Topic 8: GRAPHICAL MODELS

STAT 37710/CMSC 25400 Machine Learning
Risi Kondor, The University of Chicago

Three types of "Probability"

1. **Frequency of repeated trials:** if an experiment is repeated infinitiely many times, $0 \leq p(A) \leq 1$ is the fraction of times that the outcome will be A. Typical example: number of times that a coin comes up heads.
   $\rightarrow$ Frequentist probability.
2. **Degree of belief:** A quantity obeying the same laws as the above, describing how likely we think a (possibly deterministic) event is. Typical example: the probability that the Earth will warm by more than 5° F by 2100. $\rightarrow$ Bayesian probability.
3. **Subjective probability:** "I'm 110% sure that I'll go out to dinner with you tonight."

Mixing these three notions is a source of lots of trouble. We will start with the frequentist interpretation and then discuss the Bayesian one.

Why do we need probability for ML?

Two distinct reasons:

1. To analyze, understand and predict the performance of learning algorithms (Statistical Learning Theory, PAC model, etc.)
2. To build flexible and intuitive **probabilistic models**.

Probabilistic vs. Algorithmic learning

- Algorithmic ML (e.g., SVMs):
  - Strictly focus on the task at hand $\rightarrow$ discriminative
  - Black box
  - Algorithms often motivated directly by optimization methods $\rightarrow$ fast
  - Examples: the perceptron, SVM, etc.
  - "Frequentist"
- Probabilistic ML (e.g., graphical models):
  - Everything in the world is a random variable $\rightarrow$ generative
  - Flexible modeling framework for incorporating prior knowledge
  - Models are often expressed with graphs $\rightarrow$ efficient message passing algorithms
  - Example: k–means clustering
  - "Bayesian"

[Breiman: Statistical modeling: the two cultures]

Joint probabilities and independence

Machine learning applications often involve a large number of variables (features) $X_1, \ldots, X_n$.

- The **conditional probability** of X_i given X_j is

$$p(x_i|x_j) = \mathbb{P}(X_i = x_i \mid X_j = x_j) \qquad p(x_i, x_j) = p(x_i|x_j)\, p(x_j).$$

- X_i and X_j are **independent** (denoted $X_i \perp\!\!\!\perp X_j$) if

$$p(x_i|x_j) \text{ is indep of } x_j \qquad \Longleftrightarrow \qquad p(x_i, x_j) = p(x_i)\, p(x_j).$$

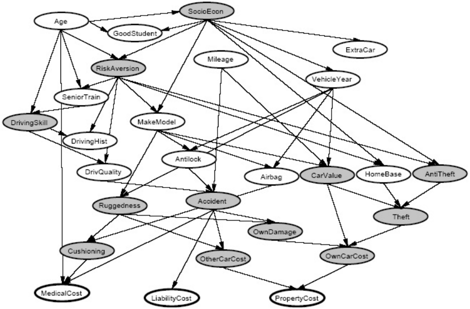
- X_i is **conditionally independent** of X_j given X_k (denoted $X_i \perp\!\!\!\perp X_j | X_k$) if

$$p(x_i, x_j|x_k) = p(x_i|x_k)\, p(x_j|x_k).$$

IDEA: When faced with a large number of features, use our prior knowledge of indepdencies to make learning easier.

Directed graphical models

Also called Bayes nets or Belief Networks. Each vertex $v \in V$ corresponds to a random variable. Graph must be acyclic but not necessarily a tree.



The general form of the joint distribution of all the variables is

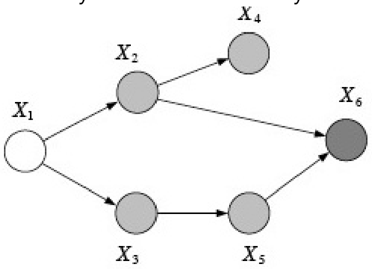
$$p(\mathbf{x}) = \prod_{v \in V} p(x_v | \mathbf{x}_{\mathrm{pa}(v)}),$$

where $\mathrm{pa}(v)$ are all the parents of v in the graph.

Directed graphical models

Assuming that $X_1, \ldots, X_6$ are binary random variables, how many numbers are need to describe their joint distribution? $2^6 - 1 = 63$.

Now what if we know that they conform to this Bayes net?



Each $p(x_i|x_j)$ corresponds to a 2×2 table, but rows sum to 1, so only 2 numbers required. $p(x_6|x_2, x_5)$ requires 4 numbers.
Total: $1 + 2 + 2 + 2 + 2 + 4 = 13$. Quite a saving!

