Bayesian statistics summary

- Bayesian statistical models
  - We prefer generative models for likelihood (and prior)
  - Conjugate priors are preferred when they are accurate enough
  - Bayesian updating for sequences of independent data
    - Yesterday's posterior becomes today's prior
- Inference uses Bayes rule to "invert" the forward model
  - Result is the posterior distribution
  - MAP estimate provides a single "best" number (often not the best)

Bayesian statistics summary

- Related topics coming up
  - Predictive distribution
    - Marginalizes out uncertainty about models
  - Model selection
  - Estimation and decision making

Bayesian Sequential Update

$$p(\theta \mid D_{1:N}) \propto \left\{ p(\theta) \prod_{i=1}^{N-1} \left(p(D_i \mid \theta) \right) \right\} p(D_N \mid \theta)$$

Posterior from 1:N-1 is now the prior

Already introduced with the example for the Bayesian estimate of the mean

Predictive Distribution

$$p(x \mid X) = \int p(x \mid \theta) p(\theta \mid X) d\theta$$

Test data

Training data

Model Selection

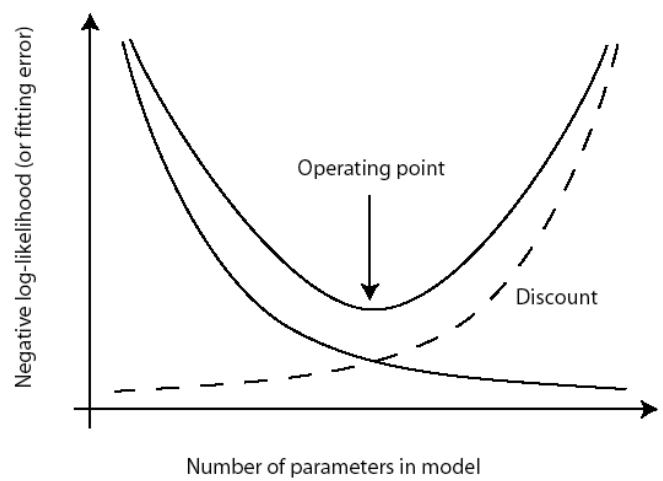
- Model selection refers to choosing among different instances within a model class (1) or different model classes (2).
- Examples:
  - The number of clusters (1)
  - The degree of a polynomial to fit a curve to data (1)
  - Polynomials versus other basis functions such as Fourier (2)

Model Comparison Difficulties

- Prior densities of different models are typically of different dimensionality (leads to expensive integration).
- Good likelihoods help select models, but constructing them is an exacting task.
  - Don't forget about the "negative space"
    - A more complex model (e.g., more objects in a scene) explains more data, but it also proposes more data where there is none.
    - Missing data must be penalized!
- Good priors over different model classes are often not obvious

Solutions (penalize complexity)

- Typical approach is to focus on the balance between fitting accuracy, and model complexity using various penalties.



Solutions (penalize complexity)

- Typical approach is to focus on the balance between fitting accuracy, and model complexity using various penalties.
- AIC (An information criterion, Akaike, 74)

Replace log likelihood, $\log(\mathrm{p}(\mathrm{D}|\theta))$, with $\log(\mathrm{p}(\mathrm{D}|\theta)) - M$ where M is the number of adjustable parameters.

Solutions (penalize complexity)

- Typical approach is to focus on the balance between fitting accuracy, and model complexity using various penalties.
- BIC (Bayesian information criterion)

Replace log likelihood, $\log(\mathrm{p}(\mathrm{D}|\theta))$, with $\log(\mathrm{p}(\mathrm{D}|\theta)) - \frac{1}{2} M \log(N)$ where M is the number of adjustable parameters, N is the number of data points. This is the usual approximation. See Bishop, page 216-217 for a more complicated version.

Often also called minimum discription length (MDL)

The dependency on N may seem confusing. Note that the likelihood typically depends on N (often N is an exponent), but the formula above does not expose this.

Solutions (penalize complexity)

- Typical approach is to focus on the balance between fitting accuracy, and model complexity using various penalties.
- DIC (Deviation information criterion)
  - Details omitted (see Google)
  - Slightly more complex, but easier to compute using MCMC sampling
  - Still relies on strong assumptions (distribution is approximately multivariate normal)

Solutions (likelihood function)

- Incorrect complex models may predict lots of data where there is none
- Solution is to model missing data
- Example --- finding asteroids from detections amidst noise
  - Predicting more asteroids explains more data, but we expect to see detections for them most of the time.
  - Good modeling the probability of noise detections and probability of missing detections has a greater affect on the posterior than a prior (necessarily not very strong) on the number of asteroids.

Solutions (integrating parameter uncertainty)

$$p(D|M_i) = \int_{\Omega_i} p(D|\theta)\, p(\theta|M_i)\, d\theta \qquad \text{(Model evidence)}$$

and we can evaluate $\mathrm{p}(M_i|D)$ by Bayes.

The dimension of the space of θ (Ω_i in the integral) is typically a function of i.

This is argued (Bishop, §3.4) to be a principled way to penalize complex models because complex models spread their probability mass over greater support (but thec skeptic asks when or why the amount of penalty is correct).

Under additional approximations and assumptions, this becomes BIC (Bishop, §4.4.1).

Solutions (integrating parameter uncertainty)

$$p(D|M_i) = \int_{\Omega_i} p(D|\theta)\, p(\theta|M_i)\, d\theta \qquad \text{(Model evidence)}$$

and we can evaluate $\mathrm{p}(M_i|D)$ by Bayes.

We can compare two models abilities to explain data by the Bayes factor

$$K_{ij} = \frac{p(D|M_i)}{p(D|M_j)} \qquad \text{(We can augment with factors for the priors } p(M) \text{ if known)}$$

Supplementary material on lecture notes page has a link to a classic reference on Bayes factors (Kass and Raftery, 95).

Solutions (integrating parameter uncertainty)

$$K_{ij} = \frac{p(D|M_i)}{p(D|M_j)} \qquad \text{(Bayes factor)}$$

Rules of thumb for K (from Jeffreys, via WikiPedia)

K	dB	bits	Strength of evidence
< 1:1	< 0		Negative (supports M_2)
1:1 to 3:1	0 to 5	0 to 1.6	Barely worth mentioning
3:1 to 10:1	5 to 10	1.6 to 3.3	Substantial
10:1 to 30:1	10 to 15	3.3 to 5.0	Strong
30:1 to 100:1	15 to 20	5.0 to 6.6	Very strong
> 100:1	> 20	> 6.6	Decisive

Solutions (model averaging)

Recall the predictive distributions

$$p(x \mid X) = \int p(x \mid \theta)\, p(\theta \mid X)\, d\theta$$

To mitigate uncertainty of different models

$$p(x|X) = \sum_i p(M_i) \int_{\Omega_i} p(x \mid \theta_i)\, p(\theta_i \mid X, M_i)\, d\theta$$

Note the assumption that M_i influences x through θ_i only, so no conditioning on M_i in the first factor in the integral.

Comments on Bayes factors, etc.

