Probabilistic graphical models
Directed (BNs) and undirected (MRFs) graphs

22-808: Generative models

Sharif University of Technology

Fall 2024

Fatemeh Seyyedsalehi

Probabilistic graphical models

- A framework to tackle with complex joint distributions
  - Representation
    - Directed graphs: Bayesian network
    - Undirected graphs: Markov random fields
  - Learning
  - Inference

- This lecture
  - Representation in PGMs

Probabilistic graphical models

- Searching in the fully generalized space of distributions even in a simple probabilistic problem is impossible!

- Learn an effective and general technique for parameterizing probability distributions using only a few parameters.

Probabilistic graphical models

- Independencies assumptions are useful
  - Simplify representation and alleviate inference complexities

- Enable us to incorporate domain knowledge and structures
  - Modular combination of heterogeneous parts
  - Combining data and knowledge (Bayesian philosophy)

Bayesian networks

- Directed graphical models are tools to present family of probability distributions that can be naturally described using a directed acyclic graph.
  - Nodes as random variables
  - Edges as dependencies

- The intention behind these parameterization is chain rule!

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2 \mid x_1) \cdots p(x_n \mid x_{n-1}, \ldots, x_2, x_1)$$

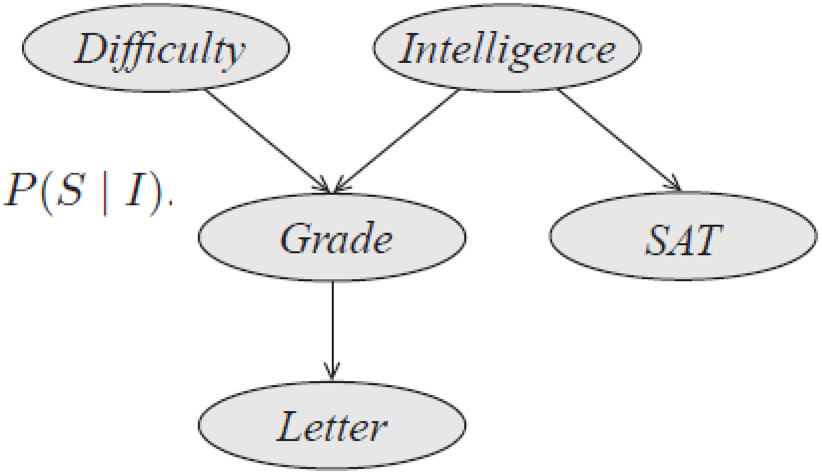
Bayesian networks

- Bayesian networks represent a joint distribution in terms of **the graph structure** and **conditional probability distributions (CPD)**

$G = (V, E)$

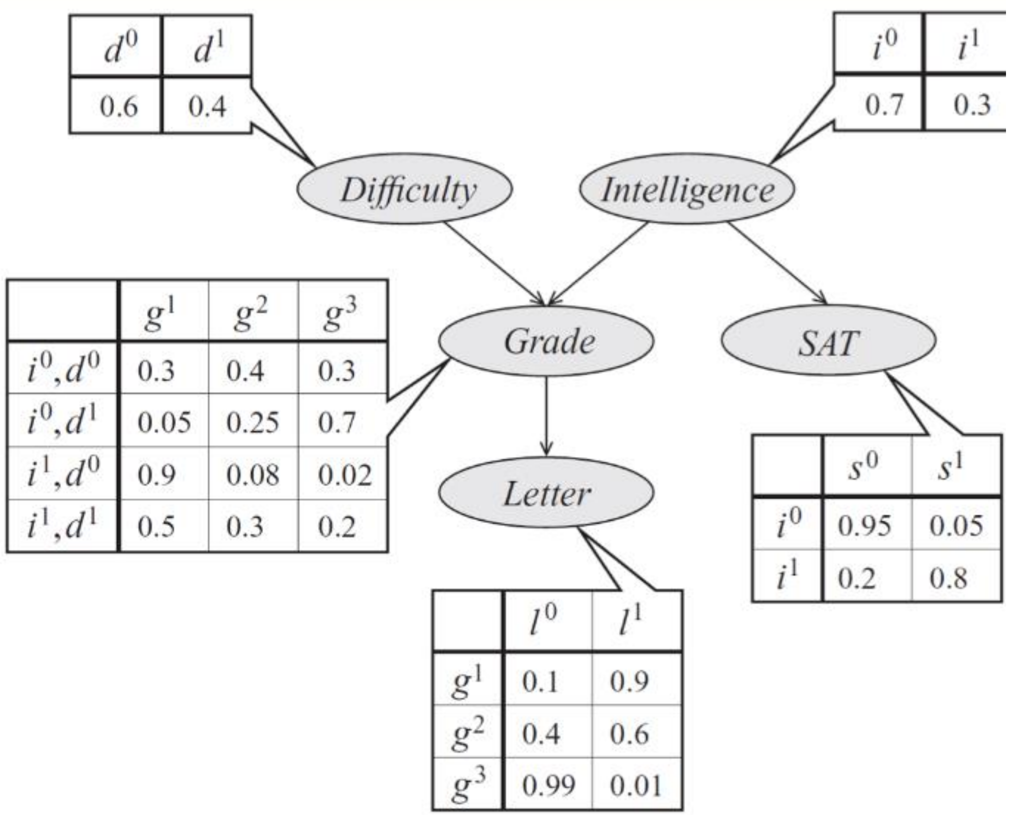
- A random variable x_i for each node $i \in V$.
- One conditional probability distribution (CPD) $p(x_i \mid x_{A_i})$ per node, specifying the probability of x_i conditioned on its parents' values.

