Graphical Models For Complex Health Data (P8124)

Daniel Malinsky

Columbia University
dsm2128@cumc.columbia.edu

Undirected Graphical Models (+ others)

Undirected Graphs

$\mathcal{G} = (V, E)$ is an undirected graph (UG), that is a graph such that E contains only undirected edges ($-$).

Factorization Property

A *clique* C is a set of vertices s.t. every pair in C is adjacent (connected by an edge). We use $\mathcal{C}$ for the set of cliques in a graph $\mathcal{G}$.

A distribution $p(x)$ satisfies the *factorization property* wrt UG $\mathcal{G}$ if

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \phi_C(x_C)$$

where each $\phi_C(x_C)$ is a non-negative function that depends only on the variables in C. These functions are called *factors*. Z is a normalizing constant called the *partition function*: $Z \equiv \sum_x \prod_{C \in \mathcal{C}} \phi_C(x_C)$.

Factorization Property

A *clique* C is a set of vertices s.t. every pair in C is adjacent (connected by an edge). We use $\mathcal{C}$ for the set of cliques in a graph $\mathcal{G}$.

A distribution $p(x)$ satisfies the *factorization property* wrt UG $\mathcal{G}$ if

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \phi_C(x_C)$$

where each $\phi_C(x_C)$ is a non-negative function that depends only on the variables in C. These functions are called *factors*. Z is a normalizing constant called the *partition function*: $Z \equiv \sum_x \prod_{C \in \mathcal{C}} \phi_C(x_C)$.

Note: A clique is *maximal* if for any superset of nodes $D \supset C$, D is not a clique. One may equivalently rephrase this factorization property in terms of maximal cliques.

Definition: Markov random field (a.k.a. Markov network model)

A pair $(\mathcal{G}, \mathcal{P})$ where $\mathcal{G}$ is an UG and $\mathcal{P}$ is a set of distributions that factorize wrt $\mathcal{G}$.

Example

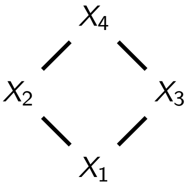
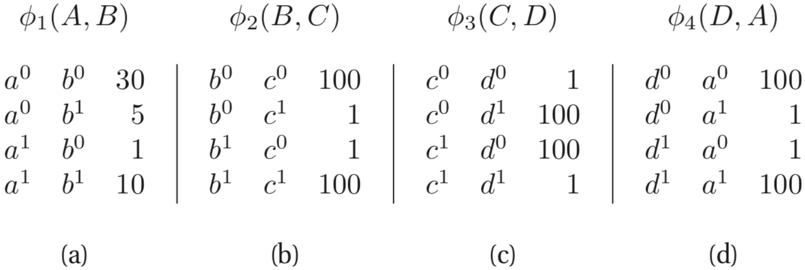


Figure: An undirected graph $\mathcal{G}$

$$\begin{aligned} p(x) &= \frac{1}{Z} \prod_{C \in \mathcal{C}} \phi_C(x_C) \\ &= \frac{1}{Z} \phi_{12}(x_1, x_2)\phi_{23}(x_2, x_3)\phi_{34}(x_3, x_4)\phi_{14}(x_1, x_4) \end{aligned}$$

where $Z = \sum_{x_1, x_2, x_3, x_4} \phi_{12}(x_1, x_2)\phi_{23}(x_2, x_3)\phi_{34}(x_3, x_4)\phi_{14}(x_1, x_4)$

