Probabilistic Time Series Analysis: Lab 0

Haresh Rengaraj Rajamohan & Caroline Haimerl

Logistics

- Time and Location
  - Wednesday 3:45 PM-4:35 PM
- Lab HW submission
  - By the end of the lab session
- TA office hour
  - TBD

Bayes Recap

Let's say we have two random variables X and Y, where X takes l distinct values and Y takes m distinct values. We observe data (x_i, y_i) , where

$i = 1,2,...,n$.

1. How many parameters do we need to represent the joint distribution ?
2. How do we estimate these parameters from the dataset that we have observed ?

Bayes Recap

Let's say we have two random variables X and Y, where X takes l distinct values and Y takes m distinct values. We observe data (x_i, y_i) , where

$i = 1,2,...,N$.

1. How many parameters do we need to represent the joint distribution ? **l*m -1**
2. How do we estimate these parameters from the dataset that we have observed ? **MLE**

Recap MLE

What’s the likelihood function in this case ?

What’s the expression for the estimate of the parameters?

Recap MLE

What's the likelihood function in this case ?

$$l = \prod_{i=1}^{n} \theta_i = \prod_{j=1}^{ml-1} \theta_j^{n_j} (1 - \sum \theta_k)^{(N - \sum n_u)}$$

Where $\theta_i = p(X = x_i, Y = y_i)$ and n_j is the number of observations of a specific combination of x and y.

What's the expression for the estimate of the parameters?

$$\theta_j = \frac{n_j}{N}$$

Bayes Recap

Now we have the joint distribution, how to find the marginal distribution of X, p(X).

$$p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_i)$$

called the sum rule of probability

Bayes Recap

- What’s the Product rule in probability ?

Bayes Recap

- What's the Product rule of probability ?

$$P(X = x_i, Y = y_j) = P(X = x_i \,|\, Y = y_j) p(Y = y_j)$$
$$= P(Y = y_j \,|\, X = x_i) p(X = x_i)$$

Bayes Recap

We get the bayes formulation to reverse the conditional probabilities, which is

$$p(Y = y_j | X = x_i) = \frac{p(X = x_i | Y = y_j) p(Y = y_j)}{p(X = x_i)}$$

Bayes Recap

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Where $P(Y)$ is the prior distribution, $P(X|Y)$ is the likelihood of the dataset and $P(Y|X)$ is the posterior distribution and $P(X)$ is the normalization.

Problem 1

Example 2.3. Jo has a test for a nasty disease. We denote Jo's state of health by the variable a and the test result by b.

$$\begin{array}{ll} a = 1 & \text{Jo has the disease} \\ a = 0 & \text{Jo does not have the disease.} \end{array} \tag{2.12}$$

The result of the test is either 'positive' ($b = 1$) or 'negative' ($b = 0$); the test is 95% reliable: in 95% of cases of people who really have the disease, a positive result is returned, and in 95% of cases of people who do not have the disease, a negative result is obtained. The final piece of background information is that 1% of people of Jo's age and background have the disease.

OK – Jo has the test, and the result is positive. What is the probability that Jo has the disease?

from MacKay (1995)

Solution

Solution. We write down all the provided probabilities. The test reliability specifies the conditional probability of b given a:

$$\begin{aligned} P(b=1 \mid a=1) &= 0.95 \qquad & P(b=1 \mid a=0) &= 0.05 \\ P(b=0 \mid a=1) &= 0.05 \qquad & P(b=0 \mid a=0) &= 0.95; \end{aligned} \tag{2.13}$$

and the disease prevalence tells us about the marginal probability of a:

$$P(a=1) = 0.01 \qquad P(a=0) = 0.99. \tag{2.14}$$

From the marginal $P(a)$ and the conditional probability $P(b \mid a)$ we can deduce the joint probability $P(a,b) = P(a)P(b \mid a)$ and any other probabilities we are interested in. For example, by the sum rule, the marginal probability of $b=1$ – the probability of getting a positive result – is

$$P(b=1) = P(b=1 \mid a=1)P(a=1) + P(b=1 \mid a=0)P(a=0). \tag{2.15}$$

Jo has received a positive result $b=1$ and is interested in how plausible it is that she has the disease (i.e., that $a=1$). The man in the street might be duped by the statement 'the test is 95% reliable, so Jo's positive result implies that there is a 95% chance that Jo has the disease', but this is incorrect. The correct solution to an inference problem is found using Bayes' theorem.