- Bayes factors can be used to derive BIC under specific conditions
- Otherwise you will normally need a numeric approximation of the integral
- $p(D|M_i)$ tells you the probability of observing the data you did under a well specified, and possibly flawed model—it is hard to know you compared the right alternatives.
- $p(D|M_i)$ does not necessarily tell you how well the model will predict other data

Cross-validation

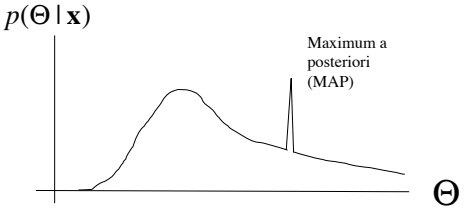
- Standard way to evaluate models
- Exclude a subset of the data while fitting model
- Compute predictions for the held-out subset.
- Evaluate predictions against actual held-out values
  - e.g., distance from truth, or class labels
- If you use k such sets, this is called k-fold cross-validation
- If you leave out 1 data point, it is called leave-one-out.

Cross-validation (2)

- Cross-validation provides
  - A way to choose models
  - A way to measure performance
  - A way to measure generalization capacity
- Held out data **must be different enough** to test the level of generality that you want
  - Consider degree of validation in a model to predict happiness
    1. How happy are you now given recent data points
    2. How happy are you now given all data points
    3. How happy are you on day X given data for other days
    4. How happy are you based on model of **other** people
    5. How happy are you based on **other** people in other experiments
    6. How happy are you based on modeling people in other cultures

More on estimation

- If the goal is to provide the model, then we often estimate the MAP value for the parameters
- This assumes that the posterior is nicely behaved
- An alternative is to average some or all (MMSE) of the posterior.



Classification

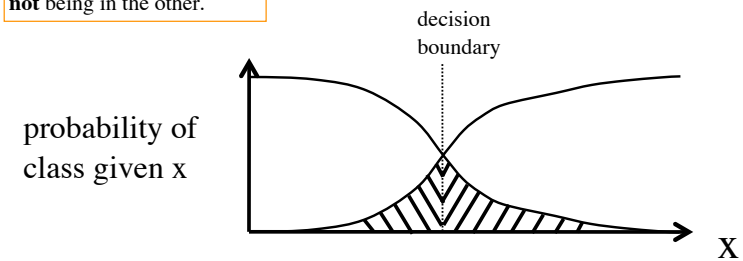
- Consider that our parameters include a discrete class variable, c.
- Assume no other variables, or that they have been marginalized out.
- Use x for the data. Then the posterior over classes is

$$p(c \mid x) \propto p(c)p(x \mid c)$$

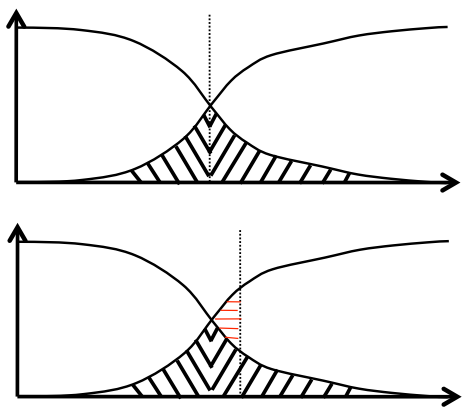
- So, given x, what is the class?

Classification

Binary case, easy to draw
Two classes, C_1 and C_2.
being in one is the same as **not** being in the other.



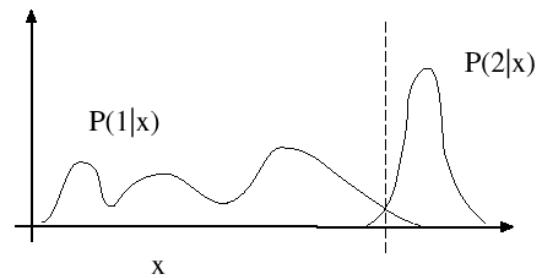
Area of intersection under curves gives expected value of making a mistake



Red shows extra that you get wrong with different boundary

Classification

Finding a decision boundary is not the same as modeling a conditional density.



Here there are more than two classes, but only two shown. Consider all animals, but you are being force to choose between “dog” and “cat”.

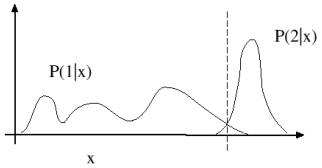
Classification

Finding a decision boundary is not the same as modeling a conditional density.

Working with the boundary might be easier (we don't care about the extra bumps).

But we loose any indication of whether the point is an outlier.

In this course we will not cover in detail finding the boundary (discriminative method).



Bishop §1.5

Decision making

Classification where the risk (loss) for each class is different.

Example: Risk of a false negative diagnosis is more than that for the risk of false positive diagnosis.

Define a loss function, Lj,k which tells us the loss of classifying a category k, as a category, j.

Example:

$$\begin{array}{c c} & \begin{array}{cc} \text{cancer} & \text{normal} \end{array} \\ \begin{array}{c} \text{cancer} \\ \text{normal} \end{array} & \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} \end{array}$$

Decision making

Now the classification boundaries for **x** are based on the loss, not just the probability.

You choice of the class, j, for x is the lowest expected loss.

This is found by:

$$\underset{j}{\operatorname{argmin}}\left\{\sum_k L_{k,j} \cdot p(C_k|x)\right\}$$

Decision making

Example to illustrate that the formula is sensible.

Suppose that at a given x*, we have

$p(C_1 \mid x^*) = 0.3 \quad p(C_2 \mid x^*) = 0.2 \quad p(C_3 \mid x^*) = 0.5$

Evaluate the assignment of x* under loss functions

$$\mathrm{L}_A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \qquad \mathrm{L}_B = \begin{pmatrix} 0 & 1 & 1 \\ 10 & 0 & 10 \\ 1 & 1 & 0 \end{pmatrix}$$

$p(C_1 \mid x^*) = 0.3 \quad p(C_2 \mid x^*) = 0.2 \quad p(C_3 \mid x^*) = 0.5$

For the first example (loss is misclassifcation rate)

$$L_B = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

Note that loss is defined for misclassifying the column item as the row item.

Declaring that x at x* is C_1 has expected loss: (0.3)*0 +(0.2)*1+ (0.5)*1=0.7

Declaring that x at x* is C_2 has expected loss: (0.3)*1 +(0.2)*0+ (0.5)*1=0.8

Declaring that x at x* is C_3 has expected loss: (0.3)*1 +(0.2)*1+ (0.5)*0=0.5

As expected, the minimum loss is for the likeliest class.

$p(C_1 \mid x^*) = 0.3 \quad p(C_2 \mid x^*) = 0.2 \quad p(C_3 \mid x^*) = 0.5$

For the second example

$$L_B = \begin{pmatrix} 0 & 1 & 1 \\ 10 & 0 & 10 \\ 1 & 1 & 0 \end{pmatrix}$$

Note that loss is defined for misclassifying the column item as the row item.

Declaring that x at x* is C_1 has expected loss: (0.3)*0 +(0.2)*10+ (0.5)*1 = 2.5

Declaring that x at x* is C_2 has expected loss: (0.3)*1 +(0.2)*0+ (0.5)*1 = 0.8

Declaring that x at x* is C_3 has expected loss: (0.3)*1 +(0.2)*10+ (0.5)*0 = 2.3

Now the heavy penalty for missing C_2 leads to C_2 being the best answer.
(Note that C_2 was the worst answer with the previous loss).

