EECS 445

Introduction to Machine Learning

Learning Bayesian Networks

Prof. Kutty

Announcements

Course evaluations are out

- Gradescope assignment to upload proof (screenshot)
  - **Please note *separate*** eval and assignment deadlines!!!
    - deadline for the assignment is *different* from the registrar's deadline
- worth 0.5% of your grade!

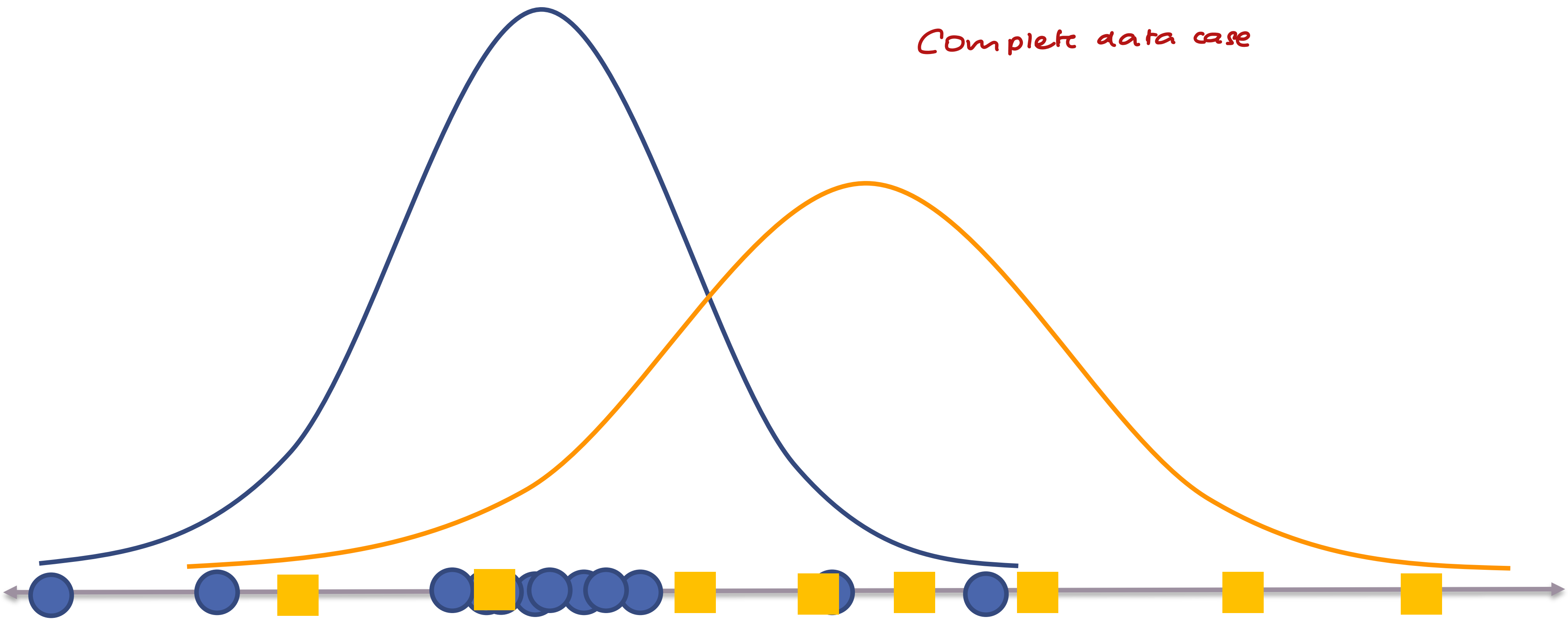
HW4 due tomorrow → please make sure you check latedays

Sample exam released on Friday:

Review on Monday 4/22 at 6:30pm → includes sample exam solutions review as well as material review