$$P(I, D, G, L, S) = P(I)P(D)P(G \mid I, D)P(L \mid G)P(S \mid I).$$



Bayesian networks
Discrete example

- When the variables are discrete, we may think of the factors (CPDs) as *probability tables*, in which rows correspond to assignments to parents and columns correspond to values of the node.



d^0	d^1
0.6	0.4

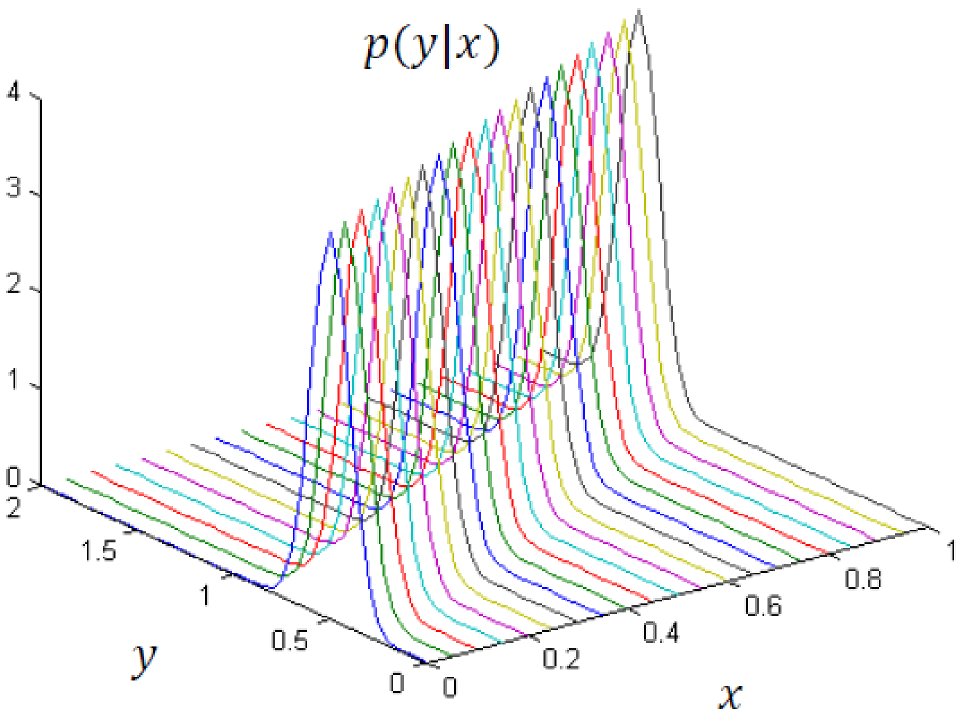
i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

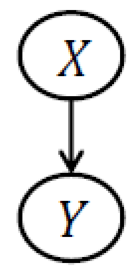
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Bayesian networks
Continues example



$b = 0.5$
$\sigma = 0.1$

$X \sim N(0,1)$



$Y|X \sim N(b + X, \sigma)$

Bayesian networks

- A probability distribution is factorized over a *DAG* G if it can be decomposed into a product of factors specified by G.

- A Bayesian network represent distributions via products of smaller, local conditional probability distributions.
  - Introduces independency assumptions over variables

- $I(p)$: denote the set of all independencies that hold for a joint distribution p.
  - $p(x,y) = p(x)p(y) \rightarrow x \perp y \in I(p)$

Bayesian networks

- Let G be a graph over $x_1, x_2, \dots, x_n$ distribution p factorizes over G if:

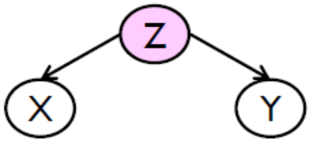
$$p(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} p(x_i | pa(x_i))$$

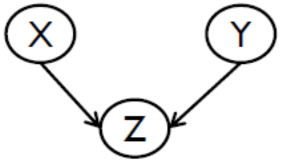
- $pa(.)$: parents of a node

- Factorization $\Leftrightarrow$ Independence
  - If p factorizes over G, then any variable in p is independent of its non-descendants given its parents (in G)
  - If any variable in the distribution p is independent of its non-descendants given its parents (in the graph G) then p factorizes over G

