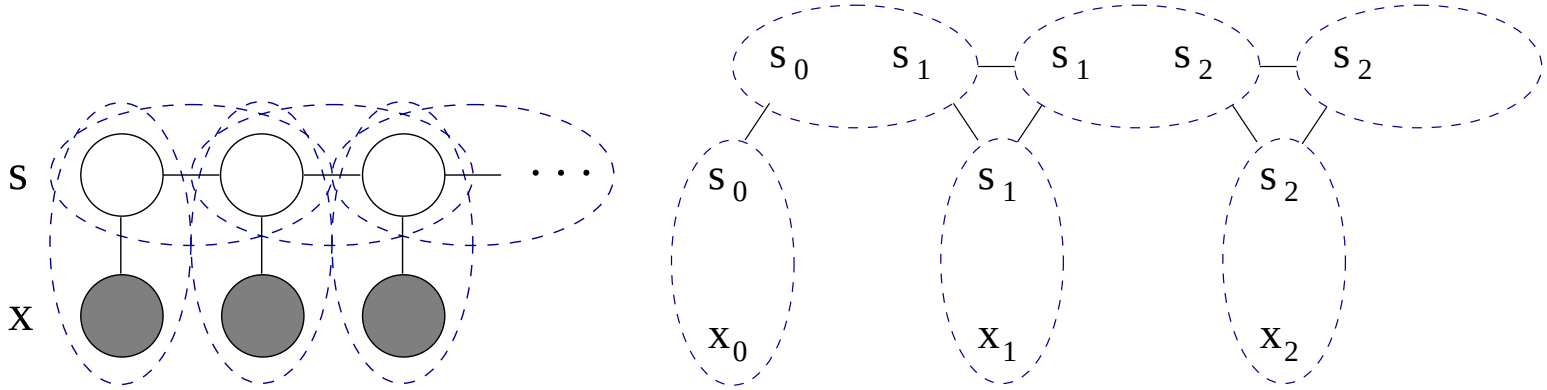
$$\begin{aligned} \psi_{01}(s_0, s_1) &\propto P(s_0, s_1|\text{evidence}) \\ \psi_1(s_1) &\propto P(s_1|\text{evidence}) \\ \psi_{12}(s_1, s_2) &\propto P(s_1, s_2|\text{evidence}) \end{aligned}$$

Example: hidden Markov model

- First we transform the HMM into an undirected graph model

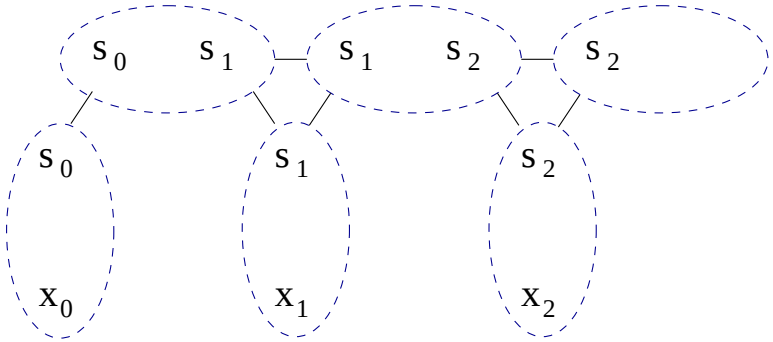


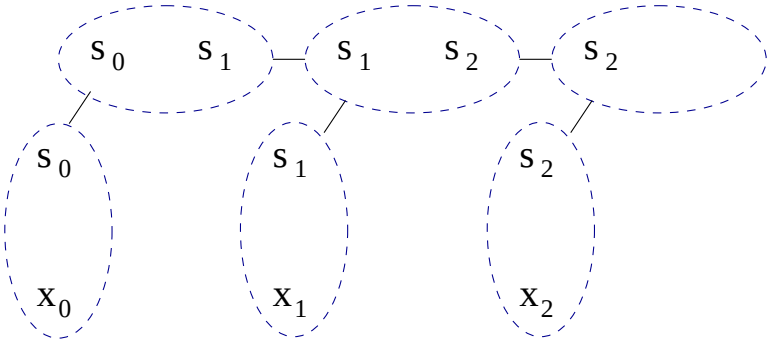
- Then we identify the cliques and construct the hyper-graph by connecting cliques that have variables in common