Example: Markov chains

- If $x_1, x_2, \ldots$ is a series of (discrete or continuous) random variables corresponding to a process evolving in time, then x_t should only depend on what happened in the past:

$$p(x_t | x_1, \ldots, x_{t-1}, x_{t+1}, \ldots) = p(x_t | x_1, \ldots, x_{t-1}).$$

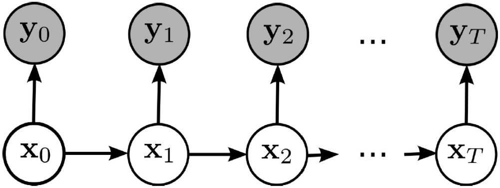
- The sequence $x_1, x_2, \ldots$ is said to be a **k 'th order Markov chain** if

$$p(x_t | x_1, \ldots, x_{t-1}, x_{t+1}, \ldots) = p(x_t | x_{t-1}, \ldots, x_{t-k}).$$

- A (first order) Markov chain is said to be **stationary** if the $p(x_t | x_{t-1})$ **transition probabilities** are independent of t,

$$p(x_t | x_{t-1}) = M_{x_t, x_{t-1}}.$$

Hidden Markov Models (HMM)



An HMM is a Markov chain of unobserved random variables $x_1, x_2, \ldots,$ each of which is related to an oberved random variable $y_1, y_2, \ldots$.

Example: Tracking, part of speech tagging, phonemes, physiological states of babies,...

Applications of HMMs

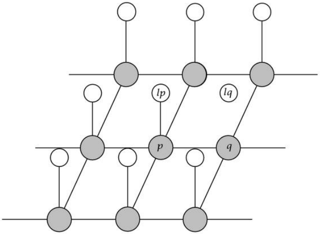
HMMs and related state space models are widely applied in

- speech recognition (which phoneme/word/etc.)
- part of speech tagging (is it a NP, VP, etc.)?
- biological sequence analysis (intron or extron)?
- time series analysis (finance, climate, etc.)
- robotics (what is the actual location of the robot)?
- tracking

Undirected graphical models

Also called Markov Random Fields. Graph can be any undirected graph.

Common example used for image segmentation:

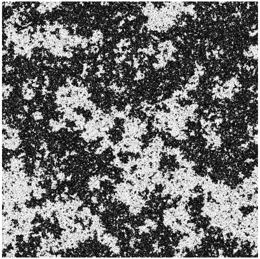


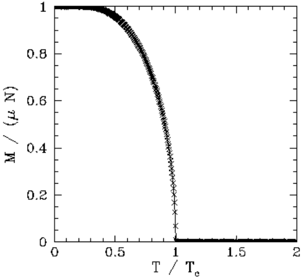
The general form of the joint distribution over all the variables is

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{c \in \text{Cliques}(\mathcal{G})} \phi_c(\mathbf{x}_c)$$

where each ϕ_c is a potentially different **clique potential** (just a positive function) and Z is the **normalizing factor** $Z = \sum_{\mathbf{x}} \prod_{c \in \text{Cliques}(\mathcal{G})} \phi_c(\mathbf{x}_c)$.

Example: the Ising model



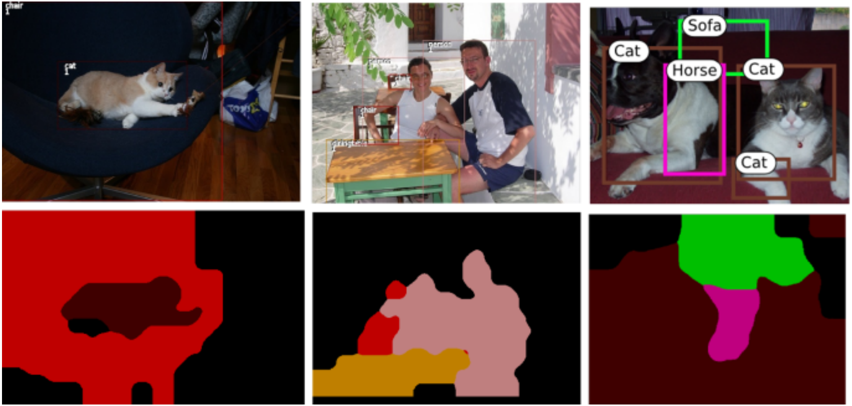


Imagine an infinite grid of $\{-1, +1\}$ valued random variables in which neighboring variables are connected by the potential

$$\phi(x_i, x_j) = e^{-\beta/2(x_i - x_j)^2}.$$

Simple model of ferromagnetism. Exhibits a **phase transition**.