Independencies described by directed graphs

- *Common parent*. If G is of the form $X \leftarrow Z \rightarrow Y$, and Z is observed, then $X \perp Y \mid Z$. However, if Z is unobserved, then $X \not\perp Y$. Intuitively this stems from the fact that Z contains all the information that determines the outcomes of X and Y; once it is observed, there is nothing else that affects these variables' outcomes.
- *Cascade*: If G equals $X \rightarrow Z \rightarrow Y$, and Z is again observed, then, again $X \perp Y \mid Z$. However, if Z is unobserved, then $X \not\perp Y$. Here, the intuition is again that Z holds all the information that determines the outcome of Y; thus, it does not matter what value X takes.
- *V-structure* (also known as *explaining away*): If G is $X \rightarrow Z \leftarrow Y$, then knowing Z couples X and Y. In other words, $X \perp Y$ if Z is unobserved, but $X \not\perp Y \mid Z$ if Z is observed.





Independencies described by directed graphs
D-separation

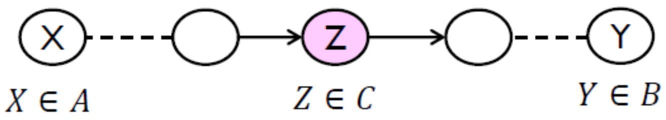
- Considering three disjoint sets of nodes:
  - *A, B, C*

- A is **d-separated** from B by C if all paths between A and B are blocked by C
  - There is no **active path** between A and B

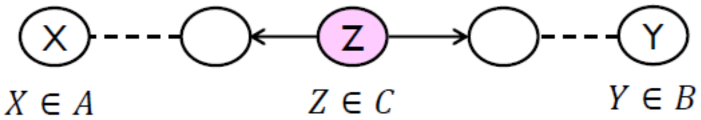
- A is d-separated from B by C if $\boldsymbol{A} \perp \boldsymbol{B} | \boldsymbol{C}$

Path blocking

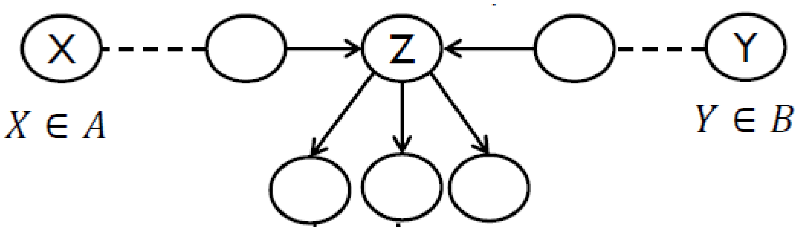
- Head to tail during path



- Tail to tail during path

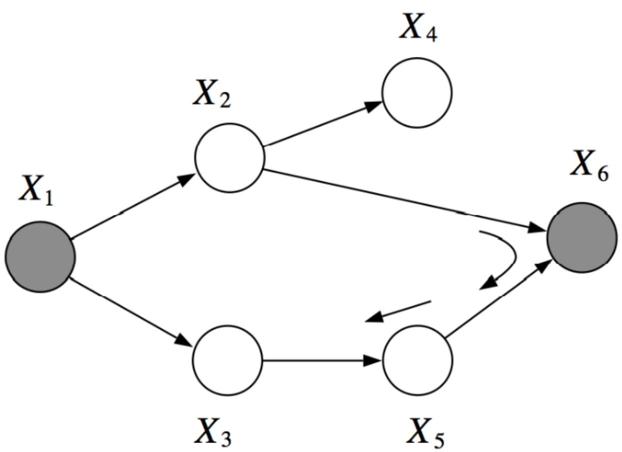


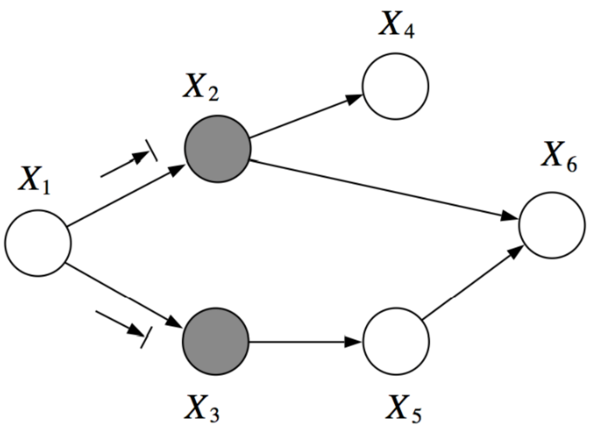
- Head to head, visiting a v-structure
  - Z and none of its descendants are observed



Independencies described by directed graphs

For example, in the graph below, X_1 and X_6 are d-separated given X_2, X_3. However, X_2, X_3 are not d-separated given X_1, X_6, because we can find an active path (X_2, X_6, X_5, X_3)