Graphical Models

Reference for much of the next topic is Chapter 8 of Bishop

Available on-line
http://research.microsoft.com/~cmbishop/PRML

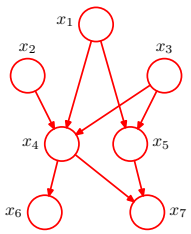
(Linked from course page).

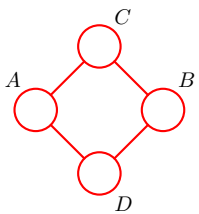
Graphical Models

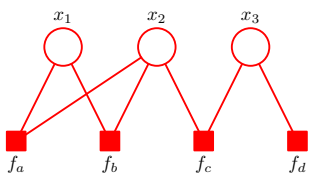
- Graphical representation of statistical models
- Nodes
  - Random variables (or groups of them)
- Edges
  - Probabilistic relationships between nodes

Graphical Models

- Various kinds
  - Directed (Bayesian networks)
  - Undirected (e.g., Markov random field)
  - Factor graphs (different representation, applicable to both)

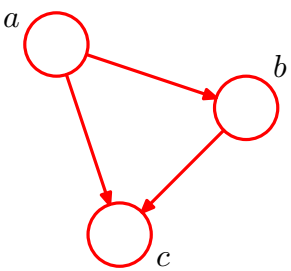






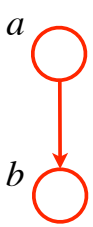
Directed Graphical Models

- Nodes represent random variables
- Edges between nodes have directed links
- No cycles



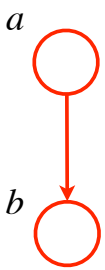
Directed Graphical Models

- Nodes represent random variables
- Edges between nodes have directed links
- No cycles
- The graph represents a **factorization** of the joint probability of all the random variables represented by the nodes.



  - An arrow from one node (a) to another one (b) means that the second node (b) is conditioned on the first (a).
  - In other words, if you have information about (a), then you have information about (b).
  - Thus the arrows tell you about information flow.

Directed Graphical Models



Here we have two nodes, a and b.

So this is a representation of the joint distribution $p(a, b)$.

In particular, it is equivalent to writing

$$p(a, b) = p(b \mid a)\, p(a)$$

Ancestral sampling version of the story:

To sample from $p(a, b)$

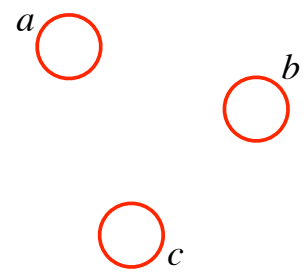
First sample $\tilde{a}$ from $p(a)$

Then sample $\tilde{b}$ from $p(b \mid \tilde{a})$

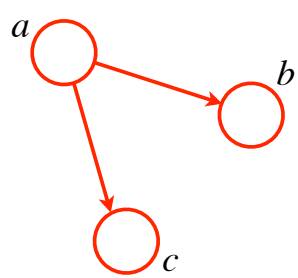
Directed Graphical Models

- A story of three random variables .... a, b, and c.
- General model is p(a,b,c) (understand this!)
- What are possible relationships of a, b, and c?
  - Independence: p(a,b,c)=p(a)p(b)p(c)
  - Some structure: e.g., p(a,b,c)=p(a)p(b|a)p(c|a)
  - Arbitrary relationship

p(a,b,c) = p(a)p(b)p(c)



p(a,b,c) = p(a)p(b|a)p(c|a)



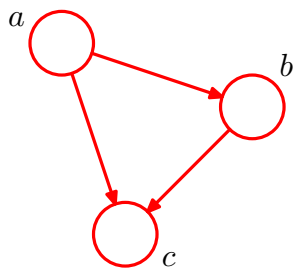
p(a,b,c) with no identified independence

$$p(a,b,c) = p(a)p(b \mid a)p(c \mid a,b)$$

$$p(a,b,c) = p(b)p(c \mid b)p(a \mid c,b)$$

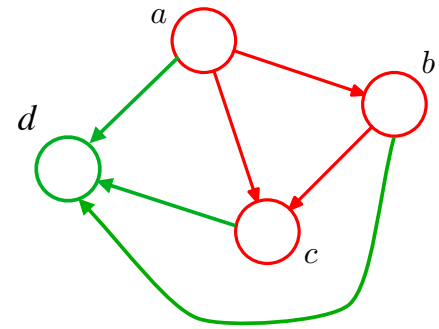
• • •

p(a,b,c) = p(a)p(b|a)p(c|a,b)



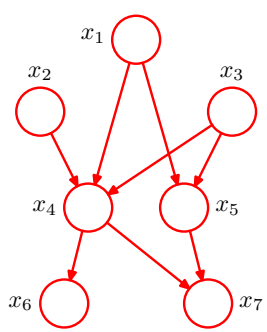
Note that the graph is fully connected

p(a,b,c,d) = p(d | a,b,c) p(a,b,c)



Note that the graph is fully connected

Another example (§8.2 in Bishop)



What is the algebraic form?

Univariate Gaussian with known variance (§8.2 in Bishop)

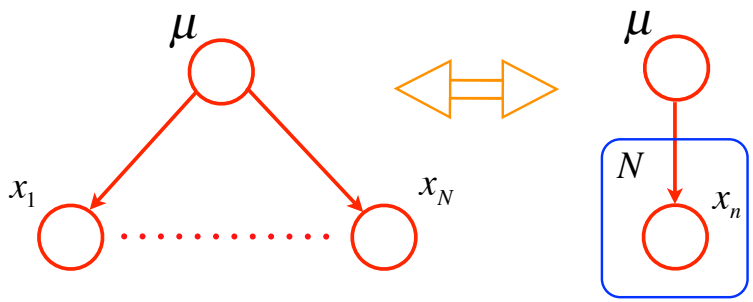
$$D = \{x_1, x_2, x_3, \ldots, x_N\}$$

$$p(D, \mu) = p(\mu) \prod_{n=1}^{N} p(x_n \mid \mu)$$

where

$$p(x_n \mid \mu) = \mathbb{N}(x_n \mid \mu; \sigma^2)$$

Univariate Gaussian with known variance (§8.2 in Bishop)

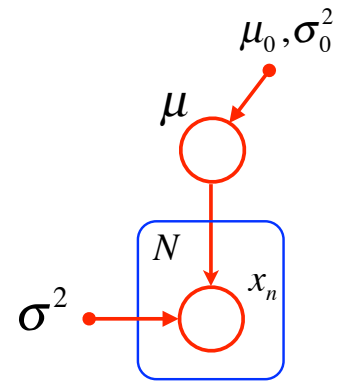


More compact notation
(plate representation)

Deterministic parameters

Our univariate Gaussian has some known parameters: the variance and the prior on the mean.

If we wish to illustrate them, we use a small filled in circle.