Example

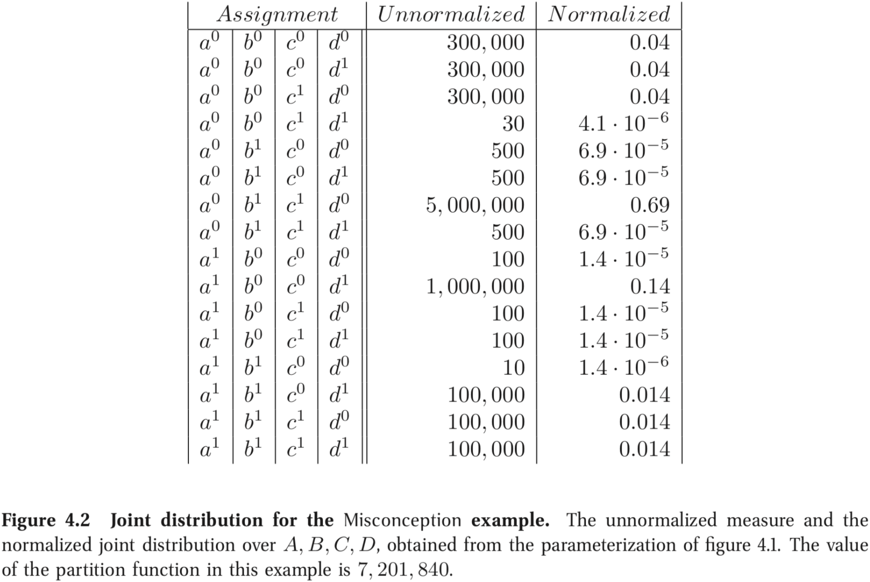


$\phi_1(A, B)$			$\phi_2(B, C)$			$\phi_3(C, D)$			$\phi_4(D, A)$		
a^0	b^0	30	b^0	c^0	100	c^0	d^0	1	d^0	a^0	100
a^0	b^1	5	b^0	c^1	1	c^0	d^1	100	d^0	a^1	1
a^1	b^0	1	b^1	c^0	1	c^1	d^0	100	d^1	a^0	1
a^1	b^1	10	b^1	c^1	100	c^1	d^1	1	d^1	a^1	100
(a)			(b)			(c)			(d)		

Figure 4.1 Factors for the Misconception **example**

This is a parameterization of the 4-variable binary MRF, from Koller & Friedman (2009, pp 83, 104).
They use A, B, C, D for X_4, X_3, X_1, X_2.

Example



Assignment				*Unnormalized*	*Normalized*
a^0	b^0	c^0	d^0	$300,000$	0.04
a^0	b^0	c^0	d^1	$300,000$	0.04
a^0	b^0	c^1	d^0	$300,000$	0.04
a^0	b^0	c^1	d^1	30	$4.1 \cdot 10^{-6}$
a^0	b^1	c^0	d^0	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^0	d^1	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^1	d^0	$5,000,000$	0.69
a^0	b^1	c^1	d^1	500	$6.9 \cdot 10^{-5}$
a^1	b^0	c^0	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^0	d^1	$1,000,000$	0.14
a^1	b^0	c^1	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^1	d^1	100	$1.4 \cdot 10^{-5}$
a^1	b^1	c^0	d^0	10	$1.4 \cdot 10^{-6}$
a^1	b^1	c^0	d^1	$100,000$	0.014
a^1	b^1	c^1	d^0	$100,000$	0.014
a^1	b^1	c^1	d^1	$100,000$	0.014

Figure 4.2 Joint distribution for the Misconception **example.** The unnormalized measure and the normalized joint distribution over A, B, C, D, obtained from the parameterization of figure 4.1. The value of the partition function in this example is $7,201,840$.

Example

Important to remember: $\phi_{12}(x_1, x_2)$ does not necessarily correspond to the marginal distribution $p(x_1, x_2)$ or to any conditional distribution. In fact it is not necessarily a distribution at all, hence the normalizing factor Z.

Example

Important to remember: $\phi_{12}(x_1, x_2)$ does not necessarily correspond to the marginal distribution $p(x_1, x_2)$ or to any conditional distribution. In fact it is not necessarily a distribution at all, hence the normalizing factor Z.

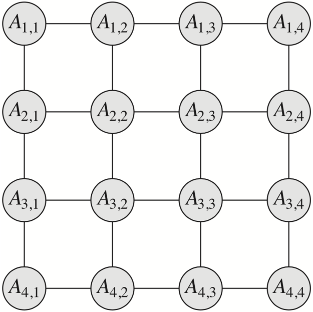
One common way to use factors or potential functions is to encode "degree of similarity" or "propensity to agree" which has roots in physics applications (lattice structures, ferromagnetism) and has become very useful in computer vision, where nodes are neighboring pixels.

Pairwise MRF

A common parameterization involves a factor for

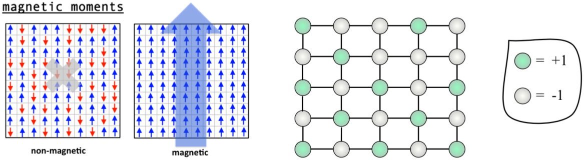
- every edge in the graph (edge potential) and
- every singleton node (node potential)
- $p(x) \propto \Pi_{i \sim j} \phi_{ij}(x_i, x_j) \Pi_i \phi_i(x_i)$

this is called a *pairwise* MRF. Often the network takes the form of a grid.