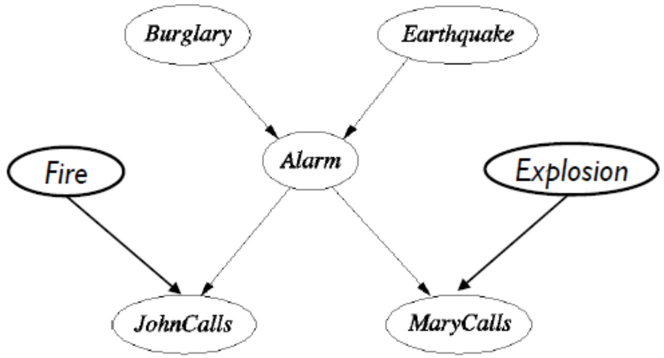
A simple d-separation simulator ☺

Markov blanket of a node

- A variable is conditionally independent of all other variables given its **Markov blanket**

- Markov blanket if a set *A* is *U* when:
  - The minimal set of nodes such that *A* is independent from the rest of the graph if *U* is observed

- Markov blanket of a node:
  - All parents
  - All children
  - Co-parents of children



Independencies described by directed graphs

- If p factorizes over G, then $\mathrm{I}(\mathrm{G}) \subseteq I(p)$. In this case, we say that G is an **I-map** (independence map) for p.
  - All independencies encoded in G are valid in p
  - However, the converse is not true:
    - a distribution may factorize over G, yet have independencies that are not captured in G.

The representation power of BNs

- Can we show all independencies in a distribution p with a DAG?
  - A **perfect map**: $\mathrm{I(G)} = I(p)$?
  - Not for every distribution exist a perfect map

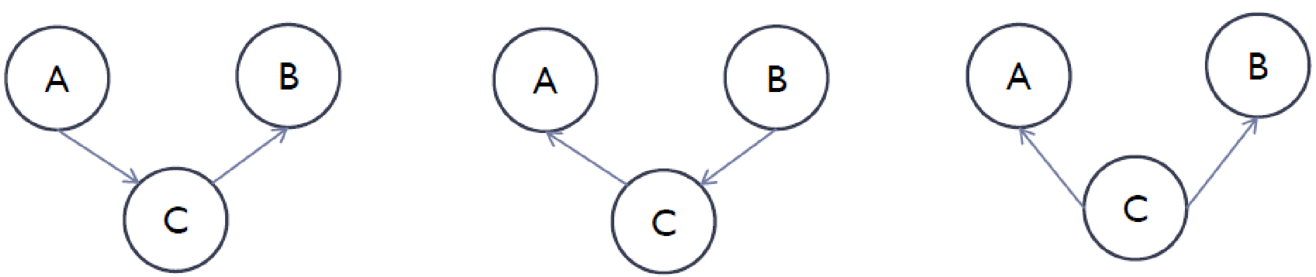
- It is easy to reach $\mathrm{I(G)} \subseteq I(p)$
  - In a complete graph: $|\mathrm{I(G)}| = 0$

- A **minimal I-map** G for p: an I-map such that the removal of even a single edge from G will result in it no longer being an I-map.

The representation power of BNs

- **I-equivalence**: When two graphs G_1 and G_2 encode a same set of dependencies: $\mathrm{I}(\mathrm{G}_1) = I(G_2)$

- **Fact:** If G and G' have the same skeleton and the same v-structures, then $\mathrm{I}(G) = I(G')$



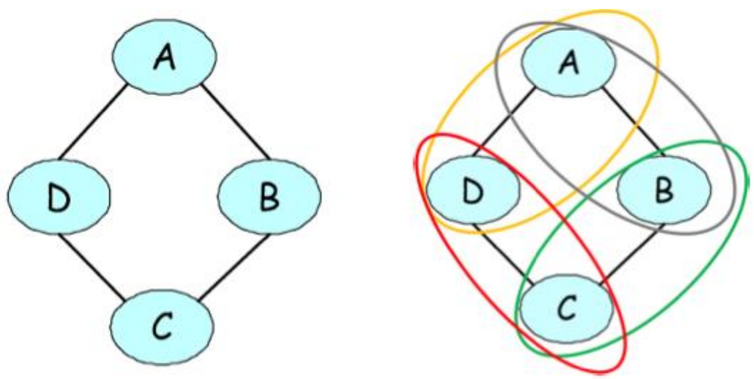
Markov random networks

- Undirected graphs for representation of joint distributions
  - Unlike in the directed case, we are not saying anything about how one variable is generated from another set of variables (as a conditional probability distribution would do).

$$\tilde{p}(A,B,C,D) = \phi(A,B)\phi(B,C)\phi(C,D)\phi(D,A)$$

$$\phi(X,Y) = \begin{cases} 10 & \text{if } X = Y = 1 \\ 5 & \text{if } X = Y = 0 \\ 1 & \text{otherwise.} \end{cases}$$