$$\begin{aligned} P(a=1 \mid b=1) &= \frac{P(b=1 \mid a=1)P(a=1)}{P(b=1 \mid a=1)P(a=1) + P(b=1 \mid a=0)P(a=0)} && (2.16) \\ &= \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} && (2.17) \\ &= 0.16. && (2.18) \end{aligned}$$

So in spite of the positive result, the probability that Jo has the disease is only 16%. □

from MacKay (1995)

Recap

Our goal is to model time series data. So we have,

$$p(X_t = x_t, X_{t-1} = x_{t-1}, \dots, X_0 = x_0)$$

Recap

- Can we estimate this in the way we did earlier ? Explain ?
- What’s the solution to this problem ?

Recap

- Can we estimate this in the way we did earlier ? Explain ?
    - No, because it’s intractable. Assuming the R.Vs take K distinct values, we have K^t -1 parameters o work with and we ‘ll need even more data
- What’s the solution to this problem ?
    - Make assumptions about the model. Ex: Markov

Recap

Using the chain rule, we can write the joint distribution as

$$p(X_t = x_t, X_{t-1} = x_{t-1}, \dots, X_0 = x_0) =$$

$$\prod_k p(X_k = x_k | X_{k-1} = x_{k-1}, \dots, X_0 = x_0)$$

Recap

The markov assumption :

$$P(X_n = x_n \mid X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = P(X_n = x_n \mid X_{n-1} = x_{n-1})$$

So the joint distribution is

$$P(X_0)P(X_1|X_0)P(X_2|X_1)...P(X_t|X_{t-1})$$

Recap

$$P(X_0)P(X_1|X_0)P(X_2|X_1)...P(X_t|X_{t-1})$$

How many parameters do we have to estimate now ?

Recap

$$P(X_0)P(X_1|X_0)P(X_2|X_1)...P(X_t|X_{t-1})$$

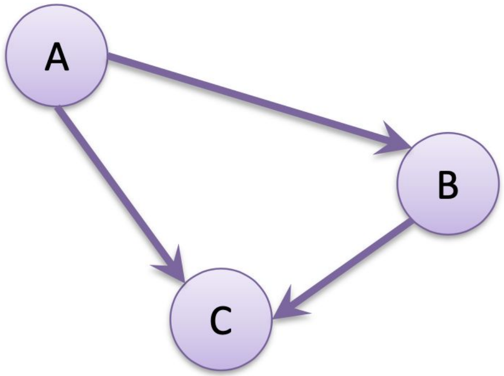
How many parameters do we have to estimate now ?

K+K(K-1)+K(K-1)+..= (t-1)(K^2-K)+K

Which doesnt blow up badly now.

Graphical Models

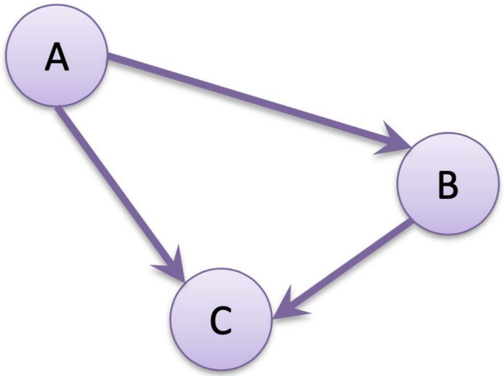
Consider joint distribution of 3 R.Vs A,B,C represented by the Bayesian network