Pairwise MRF example: Ising model

- Invented by the physicist Wilhelm Lenz (1920), who gave it as a problem to his student Ernst Ising
- Mathematical model of ferromagnetism in statistical mechanics
- The spin of an atom is biased by the spins of atoms nearby on the material:



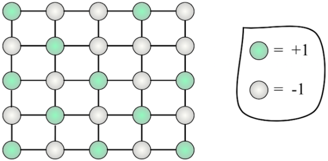
- Each atom $X_i \in \{-1, +1\}$, whose value is the direction of the atom spin
- If a spin at position i is $+1$, what is the probability that the spin at position j is also $+1$?
- Are there phase transitions where spins go from "disorder" to "order"?

(slide taken from David Sontag)

Pairwise MRF example: Ising model

Each atom $X_i \in \{-1, +1\}$, whose value is the direction of the atom spin

The spin of an atom is biased by the spins of atoms nearby on the material:



$$p(x_1, ..., x_p) = \frac{1}{Z} \exp\left(\sum_{i \sim j} w_{ij} x_i x_j - \sum_i w_i x_i \right)$$

When $w_{ij} > 0$, nearby atoms encouraged to have the same spin (called **ferromagnetic**), whereas $w_{ij} < 0$ encourages $X_i \neq X_j$

Node potentials $\exp(-w_i x_i)$ encode the bias of the individual atoms

(slide taken from David Sontag)

Pairwise MRF application: computer vision (image segementation)

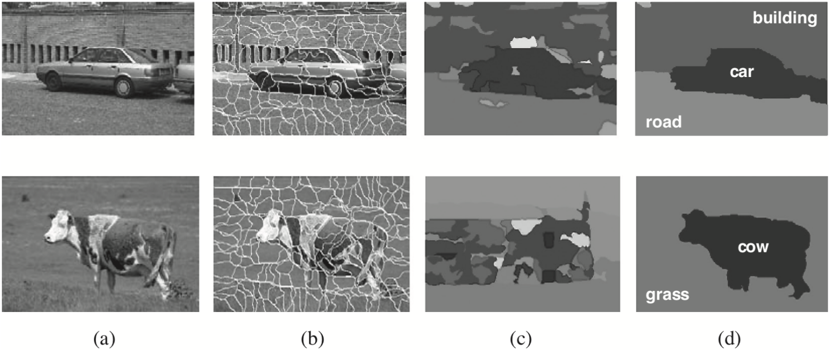


Figure 4.B.1 — Two examples of image segmentation results (a) The original image. (b) An oversegmentation known as superpixels; each superpixel is associated with a random variable that designates its segment assignment. The use of superpixels reduces the size of the problems. (c) Result of segmentation using node potentials alone, so that each superpixel is classified independently. (d) Result of segmentation using a pairwise Markov network encoding interactions between adjacent superpixels.

Log-linear models

It is common to parameterize in MRFs as log-linear models:

$$p(x_1, ..., x_p) = \frac{1}{Z} \exp[-\sum_{C \in \mathcal{C}} w_C \psi(x_C)]$$

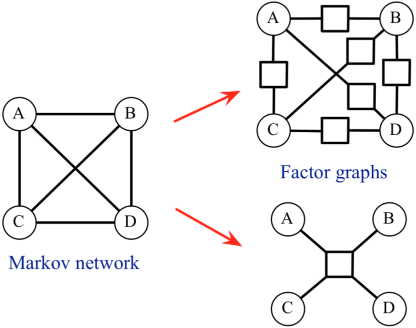
where $\psi(x_C)$ is sometimes called a "feature vector" that picks out the variable values in its scope.

For example, consider a 3-variable complete UG:

$$p(x_1, x_2, x_3) =$$
$$\frac{1}{Z} \exp[-w_1 x_1 - w_2 x_2 - w_3 x_3 - w_{12} x_1 x_2 - w_{23} x_2 x_3 - w_{13} x_1 x_3 - w_{123} x_1 x_2 x_3]$$

Often it assumed that the models are *hierarchical* so if some $w_C = 0$ then $w_D = 0$ for all $D \supset C$.

Factor graphs



May associate a factor ϕ with each *factor node* (square nodes), whose scope is the set of neighbors of the factor node. This is just a representational scheme to make explicit the structure of the factors.

What conditional independence properties does the factorization imply?

Let $X_i \sim X_j$ denote that X_i and X_j are adjacent $(X_i \in \text{Adj}(X_j, \mathcal{G}))$ and $X_i \not\sim X_j$ denote that they are not.