MLE of GMM with *known* labels: intuition

Complete data case



can also determine relative chance of each Gaussian

Log-Likelihood for GMMs with known labels

Given the training data, find the model parameters that maximize the **log-likelihood**

iid

$$P(S_n) = \prod_{i=1}^{n} p(\bar{x}^{(i)}, y^{(i)})$$

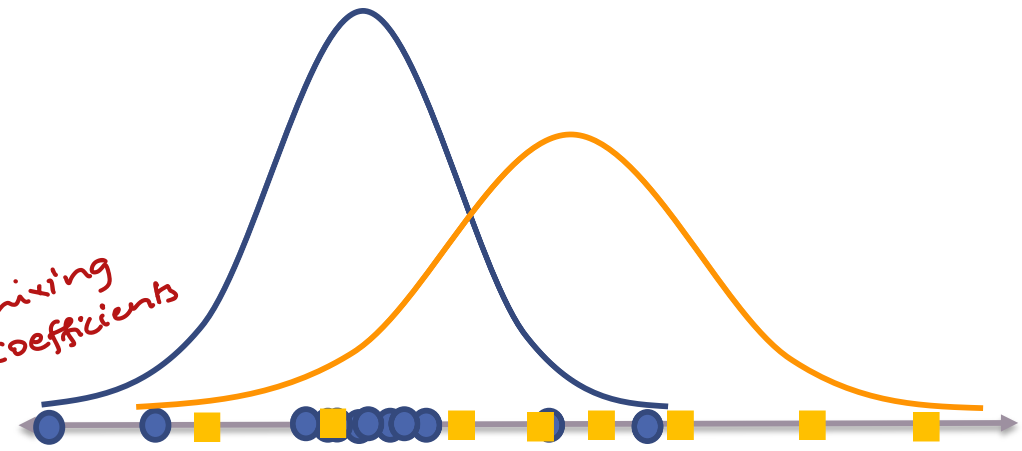
product rule

$$= \prod_{i=1}^{n} p(\bar{x}^{(i)} | y^{(i)}) p(y^{(i)})$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{k} \delta(j \mid i) (N(\bar{x}^{(i)} | \bar{\mu}^{(j)}, \sigma_j^2) \gamma_j)$$

indicator function

mixing coefficients



Maximum log likelihood objective

$$\ln P(S_n) = \ln \prod_{i=1}^{n} \sum_{j=1}^{k} \delta(j \mid i) (N(\bar{x}^{(i)} | \bar{\mu}^{(j)}, \sigma_j^2) \gamma_j)$$

$$= \sum_{i=1}^{n} \ln \sum_{j=1}^{k} \delta(j|i) N(\bar{x}^{(i)} | \bar{\mu}^{(j)}, \sigma_j^2) \gamma_j$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} \delta(j \mid i) \ln (\gamma_j N(\bar{x}^{(i)} | \bar{\mu}^{(j)}, \sigma_j^2))$$

Expectation Maximization for GMMs

(recap)

Expectation Maximization for GMMs: overview

For a mixture of k - spherical Gaussian

$$\bar{\theta} = [\underbrace{\gamma_1, \ldots, \gamma_k}_{\text{sums to } 1}, \bar{\mu}^{(1)}, \ldots, \bar{\mu}^{(k)}, \sigma_1^2, \ldots, \sigma_k^2]$$

mixing coefficients

Initialize model parameters

Iterate until convergence

– E step: use current estimate of mixture model to softly assign examples to clusters

– M step: re-estimate each cluster model separately based on the points assigned to it (similar to the "known label" case)

EM Algorithm in general

EM algorithm idea → Complete Data is easy

- Complete log likelihood

$$l_C(\bar{\theta}; X, Z) = \sum_{i=1}^{n} \log p(\bar{x}^{(i)}, z^{(i)}; \bar{\theta})$$

- Usually simpler to solve: find $\bar{\theta}$ by maximizing $l_C(\bar{\theta})$

- Issue?
  - $z^{(i)}$ is unknown in general

EM algorithm idea → Incomplete Data is hard

- Observed data X; Latent variables Z

$Z \rightarrow$ discrete r.v.

- log likelihood

$$l(\bar{\theta}; X) \overset{iid}{=} \sum_{i=1}^{n} \log p(\bar{x}^{(i)}; \bar{\theta}) \overset{\text{sum rule}}{=} \sum_{i=1}^{n} \log \sum_{z^{(i)}} p(\bar{x}^{(i)}, z^{(i)}; \bar{\theta})$$

- **Goal**: find

$$\arg\max_{\bar{\theta}} l(\bar{\theta}; X)$$

- **Issues**:
  - usually hard to find a closed form solution
  - usually non-concave --> many local optima

EM algorithm, incomplete data

- initialize parameters

p(j|i)