Example: MRFs for segmentation



Purpose of graphical models

In ML we often have a large number of variables related in complicated ways.

Graphical models

- capture prior knowledge about relationships between variables
- provide a compact representation of distributions over many variables
- define a specific hypothesis class
- help with figuring out causality
- the variables can be either discrete (e.g., “airbag yes/no”), continous (e.g., “value”) or a mixture of both types

Tasks for graphical models

- Model selection (i.e., learn the graph itself from data)
- Learn the parameters of the model from data (i.e., the individual conditionals or clique potentials)
- Deduce conditional independence relations
- Infer marginals and conditional distributions

Inference

Partition V, the set of nodes, into three sets:

1. the set O of observed nodes
2. the set Q of query nodes
3. the set L of latent nodes

Interested in $$p(\mathbf{x}_Q|\mathbf{x}_O) = \frac{\sum_{\mathbf{x}_L} p(\mathbf{x}_Q, \mathbf{x}_L, \mathbf{x}_O)}{\sum_{\mathbf{x}_L, \mathbf{x}_Q} p(\mathbf{x}_Q, \mathbf{x}_L, \mathbf{x}_O)}$$

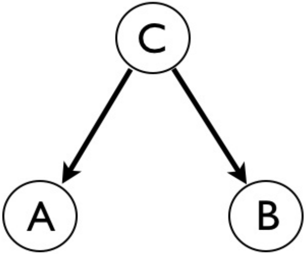
Essential for both

- **Training**, when we are trying to learn the distribution of some of the nodes from data.
- **Prediction**, when we are trying to predict the values of some nodes (the output) given the values of some other nodes (the input)

Question: How can we do this in less than $m^{|Q|+|L|}$ time?

Directed graphical models (Bayes nets)

Common cause

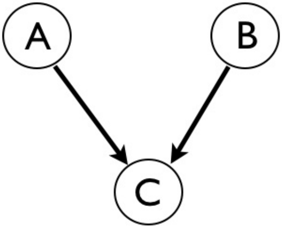


$$X_A \not\perp\!\!\!\perp X_B \qquad \text{but} \qquad X_A \perp\!\!\!\perp X_B \,|\, X_C$$

Therefore, if C is ***observed***, then A and B become independent.

Example: Lung cancer $\perp\!\!\!\perp$ Yellow teeth $|$ Smoking

Explaining away



$$X_A \perp\!\!\!\perp X_B \qquad \text{but} \qquad X_A \not\!\perp\!\!\!\perp X_B \,|\, X_C$$

Therefore, if C is ***not oberseved*** (and neither are any of its descendents) then A and B become independent.

Example: Burglary $\not\!\perp\!\!\!\perp$ Earthquake $|$ Alarm

D–separation

Is X independent of Y given the set of nodes S?

An underected path from X to Y is said to be **blocked** if

1. it includes at least one node Z from S such that the arrows along the path at Z meet head to tail or tail to tail; or
2. it includes at least one node W such that the arrows along the path at W meet head to head, and neither W nor any of its descendants are in S.

Theorem

$X \perp\!\!\!\perp Y \mid S$ if and only if all paths from X to Y are blocked.

Learning parameters in Bayes nets

Recall the general form of a discrete Bayes net:

$$p(\mathbf{x}) = \prod_{v \in V} p(x_v | \mathbf{x}_{\mathrm{pa}(v)}) \qquad x_v \in \{1, 2, \ldots, k_v\}\,.$$

Assuming for now that everyone has two parents, $(x_{m(v)}, x_{f(v)})$, the conditional distributions can be parametrized by 3D arrays $\theta_1, \ldots, \theta_k$:

$$p(x_v | x_{m(v)}, x_{f(v)}) = [\theta_v]_{x_{m(v)}, x_{f(v)}, x_v}\,.$$

To ensure normalization, $\sum_{x_v} [\theta_v]_{x_{m(v)}, x_{f(v)}, x_v} = 1$ for all $x_{m(v)}, x_{f(v)}$.

Given data $\mathcal{D} = (\mathbf{x}^1, \ldots, \mathbf{x}^T)$, what is the MLE setting of $(\theta_v)_{v \in V}$?