A distribution $p(x)$ satisfies the *pairwise Markov property* wrt UG $\mathcal{G}$ if

$$X_i \not\sim X_j \implies X_i \perp\!\!\!\perp X_j | (X \setminus \{X_i, X_j\})$$

Hammersley-Clifford Theorem (1971)

Theorem: Let $\mathcal{G}$ be an undirected graph. For any *positive* probabiliy distribution which has a density wrt product measure, the factorization and pairwise Markov properities (wrt $\mathcal{G}$) are equivalent:

Factorization $\iff$ Pairwise Markov

Non-adjacencies in BNs versus MRFs

Important: non-adjacencies do not correspond to the same independencies in BNs versus MRFs!

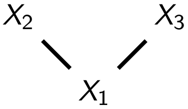


Figure: $X_2 \perp\!\!\!\perp X_3 | X_1$

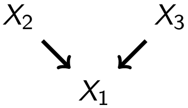


Figure: $X_2 \perp\!\!\!\perp X_3$ (and $X_2 \not\perp\!\!\!\perp X_3 | X_1$ w faithfulness)

Non-adjacencies in BNs versus MRFs

In a BN, $X_i \not\sim X_j$ means there is some set S (possibly empty) such that $X_i \perp\!\!\!\perp X_j | X_S$

In a MRF, $X_i \not\sim X_j$ means $X_i \perp\!\!\!\perp X_j |$rest

Local Markov property

$$\begin{aligned} \mathrm{Ne}(X_i, \mathcal{G}) &\equiv \{X_j : X_j \sim X_i \text{ in } \mathcal{G}\} && \text{(neighbors of } X_i\text{)} \\ \mathrm{Cl}(X_i, \mathcal{G}) &\equiv X_i \cup \mathrm{Ne}(X_i, \mathcal{G}) && \text{(closure of } X_i\text{)} \end{aligned}$$

Local Markov property

$$\begin{aligned} \mathrm{Ne}(X_i, \mathcal{G}) &\equiv \{X_j : X_j \sim X_i \text{ in } \mathcal{G}\} && \text{(neighbors of } X_i\text{)} \\ \mathrm{Cl}(X_i, \mathcal{G}) &\equiv X_i \cup \mathrm{Ne}(X_i, \mathcal{G}) && \text{(closure of } X_i\text{)} \end{aligned}$$

A distribution $p(x)$ satisfies the *local Markov property* wrt UG $\mathcal{G}$ if

$X_i \perp\!\!\!\perp X \setminus \mathrm{Cl}(X_i, \mathcal{G}) \,|\, \mathrm{Ne}(X_i, \mathcal{G})$

Global Markov property

Undirected graphs have a simpler separation criterion: C seperates A from B if all paths from (any element of) A to (any element of) B intersect C (written: $A \perp B|C$)

Let A, B, C be disjoint subsets of x. A distribution $p(x)$ satisfies the *global Markov property* wrt UG $\mathcal{G}$ if

$A \perp B|C \implies A \perp\!\!\!\perp B|C.$

Global Markov property

Undirected graphs have a simpler separation criterion: C seperates A from B if all paths from (any element of) A to (any element of) B intersect C (written: $A \perp B|C$)

Let A, B, C be disjoint subsets of x. A distribution $p(x)$ satisfies the *global Markov property* wrt UG $\mathcal{G}$ if

$A \perp B|C \implies A \perp\!\!\!\perp B|C.$

"A graphical criterion $\implies$ conditional independence in the distribution"

Equivalence of Markov properties in MRFs

Theorem. Let $\mathcal{G}$ be an undirected graph. For any *positive* probabiliy distribution which has a density wrt product measure, the factorization, local Markov, global Markov, and pairwise properities (wrt $\mathcal{G}$) are equivalent:

Factorization $\iff$ Global $\iff$ Local $\iff$ Pairwise

More generally, only Factorization $\implies$ Global $\implies$ Local $\implies$ Pairwise holds

Translating between undirected and directed graphs

Some algorithms are simpler/faster with an undirected graphical representation.

The *moral graph* (sometimes written: $\mathcal{G}^m$ or $\mathcal{M}[\mathcal{G}]$ or $\mathcal{G}^a$) of a DAG $\mathcal{G}$ over V is the undirected graph over V that contains an undirected edge between X_i and X_j if:
(a) there is a directed edge between them (in either direction), or
(b) X_i and X_j are both parents of the same node.