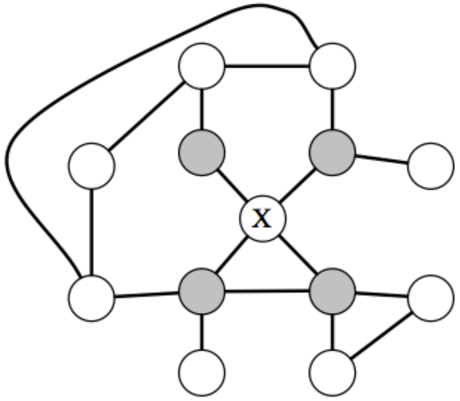
$$p(A,B,C,D) = \frac{1}{Z}\tilde{p}(A,B,C,D)$$



Markov random networks

- They specify dependent variables (but no causality relations) and define the strength of their interactions.

- This defines an energy landscape over the space of possible assignments and we convert this energy to a probability via the normalization constant.

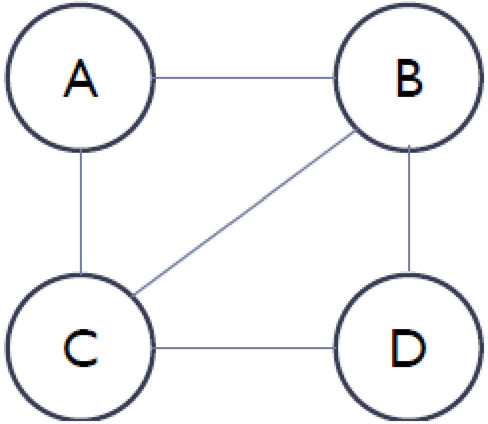


MRF factorization

- **Clique**: subsets of nodes in the graph that are fully connected (complete subgraph)
- **Maximal clique**: no superset of the nodes in a clique are also compose a clique
- Factors are functions of the variables in cliques
  - To reduce the number of factors we allow factors only for maximal cliques

Cliques: {A,B,C}, {B,C,D}, {A,B}, {A,C}, {B,C}, {B,D}, {C,D}, {A}, {B}, {C}, {D}

Max-cliques: {A,B,C}, {B,C,D}



MRF factorization

- A distribution $p(.)$ is factorized over an MRF G if it can be parameterized as follows,

$$p(x_1, x_2, \dots, x_n) = \frac{1}{Z} \prod_{i=1}^{k} \phi_i(D_i)$$

$$Z = \sum_{X} \prod_{i=1}^{k} \phi_i(D_i)$$

where each D_i is a **complete subgraph** of G

- When there is no direct edge between two nodes, x_i and x_j, there exist at least the following conditional independency between them:

$$x_i \perp x_j | X/\{x_i, x_j\}$$

  - To hold this independency in $p(.)$, these two variables are not appeared in the domain of a same factor

MRF factorization

- Potential functions:
  - The function over each clique (factor)
- Potential functions and cliques in the graph completely determine the joint distribution.
- Potentials are not necessarily marginal or conditional distributions

Markov random networks

- Formal definition

A Markov Random Field (MRF) is a probability distribution p over variables $x_1, \ldots, x_n$ defined by an *undirected* graph G in which nodes correspond to variables x_i. The probability p has the form

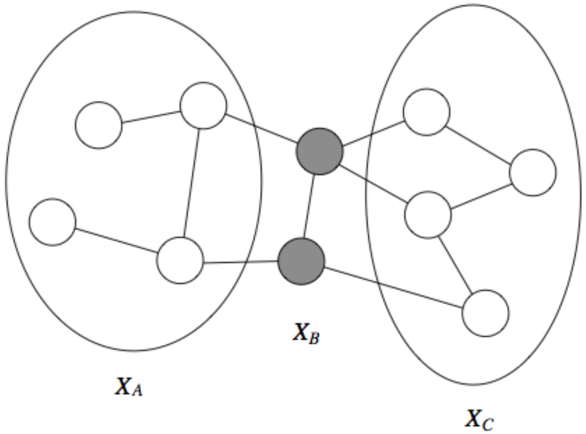
$$p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c \in C} \phi_c(x_c),$$

where C denotes the set of *cliques* (i.e., fully connected subgraphs) of G, and each *factor* ϕ_c is a non-negative function over the variables in a clique. The *partition function*