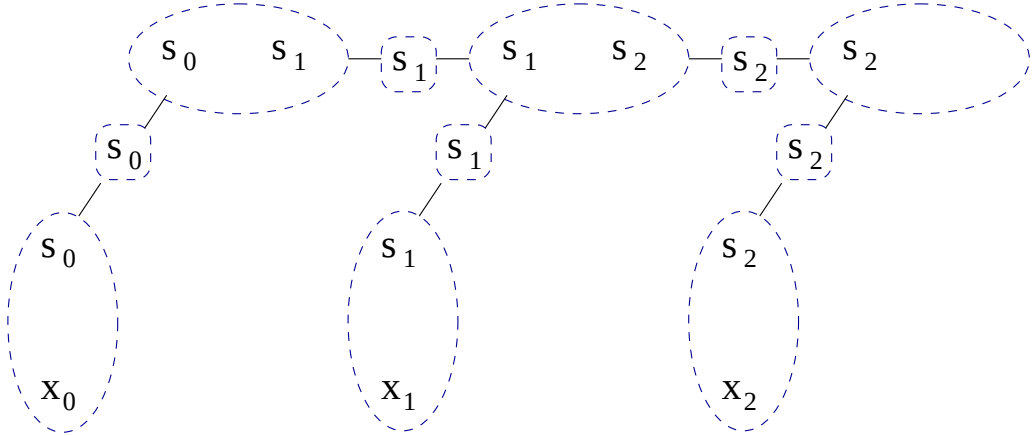
Example cont'd

- The hyper-graph can be transformed into a tree structure by dropping edges so long as any non-adjacent cliques that have variables in common still have these variables in the path between them



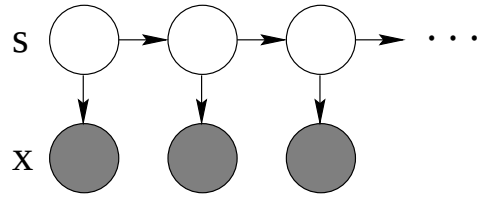


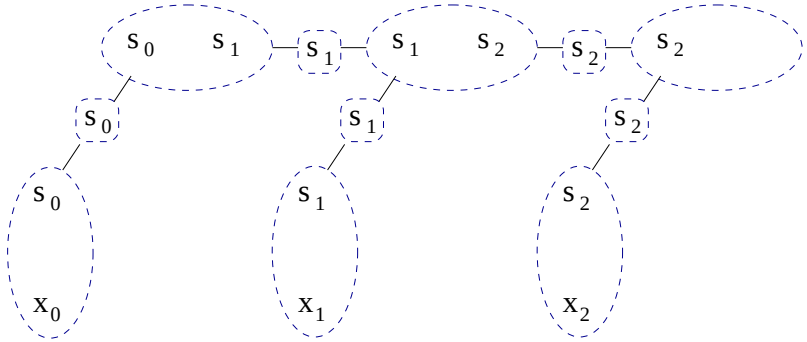
- Finally, we can construct the junction tree for HMMs



Initialization

- How do we initialize the junction tree if we cannot start from a nice consistent solution?





- We can initialize all the separators to unity and assign each conditional probability to a single clique that fully contains the associated variables. E.g.,

$$\begin{aligned} \psi_{01}(s_0, s_1) &= P(s_0)P(s_1|s_0) \\ \psi_1(s_1) &= 1 \\ \psi_{0x}(s_0, x_0) &= P(x_0|s_0) \end{aligned}$$

  and then apply the collect and distribute operations to achieve the "consistent solution"

Topics

- Model selection
  - basic ideas
  - minimum description length principle

Model selection

- We are looking for a general principle that allows us to select the best model from limited observations