[examples on board]

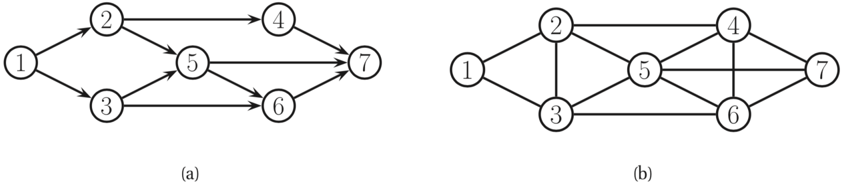
Translating between undirected and directed graphs

$\mathcal{G}_R$ stands for the induced subgraph over vertices in R, where $R \subseteq V$.

Let A, B, C be three disjoint sets of nodes in a DAG $\mathcal{G}$. C d-separates A from B in $\mathcal{G}$ if and only if C separates A from B in $(\mathcal{G}_{\mathrm{An}(A,B,C)})^m$.

(We use the convention that every vertex is in it's own ancestor set, so A, B, C are included in $\mathrm{An}(A, B, C)$.)

Moralization



(a)

(b)

Figure 19.2 (a) A DGM. (b) Its moralized version, represented as a UGM.

WARNING: you must be careful using UG properties to learn about DAG properties

Perhaps because undirected models are simpler to learn in some settings (e.g., convenient parameters in multivariate Gaussian case), there is a temptation to do structure/parameter learning with MRFs and then infer something about "underlying" DAG structure... this is tricky! and often done wrong.

WARNING: you must be careful using UG properties to learn about DAG properties

Perhaps because undirected models are simpler to learn in some settings (e.g., convenient parameters in multivariate Gaussian case), there is a temptation to do structure/parameter learning with MRFs and then infer something about "underlying" DAG structure... this is tricky! and often done wrong.

Because of colliders, the adjacencies in an MRF do not correspond to the adjacencies ("skeleton") of a BN.

WARNING: you must be careful using UG properties to learn about DAG properties

Perhaps because undirected models are simpler to learn in some settings (e.g., convenient parameters in multivariate Gaussian case), there is a temptation to do structure/parameter learning with MRFs and then infer something about "underlying" DAG structure... this is tricky! and often done wrong.

Because of colliders, the adjacencies in an MRF do not correspond to the adjacencies ("skeleton") of a BN.

Furthermore, though BNs may have a causal interpretation (or several), MRFs do not. Don't learn/reason with MRFs if what you care about is causality – you're doing the wrong thing! (There are some subtle ways to do this correctly, but it requires care...)

Markov equivalence?

Every undirected graph constitutes its own equivalence class. Why?

Graphical representations which are neither DAGs nor UGs

Why people in ML and statistics like DAGs and UGs:

- DAGs and UGs are associated with “nice” equivalent Markov properties/factorizations
- DAGs and UGs are easy to parameterize with common parametric families (next lectures)
- DAGs and UGs seem to work well (sufficiently well) in a lot probabilistic applications
- DAGs have a natural causal interpretation (later lectures)

What about other kinds of graphical representations?

Chain graphs

Chain graphs are mixed graphs with both directed ($\rightarrow$) and undirected ($-$) edges but no partially directed cycles. We can define a factorization property and some Markov properties just like with DAGs and UGs, but a bit more complicated.

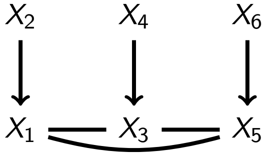


Figure: A chain graph

Chain graph properties

A block B is defined as a maximal set of vertices such that every vertex pair in $\mathcal{G}_B$ is connected by an undirected path. The set of blocks in a CG $\mathcal{G}$, denoted by $\mathcal{B}(\mathcal{G})$, partitions the vertices in $\mathcal{G}$. Given a CG $\mathcal{G}$, define the augmented graph $\mathcal{G}^a$ to be an UG constructed from $\mathcal{G}$ by replacing all directed edges with undirected edges and connecting all vertices in $\text{Pa}(B, \mathcal{G})$ for every block B in $\mathcal{G}$ by undirected edges.

Factorization:

$$p(x) = \prod_{B \in \mathcal{B}(\mathcal{G})} p(B \mid \text{Pa}(B, \mathcal{G}))$$

and

$$p(B \mid \text{Pa}(B, \mathcal{G})) = \frac{1}{Z(\text{Pa}(B, \mathcal{G}))} \prod_{\{C \in \mathcal{C}((\mathcal{G}_{\text{Bd}(B,\mathcal{G})})^a) : C \not\subseteq \text{Pa}(B,\mathcal{G})\}} \phi_C(C)$$

where $\text{Bd}(X_i) \equiv \text{Pa}(X_i) \cup \text{Ne}(X_i)$.