$$Z = \sum_{x_1, \ldots, x_n} \prod_{c \in C} \phi_c(x_c)$$

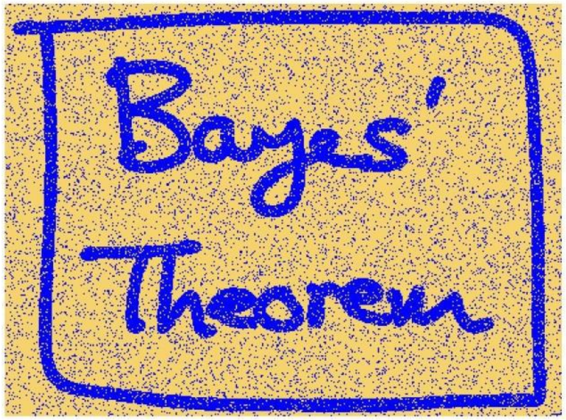
Independencies in MRFs

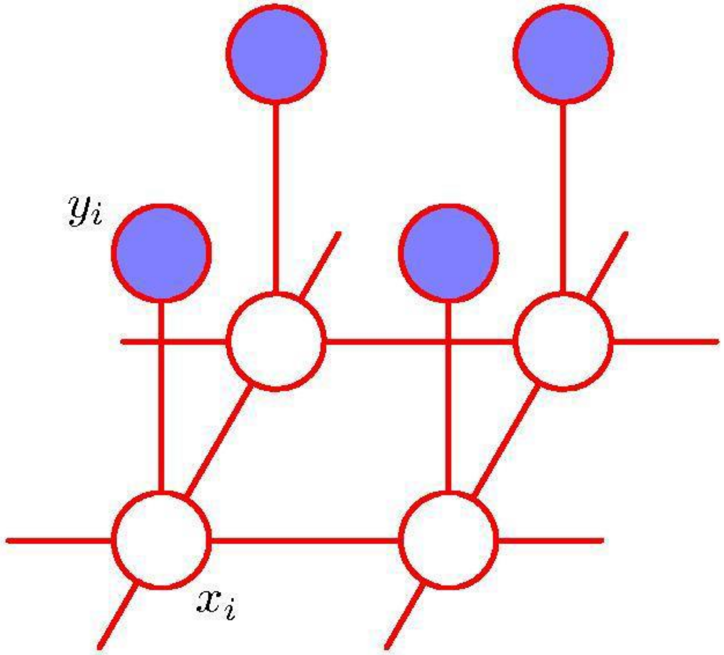
- A simple rule:
  - Variables x and y are dependent if they are connected by a path of unobserved variables.
- Markov blanket in MRFs:
  - In both BNs and MRFs
  - In MRFs: simply all neighbors of a node



MRF example: Image denoising

- Pixels are noisy observed variables: y_i
- We assume the noise free image as a latent behind the observed pixels: x_i





MRFs compared to BNs

- Pros.
  - They can be applied to a wider range of problems in which there is no natural directionality associated with variable dependencies.
  - Undirected graphs can succinctly express certain dependencies that Bayesian nets cannot easily describe (although the converse is also true)

- Cons.
  - Computing the normalization constant Z requires summing over a potentially exponential number of assignments.
    - NP-hard; thus many undirected models will be intractable and will require approximation techniques.
  - Difficult to interpret.
  - It is much easier to generate data from a Bayesian network

Hybrid graphs

- Partially directed acyclic graphs
  - A combination of both directed and undirected graphs

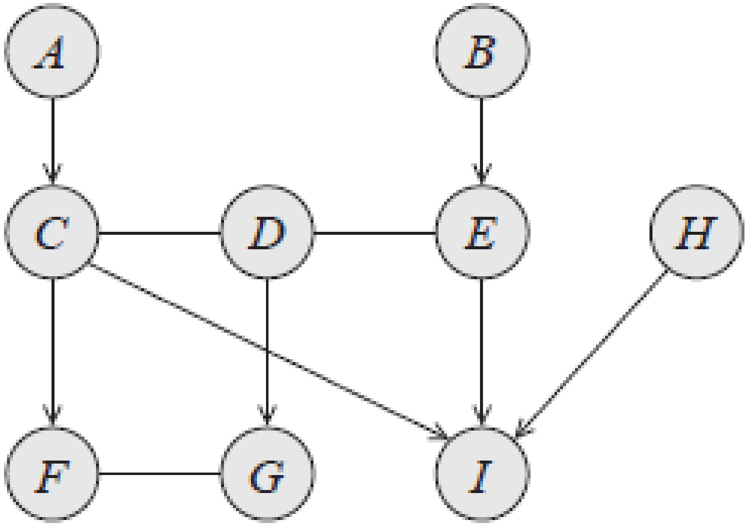
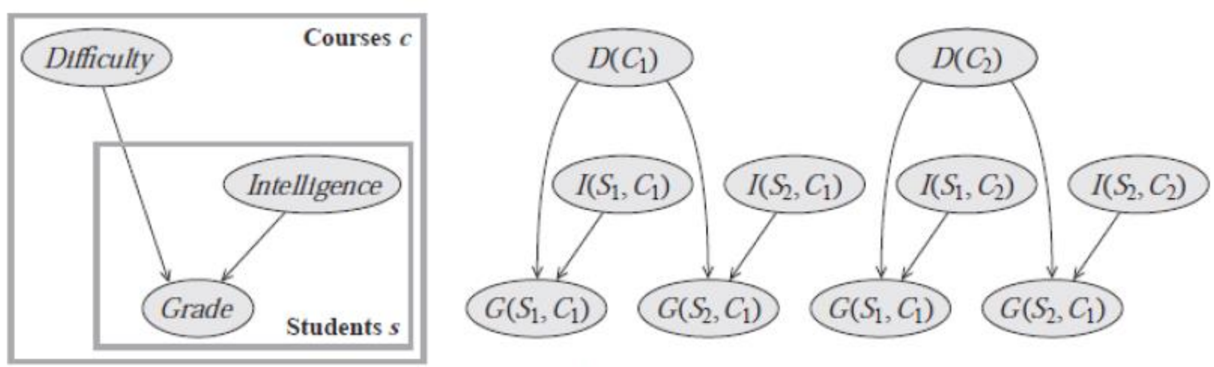