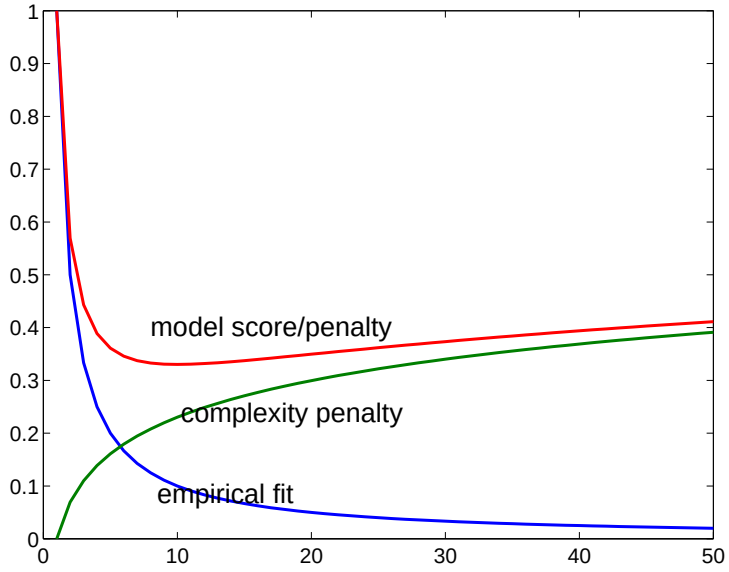
  "Explanation should be as simple as possible, but no simpler"

- What makes one explanation better than another?
  - generality?
  - fewer assumptions?
  - how it is articulated?

    ...
- We need to find quantiable measures for automated comparison

Model selection: balance

- We need a criterion that appropriately balances some measure of model complexity (simplicity) with the empirical fit
  - complexity has to be measured on the same scale as the empirical fit



What is "noise"?

Minimum description length

- Minimum description length (MDL) principle:

  We find a model that attains the minimum total encoding length: the length of the code needed to describe the observed data given the model and the length of the code needed to describe the model itself.

$$\text{Total DL} = \overbrace{(\text{DL of data given the model})}^{\text{empirical fit}} + \overbrace{(\text{DL of model})}^{\text{complexity}}$$

- A long description length (DL) may come from

  a) poor explanation (likelihood) of the observed data
  b) choosing too complex model (too many choices a priori)

Digression: encoding length

- Suppose we have a sequence of n random numbers $y_1, y_2, \ldots$ each drawn with probability $P(y)$

  `1121111314111121...`

- We need $-\log_2 P(y)$ bits to encode each number y

  Higer probability $\Rightarrow$ smaller number of bits

- On average, it would take

$$\begin{aligned} \frac{1}{n}\sum_{t=1}^{n} -\log_2 P(y_t) &= \frac{1}{n}\sum_{y} -N(y)\log_2 P(y) \\ &\to \sum_{y} -P(y)\log_2 P(y) \\ &= H(y) \end{aligned}$$

  bits to encode a single number. Here $H(y)$ is the Shannon entropy (uncertainty) of the random numbers.

Minimum description length cont'd

- Suppose we observe n i.i.d. training examples $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$
- Description length of the observed data given a specific setting of the model parameters $\bar{\theta}$:

$$-\log P(D|\bar{\theta}) = -\sum_{t=1}^{n} \log P(\mathbf{x}_t|\bar{\theta})$$

- Description length of the model parameters

$$-\log P(\bar{\theta})$$

  where $P(\bar{\theta})$ is a prior distribution over the parameters.

- Are these sufficient? How do we set the parameter $\bar{\theta}$?

Minimum description length cont'd

- We have to also decide the precision at which to encode the parameters $\bar{\theta}$ (why would be spare any bits for useless parameters?)
- We have to add a precision cost to the model description length

$$-\log P(\bar{\theta}) - \log(\delta)$$

where the parameters $\bar{\theta}$ are now described upto precision δ.