Chain graph properties

local Markov:

$$X_i \perp\!\!\!\perp \mathrm{Nd}(X_i, \mathcal{G}) \,|\, \mathrm{Bd}(X_i, \mathcal{G})$$

may define pairwise and global properties also...
prove equivalence for positive densities...

Chain graphs

Note: *causal* interpretation of CGs turns out to be quite tricky, maybe counter-intuitive. Lauritzen and Richardson (2002) show that most inuitive interpretations of the undirected edges don't formally work out.[1] They propose an interpretation of undirected components as equilibrium distributions of a Gibbs sampler – see their paper for details.

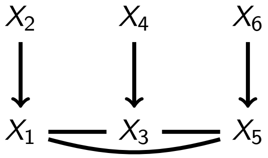


Figure: A chain graph

[1] Lauritzen, S. and Richardson, T.S. (2002) Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society B*.

Directed (cyclic) graphs

A DG is like a DAG but allowing for cycles.

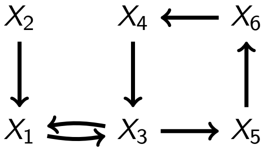


Figure: A directed graph with cycles

Again, we could define a factorization property, local and global Markov properties... but the equivalence of these properties breaks down. More importantly, the global Markov property **fails to hold** in systems of equations naturally associated with cyclic graphs! That is, d-separation does not necessarily imply conditional independence.

Directed (cyclic) graphs

$$\begin{aligned} X_1 &= f_1(X_2, X_3, \epsilon_1) \\ X_2 &= f_2(X_1, X_3, \epsilon_2) \\ X_3 &= f_3(X_1, X_2, \epsilon_3) \end{aligned}$$

One may associate directed graphs to represent systems of *non-recursive* structural equations, i.e., equations that cannot be placed in a recursive order. This is common esp. in economics, psychology, and some biological sciences.

Directed (cyclic) graphs

$$\begin{aligned} X_1 &= f_1(X_2, X_3, \epsilon_1) \\ X_2 &= f_2(X_1, X_3, \epsilon_2) \\ X_3 &= f_3(X_1, X_2, \epsilon_3) \end{aligned}$$

One may associate directed graphs to represent systems of *non-recursive* structural equations, i.e., equations that cannot be placed in a recursive order. This is common esp. in economics, psychology, and some biological sciences.

If these equations are linear, then all is (mostly) well. But if they are nonlinear, a d-seperation fact in the associated graph may fail to imply a conditional independence in the joint distribution. The global Markov property does not hold.

Nonlinear non-recursive equations violate the global Markov property (example)

$$\begin{aligned} X &= \epsilon_X \\ Y &= \epsilon_Y \\ Z &= W \times Y + \epsilon_Z \\ W &= Z \times X + \epsilon_W \end{aligned}$$

where each $\epsilon \sim N(0,1)$ (independent)

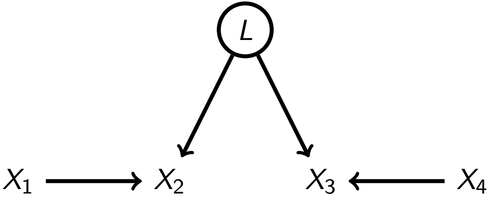
$$\begin{aligned} p(x,y,z,w) =& \frac{1}{4\pi^2} \exp(\frac{-x^2}{2}) \exp(\frac{-y^2}{2}) \\ & \times \exp(\frac{-(z-wy)^2}{2}) \exp(\frac{-(w-zx)^2}{2}) \mid \frac{1}{1-xy} \mid \end{aligned}$$

$X \perp_d Y | \{Z, W\}$ but $X \not\perp Y | \{Z, W\}$

example from Spirtes (1995, UAI)

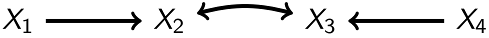
Acyclic directed mixed graphs (ADMGs)

We saw last week we may be interested in models with latent (unobserved/hidden) variables



Acyclic directed mixed graphs (ADMGs)

If we're interested in independence relations among the observed variables, we can represent these with an acyclic directed mixed graph. We can define a separation criterion (m-separation) which straightforwardly extends d-separation to mixed graphs, and proceed as before.



Latent projections

Definition. Let $\mathcal{G}$ be a DAG with vertex set $X \cup L$ where the vertices in X are observed, while those in L are latent. The latent projection $\mathcal{G}(X)$ is a directed mixed graph with vertex set X, where for every pair of distinct vertices $X_i, X_j \in X$:

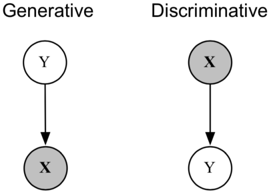
(i) $\mathcal{G}(X)$ contains $X_i \to X_j$ iff there is a directed path $X_i \to \cdots \to X_j$ on which every non-endpoint vertex is in L.