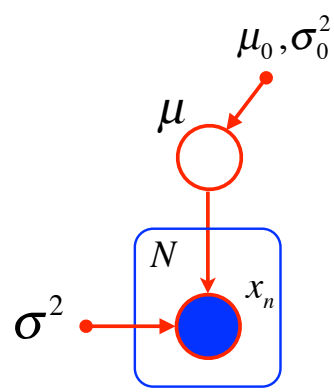
Observed variables

We indicate observed variables by shading them

Alternatively, this indicates conditioning

Observed variables

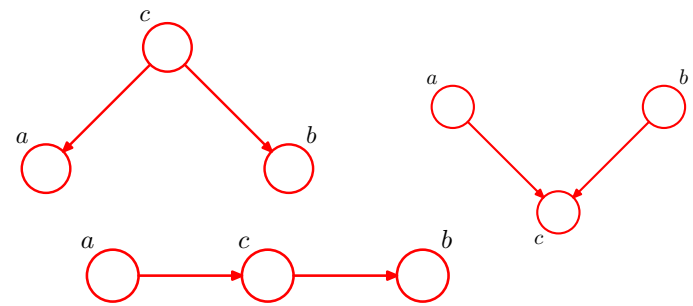
Example: Inferring the mean of the univariate



Back to three variables

What are the possible Bayes nets with three variables?

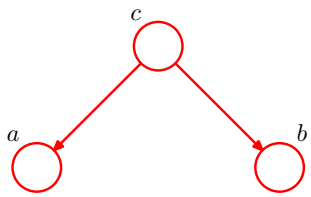
Three interesting cases



For each case, consider two questions:

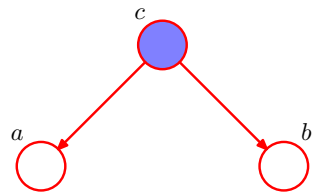
1) Is $a \perp b$?
2) Is $a \perp b \mid c$? (i.e. c is observed)

Case one



Is $a \perp b$?

Case one where c is observed



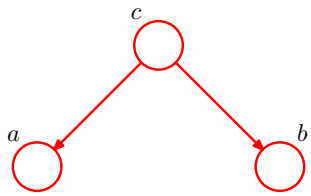
$a \perp b \mid c$

$p(a,b,c) = p(c)p(a|c)p(b|c)$ (what the graph represents in general)

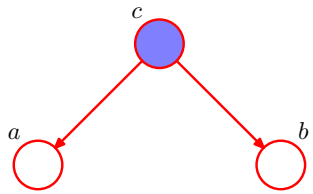
$p(a,b|c) = p(a|c)p(b|c)$ (with c observed)

This is the definition of $a \perp b|c$

Case one (tail-to-tail) summary



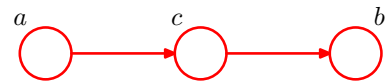
$a \not\perp b$



$a \perp b \mid c$

Tail-to-tail case
- With no conditioning, no independence
- With conditioning, we have independence

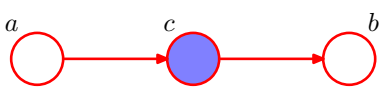
Case two



Is $a \perp b$?

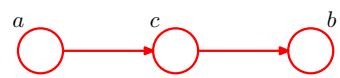
The graph represents $p(a,b,c) = p(a)\,p(c|a)\,p(b|c)$

Case two where *c* is observed

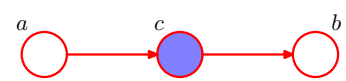


Is $a \perp b \mid c$?

Case two (head-to-tail) summary



$a \not\perp b$

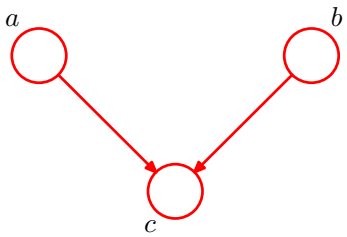


$a \perp b \mid c$

Head-to-tail case
- With no conditioning, no independence
- With conditioning, we have independence

(Same as case one)

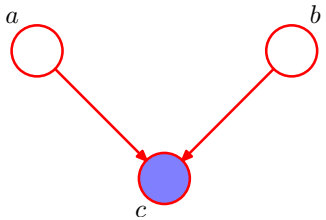
Case three



Is $a \perp b$?

Example:
- c == “strange noises at night”
- a == “burglar in the house”
- b == “deer in the back yard”

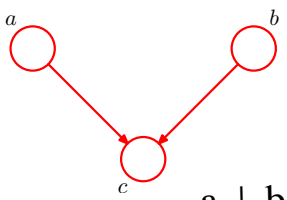
Case three with *c* observed



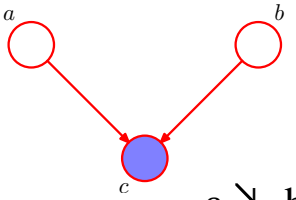
Is $a \perp b \mid c$?

Recall our example:
- c == “strange noises at night”
- a == “burglar in the house”
- b == “deer in the back yard”

Case three (head-to-head) summary



$a \perp b$



$a \not\perp b \mid c$

Head-to-head case (different than the other two)
- With no conditioning, we have independence
- With conditioning, we do not have independence

If you are having trouble with “explaining away”, please study Bishop, chapter 8, pages 378-379 (on-line).

Three random variables summary

In cases one and two, *a* and *b* were not independent until the observation of *c* “blocked” the (connection) path from *a* to *b*.

(From Koller and Friedman, a path that is not blocked is “active”)

In case three, if *c* is not observed, the path is blocked. Observing *c* made the connection (path) active.

d-Separation (Pearl, 88)

"d" stands for "directed"

Generalizes the examples we have been studying.

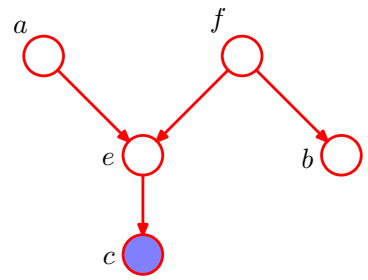
Consider non-overlapping subsets A, B, C of nodes of a graph.

Consider all paths from nodes in A to nodes in B.

A path is blocked if either:
a) The arrows meet either tail-to-tail or head-to-tail at a node in C.
b) The arrows meet head-to-head at some node that is not in C, nor are any of its descendants in C.

If all paths are blocked, then A and B are independent given C.

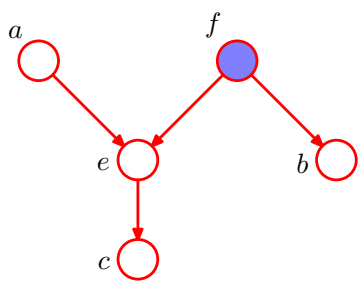
d-Separation (example one)



(A={a}, B={b}, and C={c})

Does this graph encode $A \perp B | C$?

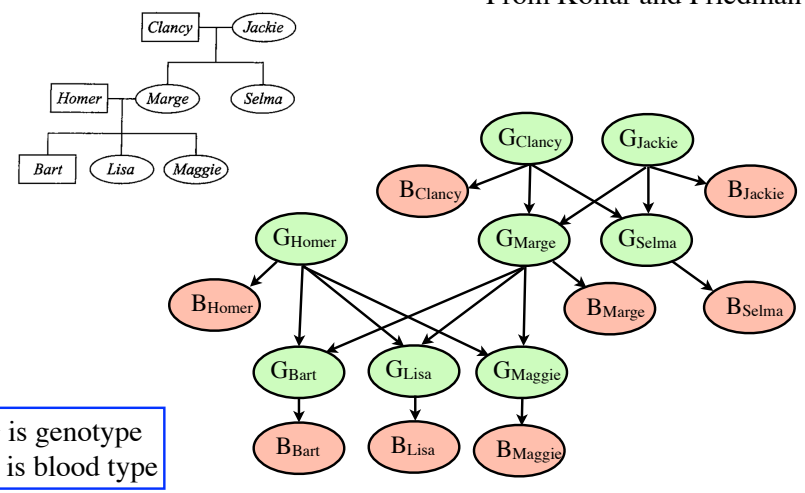
d-Separation (example two)



Does this graph encode $A \perp B | F$?