- **E-step**: compute posterior distribution according to current estimate of $\bar{\theta}$:

$$p\left(z^{(i)} = j \middle| \bar{x}^{(i)}; \bar{\theta}\right) \propto p\left(\bar{x}^{(i)} \middle| z^{(i)} = j; \bar{\theta}\right) p\left(z^{(i)} = j; \bar{\theta}\right)$$

- **M-Step**: pick parameters that maximize the *expected* log likelihood:

linearity of expectation

$$\arg\max_{\bar{\theta}} \mathbb{E}\left[\sum_{i=1}^{n} \log p\left(\bar{x}^{(i)}, z^{(i)}; \bar{\theta}\right)\right] = \arg\max_{\bar{\theta}} \sum_{i=1}^{n} \mathbb{E}\left[\log p\left(\bar{x}^{(i)}, z^{(i)}; \bar{\theta}\right)\right]$$

$$= \arg\max_{\bar{\theta}} \sum_{i=1}^{n} \sum_{j=1}^{k} p\left(z^{(i)} = j \middle| \bar{x}^{(i)} \cancel{; \bar{\theta}}\right) \log p\left(\bar{x}^{(i)}, z^{(i)} = j; \bar{\theta}\right)$$

Bayes rule:

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

- Iterate until convergence

$$P(A,B) = P(A|B)\,P(B)$$
$$\parallel$$
$$P(B,A) = P(B|A)\,P(A)$$

$$P(A|B)\,P(B) = P(B|A)\,P(A)$$

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}$$

$$P(B|A) \propto P(A|B)\,P(B)$$

$$B \rightarrow z^{(i)}$$
$$A \rightarrow \bar{x}^{(i)}$$

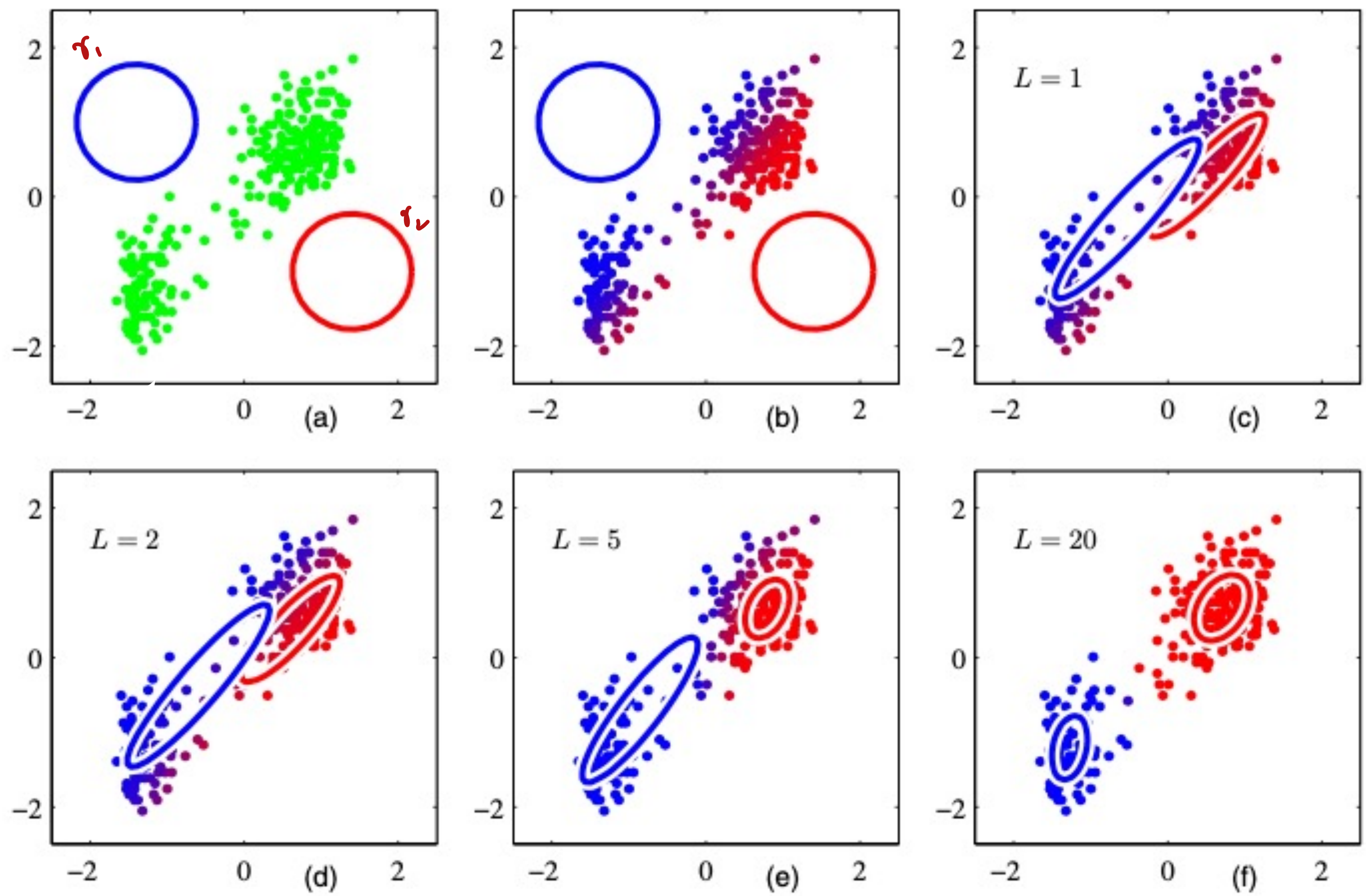
$$\bar{\Theta} = [\gamma_1, \ldots, \gamma_k, \bar{\mu}^{(1)}, \ldots, \bar{\mu}^{(k)}, \sigma_1^2, \ldots, \sigma_k]$$