(ii) $\mathcal{G}(X)$ contains $X_i \leftrightarrow X_j$ iff there exists a path of the form $X_i \leftarrow \cdots \to X_j$, on which every non-endpoint vertex is a non-collider and in L.

[examples on board]

Generative vs. discriminative modeling

Say there is an outcome (label) we are interested in predicting Y and set of features X.
In such a setting we just care about $p(y|x)$.



We may choose between

- a generative approach: modeling $p(y)p(x|y) = p(y,x)$, and then calculating the conditional
- a discriminative approach: modeling just $p(y|x)$ (does not require specifying a model for $p(x)$!)

Generative vs. discriminative modeling

In the first case we typically need to find some way to parameterize $p(x|y)$ or $p(x_i|\operatorname{Pa}(X_i, \mathcal{G}))$ (parents include Y)

In the second case we typically need to find some way to parameterize $p(y|x)$

Generative vs. discriminative modeling

In the first case we typically need to find some way to parameterize $p(x|y)$ or $p(x_i|\operatorname{Pa}(X_i, \mathcal{G}))$ (parents include Y)

In the second case we typically need to find some way to parameterize $p(y|x)$

Generative approaches include using specific Bayesian network models (e.g., naive Bayes) or specific MRFs.

Discriminative approaches include logistic regression, traditional neural networks, or *conditional* MRFs.

Generative vs. discriminative modeling

In the first case we typically need to find some way to parameterize $p(x|y)$ or $p(x_i|\,\mathrm{Pa}(X_i, \mathcal{G}))$ (parents include Y)

In the second case we typically need to find some way to parameterize $p(y|x)$

Generative approaches include using specific Bayesian network models (e.g., naive Bayes) or specific MRFs.

Discriminative approaches include logistic regression, traditional neural networks, or *conditional* MRFs.

There are trade-offs to these approach which depend on the structure of the problem (what conditional independence assumptions are reasonable, to make generative modeling more tractable? how much data do you have?). Ng and Jordan (2001) have an interesting paper where they compare generative and discriminative methods in terms of their asymptotic error and sample-size efficiency.

Generative vs. discriminative modeling

Denote a classification method for Y by $m(X)$. Naive Bayes and logistic regression are two classification methods $m_1(X), m_2(X)$. We can compare methods by their generalization error: $\varepsilon(m) = P(m(X) \neq y)$.

Ng and Jordan (2001) show, in a comparison of naive Bayes and logistic regression, that the generative method has higher asymptotic error, but approaches its (higher) asymptotic error *faster* as a function of sample size. Thus, as the number of samples increases, there are settings where we would expect naive Bayes to initially do better (because it is nearer to it's asymptotic level), but then it is overtaken in performance by logistic regression eventually.

Generative vs. discriminative modeling

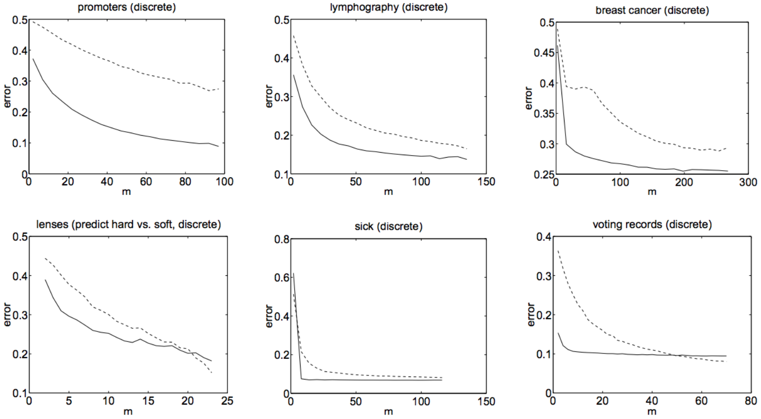
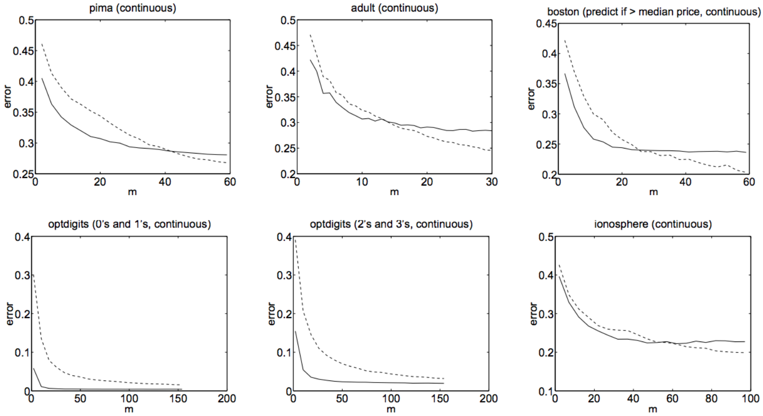