Grounded example of a Bayesian Network

From Kollar and Friedman



G is genotype
B is blood type

Bayesian network semantics

- Represents a factorization of p()
  - Random variables are nodes
  - Factors are CPD (conditional probability distributions) for child given parent (just p(`NODE`) if no parents).

Equivalent semantic specification (Proof is in K&F, ch. 3)

- For each X_i : $X_i \perp \text{NonDescendents}(X_i) \mid \text{Parents}(X_i)$
  - Notice no mention of factorization

Conditional independence in distributions and graphs

Let $I(P)$ be the set of independence assertions of the form $(X \perp Y | Z)$ that are true for a distribution P.

Let $I(G)$ be the set of independence assertions represented by a DAG, G.

G is an I-map for P if I(G) $\subseteq I(P)$

In other words, all independance represented in G are true. (There could be some more in P that G does not reveal).

A few notes on notation and independence

We sometimes write $(A \perp B | \varnothing)$ for $A \perp B$

Also, we write $(A \perp B, C | X)$ for $(A \perp B | X)$ and $(A \perp C | X)$

Recall that $(A \perp B | C)$ means that $P(A | B, C) = P(A | C)$

This generalizes to:

$$(A \perp B | \ldots, C, \ldots) \Rightarrow P(A | \ldots, B, C, \ldots) = P(A | \ldots, C, \ldots)$$

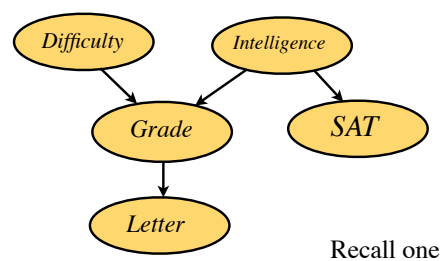
Example going from I-map to a factorization

From Kollar and Friedman

For $P(I,D,G,L,S)$, the $I(Graph)$ tells us

$(D \perp I | \varnothing)$ $(D \perp I | S)$ $(L \perp I, D, S | G)$ $(G \perp S | I, D)$ $(S \perp D, G, L | I)$

(Note that this is not necessarily all relationships that we can extract)



Recall one version of DAG semantics is

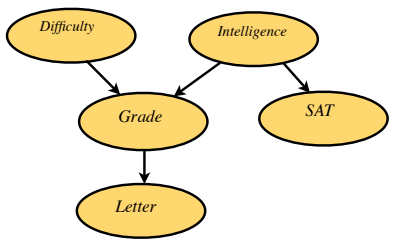
$X_i \perp \text{NonDescendents}(X_i) \mid \text{Parents}(X_i)$

We can write the joint distribution as conditioning on non-descendents if we maintain a sensible "lexigraphical order where parents occur before children.

$$P(I,D,G,L,S) = P(I)P(D|I)P(G|I,D)P(L|I,D,G)P(S|I,D,G,L)$$

This means that for each factor, all variables conditioned on are either the parents, or non-descendents.

This means that for each factor, we may have rule that gets rid of some non-descendents.



Example going from I-map to a factorization

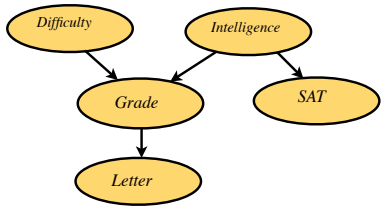
$$P(I,D,G,L,S) = P(I)P(D|I)P(G|I,D)P(L|I,D,G)P(S|I,D,G,L)$$

$$(D \perp I|\varnothing) \Rightarrow P(D|I) = P(D)$$

$$(L \perp I,D,S|G) \Rightarrow P(L|I,D,G) = P(L|G)$$

$$(S \perp D,G,L|I) \Rightarrow P(S|I,D,G,L) = P(S|I)$$

So, $P(I,D,G,L,S) = P(I)P(D)P(G|I,D)P(L|G)P(S|I)$



Summary on the equivalence of the two interpretations of directed graphical models

Factorization semantics

Factors are p(node | parents)

Abstract semantics

$$X_i \perp \text{NonDescendents}(X_i) \mid \text{Parents}(X_i)$$

These are equivalent

Proof of one direction by the one example just completed.

Interesting questions

- Does every probability distribution have a corresponding Bayesian network?

  Chain rule says yes

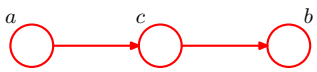
- Given the independence structure of a probability distribution, and a graph that captures them all (I(G)=I(P), is the corresponding graph unique (ignoring isomorphisms)?

  Case study of three nodes says no

- Do our graphs faithfully capture the independence structure of our distributions?

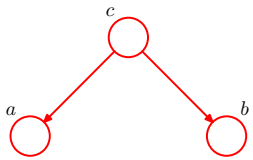
  TBA

Back to case one



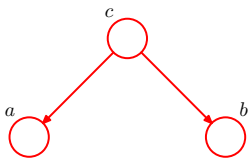
- Let a="smokes", c="high blood pressure", b="stroke"
- p(c|a) tells you probability of having high blood pressure if you smoke (for some definition of each).

Can we distinguish case two from case one?



- Let a="smokes", c="high blood pressure", b="stroke"
- p(a|c) tells you probability of being a smoker if you have high blood pressure (for some definition of each).

Can we distinguish case two from case one?



- Let a="smokes", b="high blood pressure", c="stroke"
- p(a|c) tells you probability of being a smoker if you have high blood pressure (for some definition of each).
- Data for estimating p(c|a) in first case, and p(a|c) in second case cannot tell you which model you should prefer.
  - "Correlation is not causation"
- Causality implied by our generative process is about the statistics of the data, not physical causality.

More on causality

- References
  - Kollar and Friedman, Chapter 21 which starts on page 1009!
  - Classic book by Pearl, Causality: Models, Reasoning, and Inference, 2000
    - A version is available on-line (bayes.cs.ucla.edu/BOOK-99/book-toc.html)

More on causality

- We have been focussed on the joint distribution which is adequate (arguably optimal) for answering the queries we have studied
- In particular, we know how distributions over unknowns change due to evidence
- For many problems (e.g., computer vision and much of machine learning) this is sufficient
  - Either causes are obvious or not relevant

More on causality

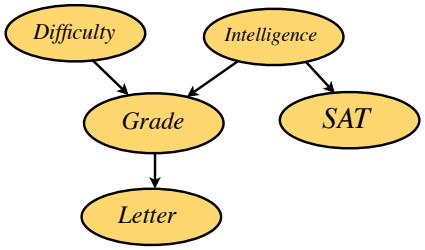
- Two correlated variables can have multiple equivalent graphs hinting at **different** causal stories able to provide the **same** joint.
  - A causes B
  - B causes A
  - C causes both A and B
  - A and B cause C (and A and B are correlated by explaining away)
- Given a choice, we prefer the Bayes net that also represents our causal theory (if we have one)
  - More natural, easier to understand
  - Helps tell you whether observed statistics are consistent with your theory
    - (Covered briefly next)