Plate notation

- Plate notation is a rectangle in graphical model representation which shows random variables generated from the same distribution
- Plate notation present a replication of random variables that share same parameters

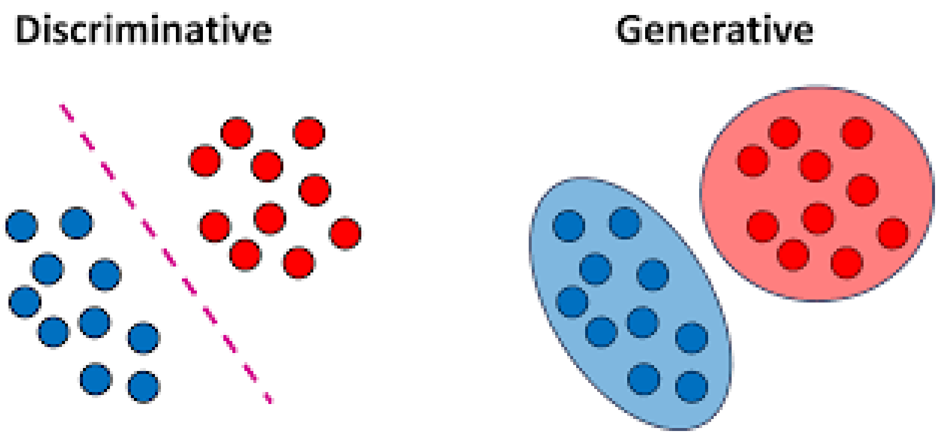
(a)



(b)

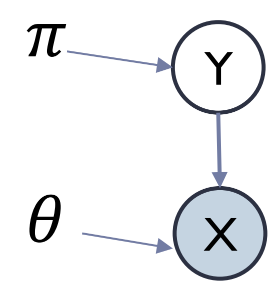
Generative vs. discriminative models

- In generative models we describe the generation process of observed variables
- In discriminative models, we learn how samples are discriminated
  - Decision boundaries in classifiers



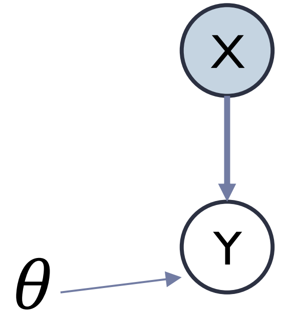
Generative vs. discriminative models
Example

- Generative classifier
  - We should learn $p(y),\ p(x|y)$



- Discriminative classifier
  - We should learn $p(x), p(y|x)$
  - However, for classification task

  $p(y|x)$ is the only thing we need.
    - Less parameters are needed to be learned