Figure 1: Results of 15 experiments on datasets from the UCI Machine Learning repository. Plots are of generalization error vs. m (averaged over 1000 random train/test splits). Dashed line is logistic regression; solid line is naive Bayes.

(partial figure from Ng and Jordan (2001))

Generative vs. discriminative modeling



(partial figure from Ng and Jordan (2001))

Generative vs. discriminative modeling

The point of this is that sometimes generative modeling will be better, and sometimes discriminative modeling: it depends on the circumstances and the regime (sample size + assumptions)

Conditional Random Fields (CRFs)

CRFs are undirected graphical models for conditional distributions $p(y|x)$. The undirected graph $\mathcal{G}$ has vertices V which index random variables in $X \cup Y$.

$$p(y|x) = \frac{1}{Z(x)} \prod_{C \in \mathcal{C}} \phi_C(x, y_C)$$

where

$$Z(x) = \sum_y \prod_{C \in \mathcal{C}} \phi_C(x, y_C)$$

so now the partition function is indeed a *function* of x.

Here, we don't say anything about $p(x)$! (Note: the notation is a bit weird. y_C is supposed to mean that we only consider cliques containing Y's, i.e., no potential functions that involve only variables in X.)

Conditional Random Fields (CRFs)

A common parameterization of the potential functions is

$$\phi_C(x, y_C) = \exp(w^T \psi(x, y_C))$$

where w is a vector of weight parameters to be learned (estimated) from data and ψ extracts the vector features x, y_C.

Example: chain structured CRF

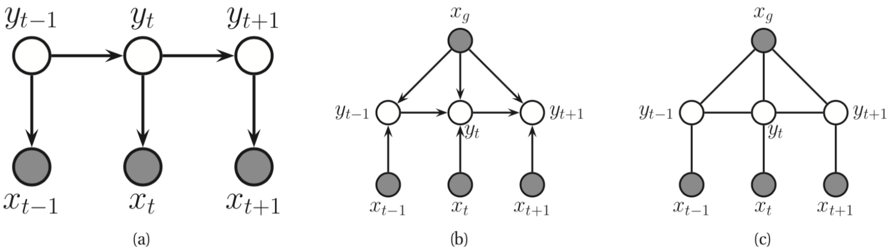


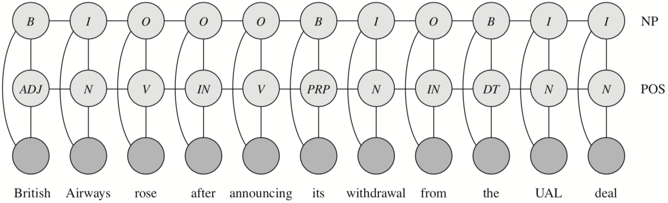
Figure 19.14 Various models for sequential data. (a) A generative directed HMM. (b) A discriminative directed MEMM. (c) A discriminative undirected CRF.

(fig from Murphy (2012), x_g denote "global features," e.g., language of text)

$$p(y|x) = \frac{1}{Z(x)} \prod_{t=1}^{T} \phi(x, y_t) \prod_{t=1}^{T-1} \phi(x, y_t, y_{t+1})$$

where $x = (x_1, ..., x_T, x_g)$

Example: chain structured CRF


B I O O O B I O B I I
NP
ADJ N V IN V PRP N IN DT N N
POS
British Airways rose after announcing its withdrawal from the UAL deal

KEY

B	Begin noun phrase	*V*	Verb
I	Within noun phrase	*IN*	Preposition
O	Not a noun phrase	*PRP*	Possesive pronoun
N	Noun	*DT*	Determiner (e.g., a, an, the)
ADJ	Adjective		

A chain CRF that performs joint part-of-speech labeling and noun-phrase segmentation. Here, B indicates the beginning of a noun phrase, I other words in the noun phrase, and O words not in a noun phrase. The labels for the second chain are parts of speech.