Intervention

- Two Bayes nets that give the same joint distribution can differ in what they say about an intervention.
- We represent an intervention, x, as setting some subset of the variables, X, to the value, x, denoted by $do(X=x)$.
  - Example 1: Creating an experimental group that will not smoke
  - Example 2: Setting your grade to A by hacking into a computer
- On the surface, this might look like conditioning on X, but it is different --- the graph needs to change also
  - We need to "mutilate" the graph

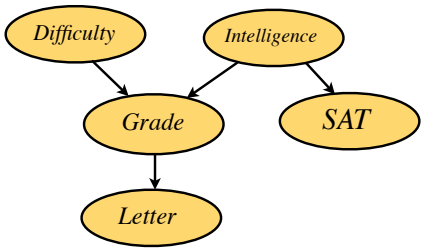
Representing Intervention

- Example one (students and grades, again)
  - Does observing grade change your belief about SAT?



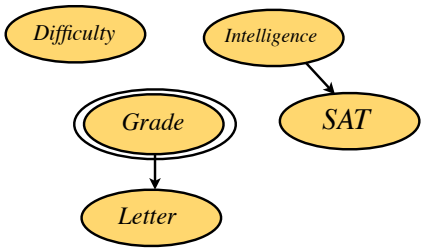
Representing Intervention

- Example one (students and grades, again)
  - Does observing grade change your belief about SAT?
- Now, suppose we intervene on the *Grade* random variable
  - E.G., we fix it by hacking into the grade computer
  - Now does observing grade change your belief about SAT?



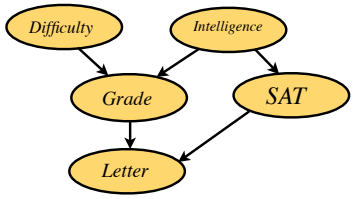
Representing Intervention

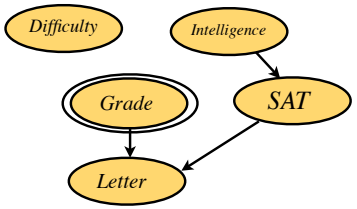
- The intervention not only conditions on the variable, it cuts the links that influence it. This is the mutilated graph.



Representing Intervention

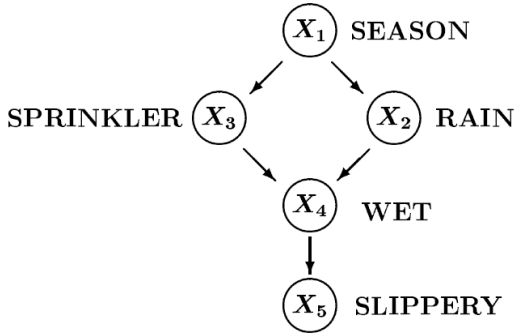
- Another example --- the student from before with a link between SAT and letter. Now we expect that the intervention does not entirely explain the letter, but that the influence of *grade* is direct (only).





Representing Intervention

- Another example --- from Pearl, 2000.
  - Consider the intervention of turning the sprinkler "on"



Representing Intervention

- Representation of the intervention of turning the sprinkler on.

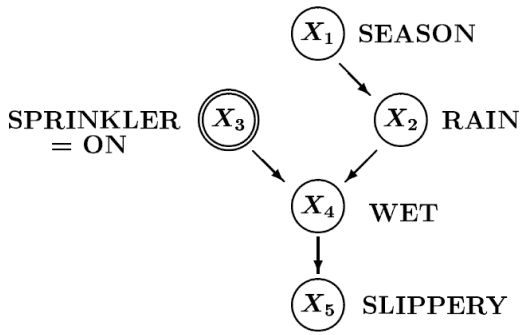


Figure 1.4: Network representation of the action "turning the sprinkler On."

Can graphs capture all independence?

- Do our graphs faithfully capture the independence structure of our distributions?
- Recall that

  G is an I-map for P if $I(G) \subseteq I(P)$

  In other words, all independence represented in G are true.
  (There could be more independence in P that G does not reveal).

- Hence we are asking if $I(G) \equiv I(P)$
  Since $I(G) \subseteq I(P)$ this amounts to asking if $I(P) \subseteq I(G)$

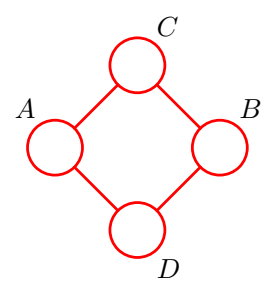
Perfection

G is an P-map for P if $I(G) \equiv I(P)$ (perfect map)

In other words, all independence represented in G are true, and there are no other independence relations.

Do all distributions have perfect maps?

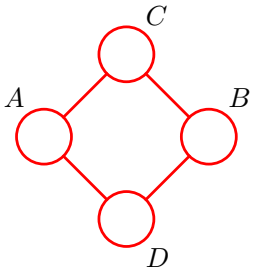
Perfection may not be attainable



The "misconception" example in K&F (pp. 82-3), where Alice, Bob, Charles, and Debbie study in pairs shown, but A and B never work together, nor do C and D.

Note **no arrows**, but a link still means some probabilistic relation.

Perfection may not be attainable



Suppose that we have

$$(A \perp B | C, D)$$

and

$$(C \perp D | A, B)$$

Now, draw the Bayes net (have fun!).

Note **no arrows**, but a link still means some probabilistic relation.

Interesting questions

- Does every probability distribution have a corresponding Bayesian network?

  Chain rule says yes

- Given the independence structure of a probability distribution, and a graph that captures them all (I(G)=I(P), is the corresponding graph unique (ignoring isomorphisms)?

  Case study of three nodes says no

- Do our graphs **always** faithfully capture the independence structure of our distributions?

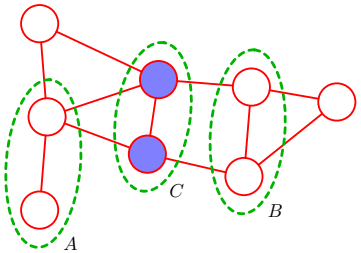
  Misconception example says no

Undirected graphical models

- Also referred to as
  - Markov Networks
  - Markov Random Fields
- Nodes represent (groups of) random variables
- Edges represent probabilistic relations between connected nodes.
- We have already seen an example suggestive that arrows are not always helpful.

Undirected graphical models

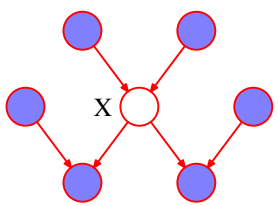
- The analog to d-separation is simper
  - Disjoint sets A and B are independent conditioned on C if all paths from nodes in A to nodes in B pass through C.

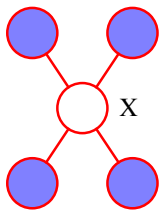


Here $(A \perp B | C)$ for all probability distributions represented by this graph.

Markov Blanket

- The Markov blanket of a node, X, is a particular set of (nearby) nodes B where $X \perp X_i | B$ for all X_i
- For directed graphs the Markov blanket is the parents, children, and co-parents of X.
- For undirected graphs this is simply the set of nodes connected to X.