High precision (small δ) $\Rightarrow$ large precision cost

- The total description length is found by adding all the terms and optimizing the sum with respect to $\bar{\theta}$ and the precision δ

$$\text{Total DL} = \min_{\bar{\theta},\delta} \left\{ -\sum_{t=1}^{n} \log P(\mathbf{x}_t|\bar{\theta}) - \log P(\bar{\theta}) - \log(\delta) \right\}$$

(note that the possible choices of $\bar{\theta}$ depend on the precision δ)

Minimum description length cont'd

- Asymptotically (when the number of observations n is large) the description length reduces to

$$\text{Total DL} \approx -\sum_{t=1}^{n} \log P(\mathbf{x}_t|\widehat{\theta}) + \frac{d}{2}\log(n)$$

  where d is the number of parameters in the model and $\widehat{\theta}$ is the maximum likelihood parameter estimate.

- This still has the right flavor...

Minimum description length: example

- We have two binary variables x_1 and x_2 and two competing models

Model 0: two parameters needed for $P(x_1)P(x_2)$
Model 1: three parameters needed for $P(x_1, x_2)$

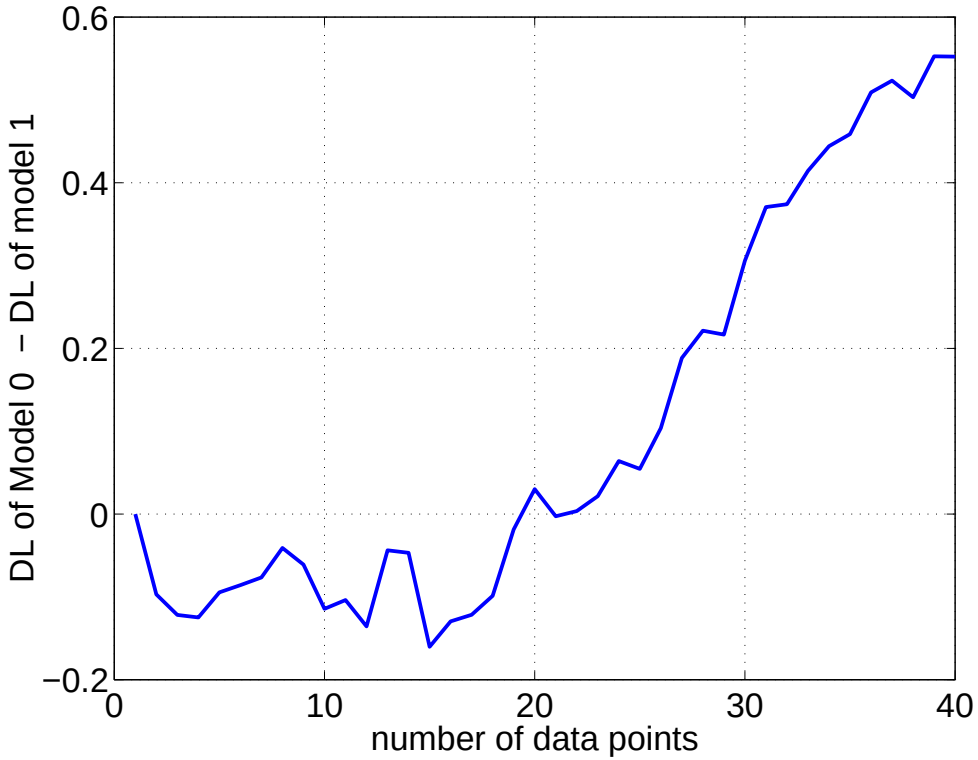
- Observed data:

x_1	x_2
1	0
0	1
1	1
...	...

(here $x_1 = 1$ and $x_2 = 1$ with probability 0.5 and otherwise x_1 and x_2 are selected uniformly at random)

Minimum description length: example

- The difference between the description lengths of the two models as a function of the number of data points:



The figure is averaged over several random generation of the training set

Review

- Main topics:
  - regression, classification
    - linear/additive, Boosting, SVMs
  - generalization, regularization, feature selection
  - complexity, model selection
  - active learning, clustering
  - density models: mixture models, mixtures of experts
    - the EM-algorithm
  - dynamic models: Markov models, hidden Markov models
    - forward-backward, viterbi
  - graphical models: representation, inference