$$\underbrace{P(z^{(i)} \mid \bar{x}^{(i)})}_{\text{posterior}} \propto \underbrace{P(\bar{x}^{(i)} \mid z^{(i)})}_{\text{likelihood}}\ \underbrace{P(z^{(i)})}_{\text{prior}}$$

Properties of the EM algorithm

- each iteration improves log-likelihood
  - E step never decreases log-likelihood
  - M step never decreases log-likelihood
- EM converges to a (local) optimum

Expectation Maximization



multivariate Gaussian Mixture Model

Model Selection: how to pick k?

Bayesian Information Criterion (BIC)

number of training data

Log-likelihood

$$BIC(D;\bar{\theta}) = \underbrace{\boxed{l(D;\bar{\theta})}}_{\text{Log-likelihood}} - \underbrace{\boxed{\frac{\#param}{2}}}_{\text{model complexity}} log(\underbrace{n}_{\text{number of training data}})$$

model complexity

Here we'd want to maximize the BIC.

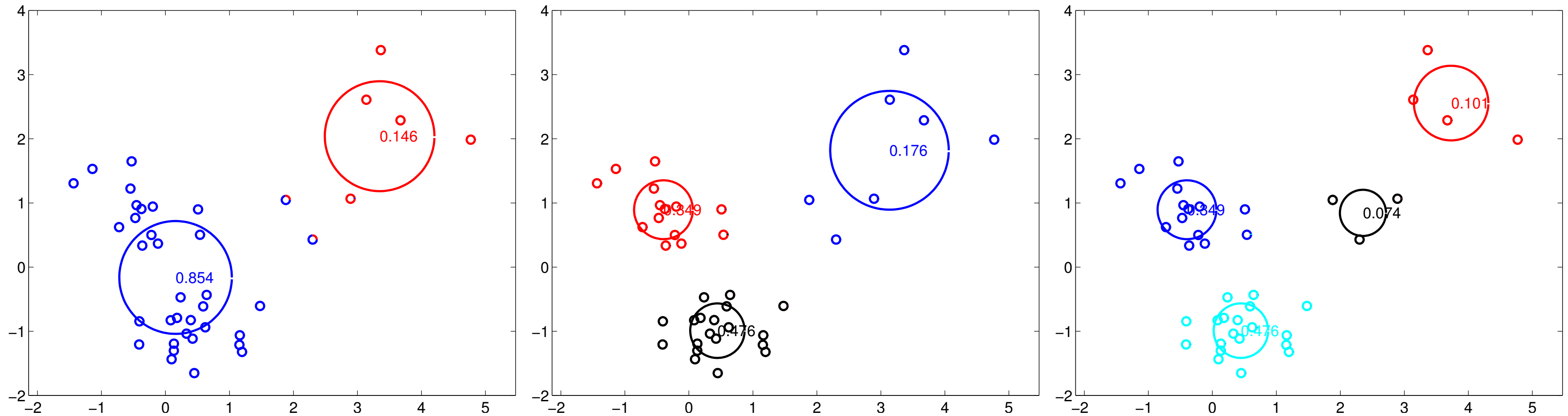
Sometimes defined as the negative of above definition. In such cases, we want to minimize.