Simpson's paradox: word of caution

You are trying to determine whether a particular treatment for a serious disease is beneficial. Given the following observations would you recommend it?

	Survived	Did not survive	Survival rate
Treatment	20	20	50%
No treatment	16	24	40%

Now what if you discovered that the breakdown by gender was this?

Males	Survived	Did not survive	Survival rate
Treatment	18	12	60%
No treatment	7	3	70%

Females	Survived	Did not survive	Survival rate
Treatment	2	8	20%
No treatment	9	21	30%

Simpson's paradox

- A graphical model can never capture all the variables that might possibly be relevant. In the first case we ignored gender. This can affect what interpretation the model suggests.
- The fact that there is an arrow from A (treatment) to B (outcome) does not imply that A causes B. In our case we had a hidden common cause, gender, of the opposite effect on B.
- To tease out causal structure we need more sophisitcated tools than just ordinary graphical models: need to introduce **interventions**.
- Observational studies are not sufficient. The gold standard in medicine is **randomized controlled trials (RCTs)**.

Learning parameters in Bayes nets

$$p(x_v|x_{m(v)}, x_{f(v)}) = [\theta_v]_{x_{m(v)}, x_{f(v)}, x_v}\,.$$

$$\ell(\theta|\mathcal{D}) = \prod_{t=1}^{T} \prod_{v \in V} [\theta_v]_{x^t_{m(v)}, x^t_{f(v)}, x^t_v} = \prod_{v \in V} \ell_v(\theta_v|\mathcal{D})$$

$$\ell_v(\theta_v|\mathcal{D}) = \prod_{t=1}^{T} [\theta_v]_{x^t_m, x^t_f, x^t_v} =$$

$$\prod_a \prod_b \frac{N_{a,b}!}{N_{a,b,1}!\, N_{a,b,2}! \ldots N_{a,b,k_v}!}\, [\theta_v]^{N_{a,b,1}}_{a,b,1}\, [\theta_v]^{N_{a,b,2}}_{a,b,2} \ldots [\theta_v]^{N_{a,b,v_k}}_{a,b,v_k}$$

$$N_{a,b,c} = \left|\, \{\ t \mid x^t_m = a,\ x^t_f = b,\ x^t_v = c\ \}\, \right|$$

Learning parameters in Bayes nets

Each

$$\ell_{v,a,b}(\theta_v|\mathcal{D}) = \frac{N_{a,b}!}{N_{a,b,1}!\, N_{a,b,2}! \ldots N_{a,b,k_v}!}\, [\theta_v]_{a,b,1}^{N_{a,b,1}} \ldots [\theta_v]_{a,b,v_k}^{N_{a,b,v_k}}$$

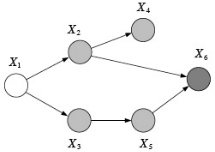
is just a multinomial like in Naive Bayes, so we know the MLE is

$$[\widehat{\theta}_v]_{a,b,c} = \frac{N_{a,b,c}}{\sum_c N_{a,b,c}} \,.$$

As before, can also use biased estimator

$$[\widehat{\theta}_v]_{a,b,c} = \frac{N_{a,b,c} + \gamma}{\sum_c (N_{a,b,c} + \gamma)} \,.$$

Inference in Bayes nets: example



The key is to factor and then apply the distributive law.

$$
\begin{aligned}
p(\mathbf{x}_1|\bar{\mathbf{x}}_6) &= p(\mathbf{x}_1, \bar{\mathbf{x}}_6)/p(\bar{\mathbf{x}}_6) \\
&= p(\mathbf{x}_1, \bar{\mathbf{x}}_6)/\sum_{\mathbf{x}_1'} p(\mathbf{x}_1', \bar{\mathbf{x}}_6)
\end{aligned}
$$

