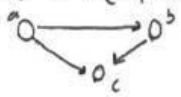
# Graphical Models (PRML, Ch 8)

- Simple way to visualize the structure of a probabilistic model → Design and motivate new models
- Understanding of a model's properties, conditional independencies for example
- Complex computations expressed in terms of graphical manipulation.

- Each node represents a random variable
- Links express probabilistic relationships between the variables
- Bayesian networks, also called DAG's, are graphs where all links have a particular directionality indicated by arrows.
- DAGs are useful for expressing causal relationships
- $P(a,b,c) = P(c|a,b)\,P(b|a)\,P(a)$ gives the following DAG



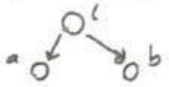

$p(c|a,b)$ = Two incoming links for c

- Since there is a link from a to b, a is the parent of b and b is the child of a.
- A node can represent a single variable as well as a set of variables.
- There must be no directed cycles (follow links and end up at start node not allowed to be possible).
- A graphical model captures the causal process and is therefore often called generative model
- If there are K states and M variables, a fully connected graph (the completely general distribution) will have $K^M - 1$ parameters
- No links in the graph gives the product of the marginals and $M(K-1)$ parameters
- A graph over discrete variables turned into a bayesian model by setting dirichlet priors for the parameters.

- Notation for conditional independence: $a \perp\!\!\!\perp b \mid c$

a is conditionally independent of b given c

- Is equivalent to $p(a|b,c) = p(a|c)$
- General framework for achieving a graphical model that shows conditional independencies is called d-separation
- $p(a,b|c) = p(a|b,c)\,p(b|c) = p(a|c)\,p(b|c)$

gives the graph  The path makes the nodes a and b dependent given the empty set. Given an observed value of c are they independent (conditionally)

- The conditional node then "blocks" the path.
- Another example: $p(a,b,c) = p(a)\,p(b)\,p(c|a,b)$.

In a case where none of the variables are observed $\Rightarrow p(a)\,p(b)$ $\Rightarrow a \perp\!\!\!\perp b \mid \emptyset$. However, if conditioned on the observed value of c they are dependent: $a \not\perp\!\!\!\perp b \mid c$.

- If all paths are blocked, then A is said to be d-separated from B by C.
- A Markov network, or an undirected graphical model, is equivalent to a DAG except that all links are undirected.
- If path from A to B always passes through C, A and B are conditionally independent
- Same general idea as d-separation, but simpler to check for undirected graphs.
- The Markov blanket of a node consists of the set of neighbouring nodes
- A clique is a set of nodes that are fully connected.
- A maximal clique is a clique such that it not is possible to include any other nodes from the graph.
- The Moral graph is the conversion from an directed graph to an undirected.
- If one parent - just remove arrow.

• If more than one parent, the node and its parents must all belong to a single clique.

• The process of moralization adds the fewest needed extra links and retains the maximum number of independence properties.

# Koski & Noble

• Fork connection  $X \perp Y \mid Z$

• Chain connection  $X \perp Y \mid Z$

• Collider connection  $X \perp Y$, $X \not\perp Y \mid Z$

• All connections between nodes in a DAG are of one of these types.

• The Markov blanket of a node in a DAG is the parents, the children and the parents of the children.

• Aalborg algorithm (by L.S)

1. DAG is moralized
   - The Moral graph is the graph where each variable-parent set in the original DAG is a clique.
2. Moral graph is triangulated
3. Cliques are organized into a junction tree
   - Can only be done if decomposable, which it is if triangulated
   - Junction tree is a tree with all the cliques
   - The nodes present in two adjacent cliques are called separators.

• Evidence is the information that certain nodes take specific values.

• Parameter learning: Using proportions from observed data, use dirichlet prior and update with bayes rule.

# Hidden Markov Models (PRML, Ch. 13)

- Sequential data and stationary distributions. Stationarity means that underlying distribution for the sequence remains the same over the whole time period.
- HMM useful when there is a need of relaxing the iid assumption.
- In a simple HMM is the next observation independent of all earlier observations except the preceding one.
- Higher-order Markov Chains allow more than the preceding observation to have impact on the prediction.
- In a second-order Markov chain is information from the two preceding observations used.
- Higher-order models increases the flexibility but also the complexity in terms of number of parameters.
- If the variables are discrete, the number of parameters in the model will be $K^{M-1}(K-1)$ ($M$ = order, $K$ = states)
- To get past this problem is a latent variable, $z_n$, introduced for each observation.
- There is then always a path open between two observed values, so all previous observations can be used for making predictions (unless $z$ is observed, which it not is)
- If latent variables are discrete is a HMM obtained.
- The observed values can be discrete or continuous.
- If latent variables not are discrete → State space model
- The value of $z_n$ depends on $z_{n-1}$ and the probability for each state is specified in the transition matrix.
- The combination probabilities for $x_n$ is given by the emission matrix
- When sampling: Init value for $z_1$ => Sample $x_1$ => Sample $z_2$ using transition matrix and $z_1$.
- Forward-Backward algorithm

- $\alpha$ uses data up to time $n$
- $\beta$ uses data from $n+1$ to $N$
- $\alpha$ works forward and $\beta$ backwards

- If latent variables do have some meaningful interpretation is often the most probable sequence of hidden states for a given observation sequence of interest to find.
- The problem of finding the most probable sequence of latent states is not the same as that of finding the the set of states that individually are the most probable.
- The latter might give sequences with very low, or zero, probability.
- The Viterbi algorithm is used for finding the most probable sequence/path.
- The algorithm only keeps the path that at each time step is the most probable. Then, at next step, selects the state which gives the, so far, most probable path and keeps only that sequence.

# Ghahramani

- Markov property: Given the state $S_{t-1}$ is $S_t$ independent of all states prior to $t-1$
- Observations was generated by some process whose state is hidden from the observer.
- The properties above defines the HMM.
- HMM is a dynamic bayesian network, which is a bayesian network that works for modeling time series data.