Model Selection for Mixtures

Bayesian Information Criterion (BIC)

Example:

$k = 2$ $\quad$ $k = 3$ $\quad$ $k = 4$



$l(D;\hat{\theta}) = -118.25$ $\quad$ $l(D;\hat{\theta}) = -98.64$ $\quad$ $l(D;\hat{\theta}) = -94.11$

$BIC(D;\hat{\theta}) = -131.16$ $\quad$ $BIC(D;\hat{\theta}) = -118.93$ $\quad$ $BIC(D;\hat{\theta}) = -121.78$

From Jaakkola

Graphical Models:
Bayesian Networks

Bayesian Networks by Example

nodes: variable

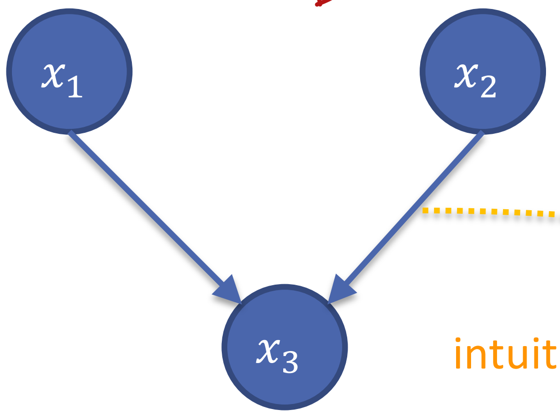
directed edges: dependencies

CPT

1 row

x_1

H	T
0.5	0.5



DAG

x_2

H	T
0.5	0.5

1 row

intuitively, read this edge as "influences"

x_1 is a parent of x_3

x_3 is a child of x_1

$x_3 \colon x_1 == x_2$

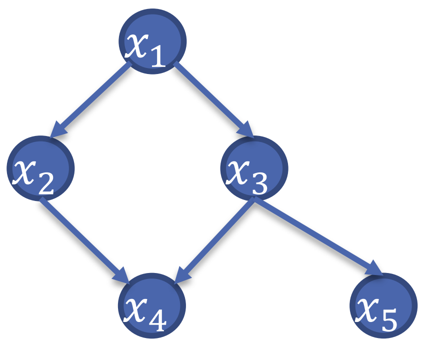
x_1	x_2	$\Pr(x_3 = T \mid \boldsymbol{x_1}, \boldsymbol{x_2})$	$\Pr(x_3 = \boldsymbol{F} \mid \boldsymbol{x_1}, \boldsymbol{x_2})$
H	H	1	0
T	H	0	1
H	T	0	1
T	T	1	0

Factorization *based on given graph*: $\Pr(x_1, x_2, x_3) = \Pr(x_1)\Pr(x_2)\Pr(x_3 \mid x_1, x_2)$

Factorization: Example

For a given graph, the joint distribution can be written as a product of the conditional probability of each variable given its parents

$$P(X_1, \dots, X_d) = \prod_{i=1}^{d} P(X_i | X_{pa_i})$$



From the chain rule:

$$\Pr(x_1, x_2, x_3, x_4, x_5)$$
$$= \Pr(x_1) \Pr(x_2|x_1) \Pr(x_3|x_1, x_2) \Pr(x_4|x_3, x_2, x_1) \Pr(x_5|x_4, x_3, x_2, x_1)$$

vs.

From the factorization based on the graph:

$$\Pr(x_1, x_2, x_3, x_4, x_5)$$
$$= \Pr(x_1) \Pr(x_2|x_1) \Pr(x_3|x_1) \Pr(x_4|x_3, x_2) \Pr(x_5|x_3)$$

Two notions of Independence

Marginal independence

$$\Pr(X_1, X_2) = \Pr(X_1)\Pr(X_2)$$

$$X_1 \perp X_2$$

Alternately, $\Pr(X_1 | X_2) = \Pr(X_1)$

Bayesian Networks encode independencies

Conditional independence

$$\Pr(X_1, X_2 | X_3) = \Pr(X_1 | X_3)\Pr(X_2 | X_3)$$

$$X_1 \perp X_2 | X_3$$

Alternately, $\Pr(X_1 | X_2, X_3) = \Pr(X_1 | X_3)$

d-separation:
Inferring independence

Bayesian Networks provide us a way to determine these via the dependency graph

Independence from the Graph (d-separation)