$$
\begin{aligned}
p(\mathbf{x}_1, \bar{\mathbf{x}}_6) &= \sum_{\mathbf{x}_2}\sum_{\mathbf{x}_3}\sum_{\mathbf{x}_4}\sum_{\mathbf{x}_5} p(\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{x}_1)p(\mathbf{x}_3|\mathbf{x}_1)p(\mathbf{x}_4|\mathbf{x}_2)p(\mathbf{x}_5|\mathbf{x}_3)p(\bar{\mathbf{x}}_6|\mathbf{x}_2, \mathbf{x}_5) \\
&= p(\mathbf{x}_1)\sum_{\mathbf{x}_2} p(\mathbf{x}_2|\mathbf{x}_1)\sum_{\mathbf{x}_3} p(\mathbf{x}_3|\mathbf{x}_1)\sum_{\mathbf{x}_4} p(\mathbf{x}_4|\mathbf{x}_2)\sum_{\mathbf{x}_5} p(\mathbf{x}_5|\mathbf{x}_3)p(\bar{\mathbf{x}}_6|\mathbf{x}_2, \mathbf{x}_5) \\
&= p(\mathbf{x}_1)\sum_{\mathbf{x}_2} p(\mathbf{x}_2|\mathbf{x}_1)\sum_{\mathbf{x}_3} p(\mathbf{x}_3|\mathbf{x}_1)\Phi_5(\mathbf{x}_2, \mathbf{x}_3)\sum_{\mathbf{x}_4} p(\mathbf{x}_4|\mathbf{x}_2) \\
&= p(\mathbf{x}_1)\sum_{\mathbf{x}_2} p(\mathbf{x}_2|\mathbf{x}_1)\Phi_4(\mathbf{x}_2)\sum_{\mathbf{x}_3} p(\mathbf{x}_3|\mathbf{x}_1)\Phi_5(\mathbf{x}_2, \mathbf{x}_3) \\
&= p(\mathbf{x}_1)\sum_{\mathbf{x}_2} p(\mathbf{x}_2|\mathbf{x}_1)\Phi_4(\mathbf{x}_2)\Phi_3(\mathbf{x}_1, \mathbf{x}_2) \\
&= p(\mathbf{x}_1)\Phi_2(\mathbf{x}_1)
\end{aligned}
$$

Is there a general algorithm that allows us to find factorizations like this?

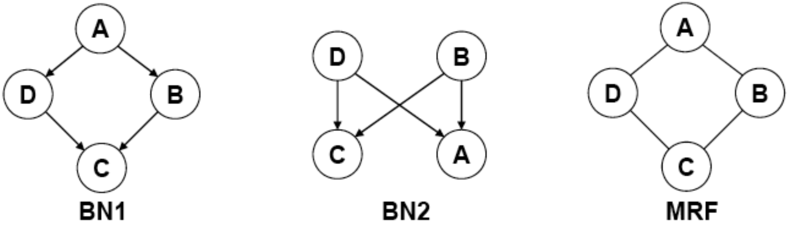
$\rightarrow$ Message passing algorithms

Undirected graphical models

Undirected graphical models

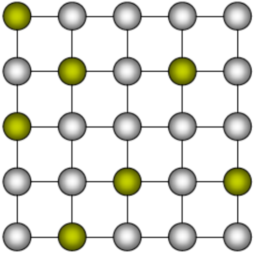
Not every type of conditional dependency structure can be represented by a Bayes net. Example:

$$X_A \perp\!\!\!\perp X_C | \{X_B, X_D\}, \qquad X_B \perp\!\!\!\perp X_D | \{X_A, X_C\}.$$

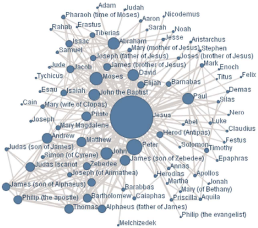


Exercise: Give an example of a structure that cannot be represented by a directed model either.

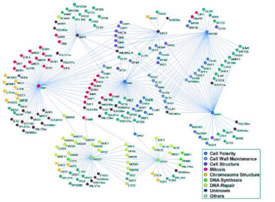
Examples of undirected models



Grid model (e.g., Ising)



Social Network



Protein interaction net

Ordinary separation

Recall the general form of the undirected models:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{c \in \text{Cliques}(\mathcal{G})} \phi_c(\mathbf{x}_c)$$

Is X independent of Y given the set of nodes S?

Theorem

$X \perp\!\!\!\perp Y \mid S$ if and only if all paths from X to Y contain at least one node in S.

This is simpler than in the directed case.

Parameter estimation and inference

In undirected models

- Parameter estimation: Not as easy as in the directed case!
- Inference : message passing algorithms.

FURTHER READING

- David Barber: **Bayesian Reasoning and Machine Learning** (online)
- Daphne Koller and Nir Friedman: **Probabilistic Graphical Models**
- Tutorial by Sam Roweis:
  http://videolectures.net/mlss06tw_roweis_mlpgm/
- Coursera course "Probabilistic Graphical Models" by Daphne Koller