Undirected graphical models

- Bayes nets where nodes only have one parent are easily converted to undirected graphs without changing links.
- (Discussed in more detail soon)

Semantics of undirected graphical models

- Intuitively, for any two nodes, x_i and x_j, not connected by a link, $x_i \perp x_j | \mathbf{x}/\{i,j\}$.

Draw this on the board!

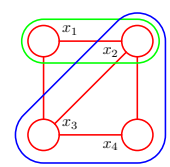
- So, $p(\ldots,x_i,\ldots,x_j,\ldots) = p(x_i|\mathbf{x}/\{i,j\})\,p(x_j|\mathbf{x}/\{i,j\})\,p(\mathbf{x}/\{i,j\})$
- This suggests that an appropriate factorization should not have factors with these two nodes together.
- Direct links imply that we have a relation, and so we cannot put directly linked nodes into the same factor.
- A group of nodes that are all connected cannot be factored by the above rule.

Semantics of undirected graphical models

- So, we add nodes into factors, provided that they are all connected.

Draw this on the board!

- This leads to describing the semantics in terms of maximal cliques.
  - A clique is fully connected subset of nodes from the graph
  - A maximal clique is a clique where no node in the graph can be added to it without it ceasing to be a clique.



All parwise linked nodes are cliques. For example $\{x_1,x_2\}$ is a clique (green). However, it is not a maximal clique. $\{x_2,x_3,x_4\}$ is a maximal clique (blue). If we add another no no longer have a clique.

Examples on the board!

Semantics of undirected graphical models (2)

Let C index maximal cliques. Then

$$p(\mathrm{x}) = \frac{1}{Z}\prod_c \psi_C(\mathrm{x}_C)$$

where $Z = \sum_x \prod_c \psi_C(\mathrm{x}_C)$ (or $\int_x \prod_c \psi_C(\mathrm{x}_C)$) is the partition function,

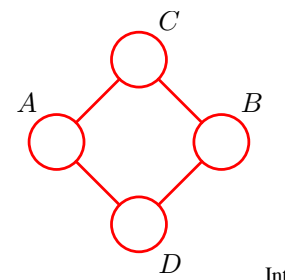
and $\psi_C(\mathrm{x}_C)$ are the clique potentials.

If x_i and x_j do not share an edge, then they do not share cliques.

So $p(\mathrm{x}) = \frac{1}{Z}\prod_{c(i)} \psi_C(\mathrm{x}_C) \prod_{c(j)} \psi_C(\mathrm{x}_C) \prod_{c \notin c(i) \cup c(j)} \psi_C(\mathrm{x}_C)$

Draw on the board.

Misconception example



$$p(A,B,C,D) \propto \psi(A,C)\psi(C,B)\psi(B,D)\psi(D,A)$$

Intuitively we have $(A \perp B | C,D)$ and $(C \perp D | A,B)$ because if C,D are fixed, then factors for A and B have no shared variables.

However, let us derive a result to confirm this

$p(X,Y,Z) = \varphi(X,Z)\varphi(Y,Z) \quad \Leftrightarrow \quad X \perp Y | Z$ (algebra for $\Rightarrow$, other direction is easier)

$$p(X|Z)p(Y|Z) = \frac{\sum_X \varphi(X,Z)\varphi(Y,Z) \sum_Y \varphi(X,Z)\varphi(Y,Z)}{\sum_{X,Y} \varphi(X,Z)\varphi(Y,Z) \sum_{X,Y} \varphi(X,Z)\varphi(Y,Z)}$$

$$= \frac{\varphi(Y,Z)\sum_X \varphi(X,Z)}{\sum_X \varphi(X,Z)\sum_Y \varphi(Y,Z)} \frac{\varphi(X,Z)\sum_Y \varphi(Y,Z)}{\sum_X \varphi(X,Z)\sum_Y \varphi(Y,Z)}$$

$$= \frac{\varphi(Y,Z)\sum_X \varphi(X,Z)}{\left(\sum_X \varphi(X,Z)\right)\left(\sum_Y \varphi(Y,Z)\right)} \frac{\varphi(X,Z)\sum_Y \varphi(Y,Z)}{\left(\sum_X \varphi(X,Z)\right)\left(\sum_Y \varphi(Y,Z)\right)}$$

$$= \frac{\varphi(Y,Z)}{\left(\sum_Y \varphi(Y,Z)\right)} \frac{\varphi(X,Z)}{\left(\sum_X \varphi(X,Z)\right)}$$ (canceling green and red pairs)

$$= \frac{\varphi(Y,Z)\varphi(X,Z)}{\sum_{X,Y} \varphi(Y,Z)\varphi(X,Z)}$$

$$= \frac{p(X,Y,Z)}{p(Z)}$$

From directed to undirected

- Easy case (all nodes have at most one parent).
- Example:

- Becomes:

From directed to undirected

- Convert:

$$p(\mathrm{x}) = p(x_1)\,p(x_2|x_1)\,p(x_3|x_2)\;\cdots\;p(x_{N-1}|x_{N-2})\,p(x_N|x_{N-1})$$

- To:

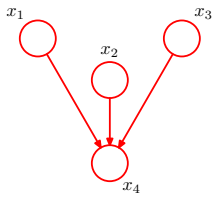
$$p(\mathrm{x}) = \Psi(x_1,x_2)\,\Psi(x_2,x_3)\;\ldots\;\Psi(x_{N-2},x_{N-1})\,\Psi(x_{N-1},x_N)$$

- Inspection suggests:

$$\Psi(x_1,x_2) = p(x_1)\,p(x_2|x_1)$$
$$\Psi(x_2,x_3) = p(x_3|x_2)$$
$$\cdots$$
$$\Psi(x_{N-2},x_{N-1}) = p(x_{N-1}|x_{N-2})$$
$$\Psi(x_{N-1},x_N) = p(x_N|x_{N-1})$$

From directed to undirected

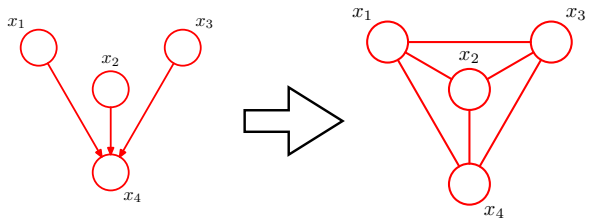
- Harder case (some nodes have multiple parents).
- Example:



- Because this implies conditioning on three variables, the potentials for the clique are a function of four variables.
- These nodes need to be part of a clique (but they are not).

From directed to undirected

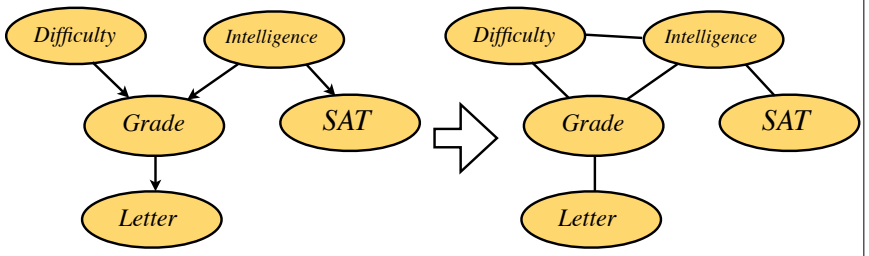
- Solution is to marry the parents.
- This makes the graph "moral".
- Note that moralization looses conditional independence information.