Steps

1. keep only "ancestral" graph

2a. connect nodes with common child

2b. make undirected

3. read off property

If there is no path between variables of interest, then they are marginally independent

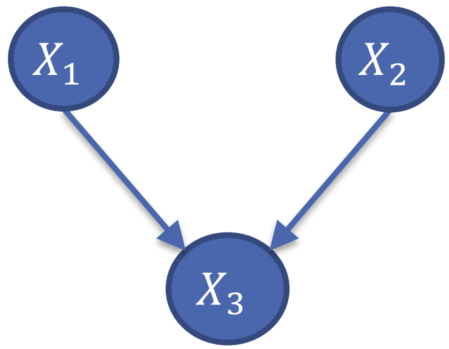
If all paths between variables of interest go through a particular node, then the variables are independent given that node

intuitively can say that that node "blocks" the influence from the first variable to the second

Note: *for* $X_1 \perp X_2 | \{X_3, X_4\}$ *each path has to go through at least one of the nodes in the set* $\{X_3, X_4\}$

d-separation Examples

Does the graph imply $X_1 \perp X_2 \,|X_3$? Does the graph imply $X_1 \perp X_2$?



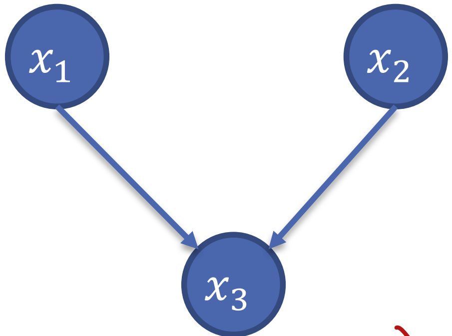
$X_1 \perp X_2 \,|X_3$?

$X_1 \perp X_2$?

$x_1 \perp x_2$? Y

$x_1 \perp x_2 \,| x_3$? N

Marginal Independence: Example



Claim: x_1 and x_2 are (marginally) independent of each other

That is, want to show that $\Pr(x_1, x_2) = \Pr(x_1)\Pr(x_2)$

$$\Pr(x_1, x_2) = \sum_{x_3} \Pr(x_1, x_2, x_3) = \sum_{x_3} \Pr(x_1)\Pr(x_2)\Pr(x_3 \mid x_1, x_2)$$

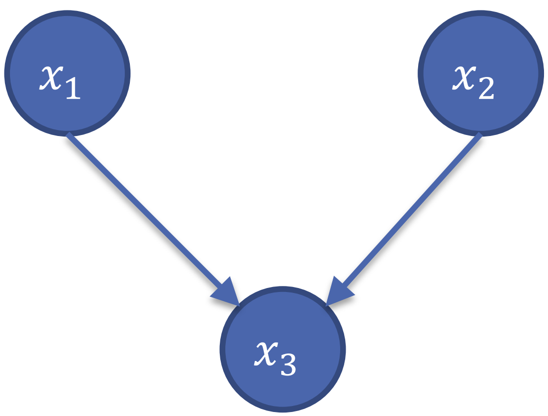
$$= \Pr(x_1)\Pr(x_2) \underbrace{\boxed{\sum_{x_3} \Pr(x_3 \mid x_1, x_2)}}_{=1}$$

$$= \Pr(x_1)\Pr(x_2)$$

Conditional Independence : Example

However, x_1 and x_2 are *conditionally dependent* given x_3

To see this, note that if we knew $x_3 = T$ then we know that either $x_1 = x_2 = H$ or $x_1 = x_2 = T$



x_1

H	T
0.5	0.5

x_2

H	T
0.5	0.5

x_3

x_1	x_2	$\Pr(x_3 = T \mid \boldsymbol{x_1}, \boldsymbol{x_2})$	$\Pr(x_3 = \boldsymbol{F} \mid \boldsymbol{x_1}, \boldsymbol{x_2})$
H	H	1	0
T	H	0	1
H	T	0	1
T	T	1	0

Learning Bayesian Networks

Learning Bayesian Networks

Two Main Problems

1. estimate parameters given graph structure (and data)
2. search over possible graph structures (model sel.)