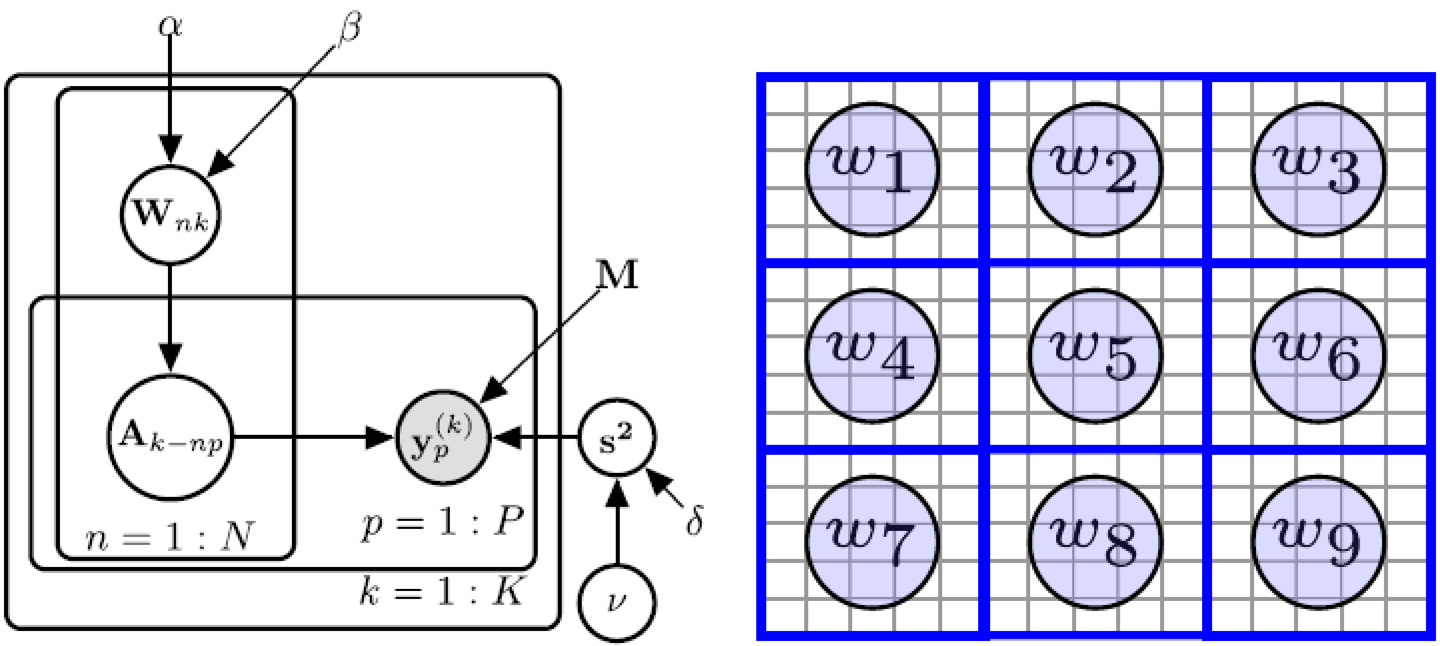


- When we only need to discriminate

Between samples, discriminative models are preferred.

Generative PGM example
Hyperspectral unmixing with PGMs

- A generative model
  - K = number of patches
  - P = number of pixels in each patch
  - N = the dimension of vector A

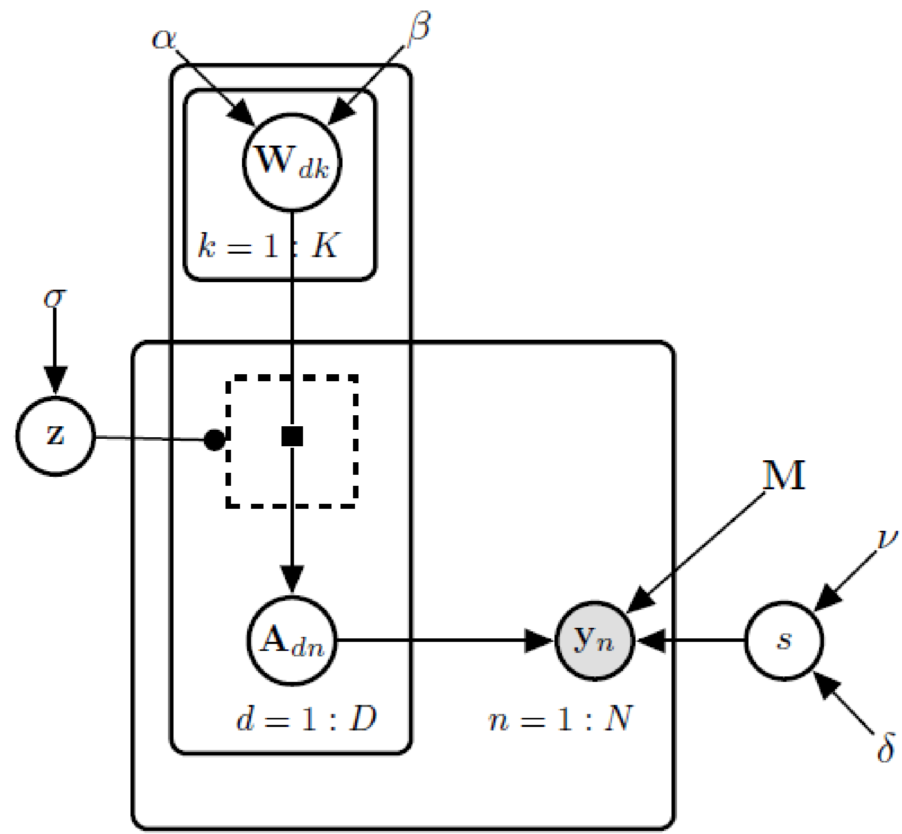


(a) (b)

Generative PGM example
Hyperspectral unmixing with PGMs

- A generative model
- A partially directed model
  - With a plate notation

  and gate structure



Next topic

- Probabilistic graphical models
  - Exact and approximate inference