From directed to undirected

- Complete algorithm
  - Make the graph moral.
  - Initialize each maximal clique potential to one.
  - Multiply each factor in p() into an appropriate clique potential.
  - Note that Z=1

Example of converting directed to undirected



$$P(I,D,G,L,S) = P(I)P(D)P(G|I,D)P(L|G)P(S|I)$$

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

$$\psi(D,G,I) = P(I)P(D)P(G|I,D) \quad \psi(S,I) = P(S|I) \quad \psi(L,G) = P(L|G)$$

$$\psi(D,G,I) = P(D)P(G|I,D) \quad \psi(S,I) = P(I)P(S|I) \quad \psi(L,G) = P(L|G)$$

Energy function encoding

We will assume that all $\psi_C(\mathrm{x}_C) > 0$.

In general, we leave the semantics of $\psi_C(\mathrm{x}_C)$ open, but for undirected graphs that come from directed graphs where each node has one parent, the semantics follows that for the directed graphs (as we have just done).

Since $\psi_C(\mathrm{x}_C) > 0$ we will often write $\psi_C(\mathrm{x}_C) = \exp\{-E(\mathrm{x}_C)\}$ where E() is the energy function.

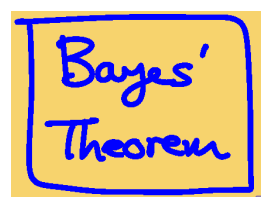
Energy function encoding (2)

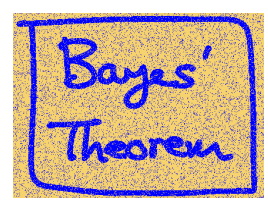
Writing $\psi_C(\mathrm{x}_C) = \exp\{-E(\mathrm{x}_C)\}$ means that

$$\begin{aligned} p(x) &= \frac{1}{Z}\prod_c \psi_x(\mathrm{x}_C) \\ &= \frac{1}{Z}\prod_c \exp\{-E(\mathrm{x}_C)\} \\ &= \frac{1}{Z}\exp\left\{\sum_c -E(\mathrm{x}_C)\right\} \\ &= \frac{1}{Z}\exp\{-E(x)\} \qquad \text{Where} \quad E(x) = \sum_c E(\mathrm{x}_C) \end{aligned}$$

Example of a Markov random field

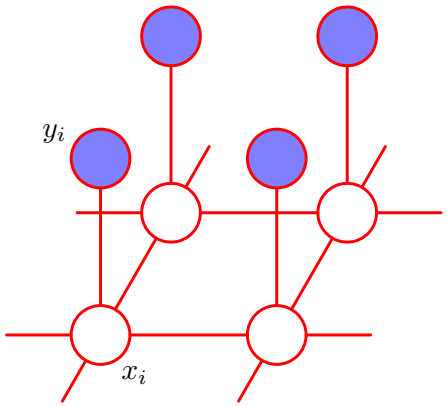
- Consider a binary image (pixels are either black or white).
- Pixels are represented by {-1,1}.
- Suppose the image have is an underlying accurate image where some of the bits have been flipped by a noise process.





Example of a Markov random field (2)

- Undirected graphical model.



Example of a Markov random field (2)

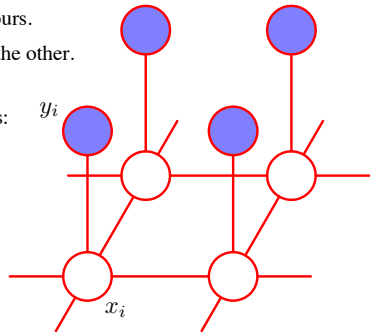
- For low energy (high probability)

$x_i = y_i$ most of the time (set by noise level)
$x_i = x_j$ most of the time if i and j are neighbours.
x_i could be biased to have one value or the other.

A simple energy function for the entire grid is:

$$E(\mathbf{x},\mathbf{y}) = h\sum x_i - \beta\sum_{i,j} x_i x_j - \eta\sum_i x_i y_y$$

Because values are 1 and -1, being the same makes the sums bigger, being different makes them smaller.



Example of a Markov random field (3)

$x_i = y_i$ most of the time (set by noise level)
$x_i = x_j$ most of the time if i and j are neighbours.
x_i could be biased to have one value or the other.

Additional details glossed over in class provided in notes.

For each $\{x_i, y_i\}$ maximum clique, $\mathrm{E}(x_i, y_i) = -\eta \bullet x_i \bullet y_i \quad (\eta > 0)$
(high probablity corresponds to low energy)

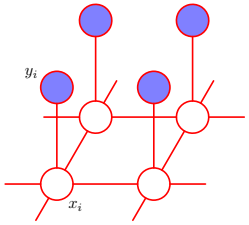
For unique $\{x_i, x_{j \in neighbor(i)}\}$ max clique, $\mathrm{E}(x_i, x_j) = -\beta \bullet x_i \bullet x_j \quad (\beta > 0)$

For a subset of the above cliques, one for each i, add in a term $h \bullet x_i$.

Example of a Markov random field (4)

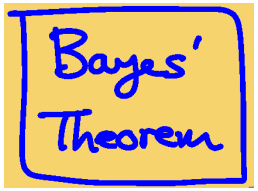
- Notice in the previous analysis we assigned arguably symmetric cliques different potentials
  - Left boundary x_i might get different potentials than right boundary x_i.
  - Some x_{ij} get a factor for the bias, other do not.
- Notice that exact assignment to clique potentials may not matter
- We can jump readily quickly to the overall picture, hence:

$$E(\mathbf{x},\mathbf{y}) = h\sum x_i - \beta\sum_{i,j} x_i x_j - \eta\sum_i x_i y_y$$

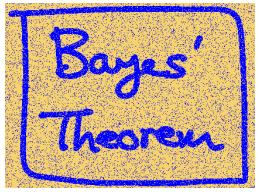


Example of a Markov random field (3)

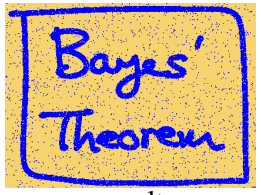
- Finding a low energy (high probability) state using ICM (iterated conditional modes).
  - Initialize x_i to y_i.
  - For each i, change x_i if energy decreases.
  - Repeat until energy no longer can be decreased.
- Converges to a local minimum because we only decrease.



original

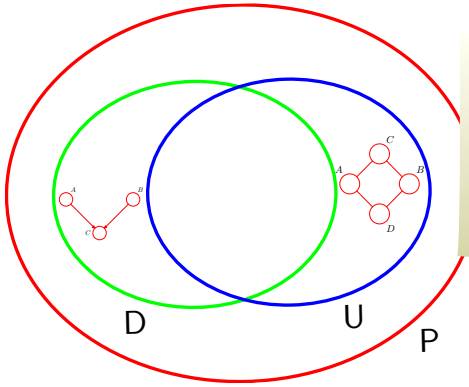


with noise



result

Directed and undirected perfect maps



Here we are in the first few slides of lecture 13.

D is subset of distributions in P that are perfectly represented by directed graphs; similarly U for undirected graphs.