$$p(A,B,C) = p(C|A,B)p(B|A)p(A)$$

Graphical Models

So we only need to condition on the parent nodes

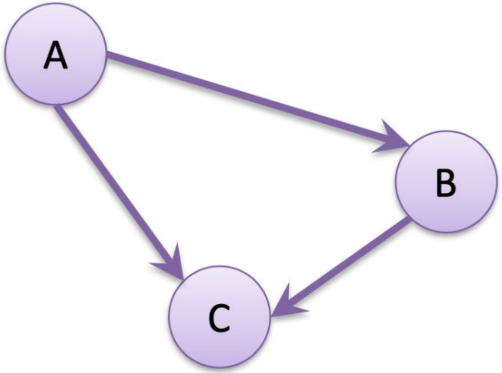


$$p(A, B, C) = p(C|A, B)p(B|A)p(A)$$

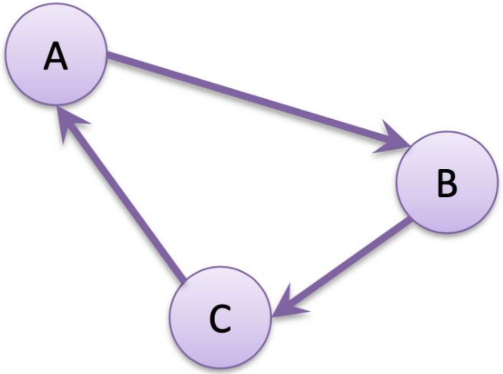
Graphical Models

Bayesian networks

directed acyclic graphs (DAG)

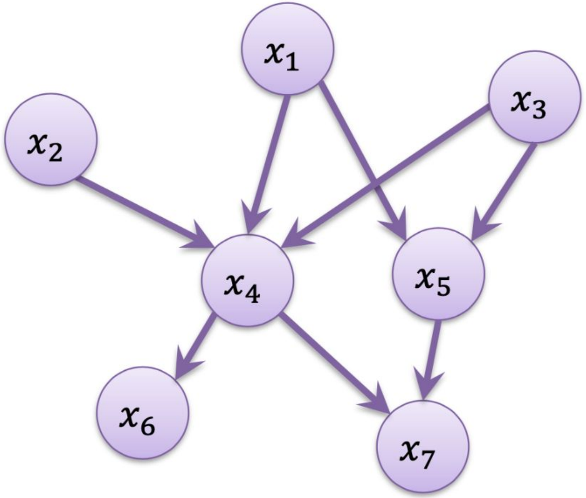


directed cyclic graphs



Graphical Models

Bayesian networks



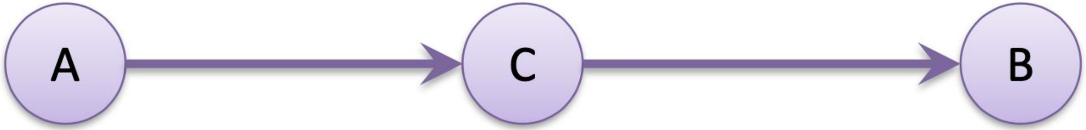
$$p(x_1, x_2, x_3, x_4, x_5, x_6, x_7) =$$
$$p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$

Conditional Independence

- A and B are independent if $P(A,B) = P(A)P(B)$
- A and B are conditionally independent given C if $P(A,B|C)=P(A|C)P(B|C)$

Graphical models can abstract conditional independencies

Conditional Independence



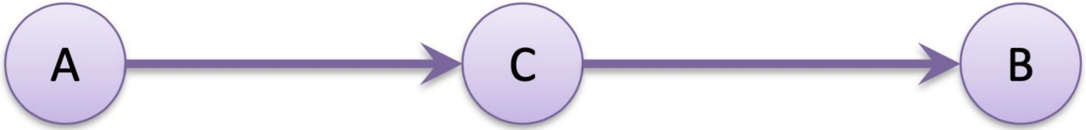
We know that,

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

What about $P(A, B|C)$?

Are A and B independent ?

Conditional Independence

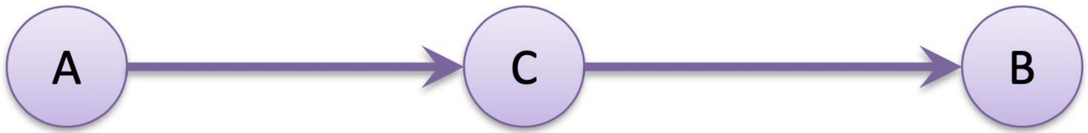


We know that,

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

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

Conditional Independence

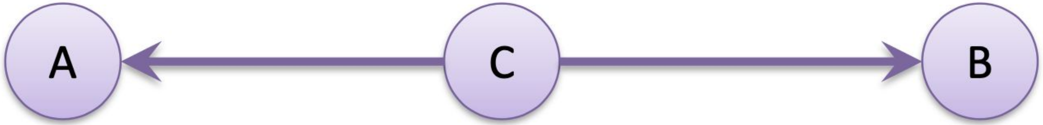


So is $P(A, B|C) = P(A|C)P(B|C)$?

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

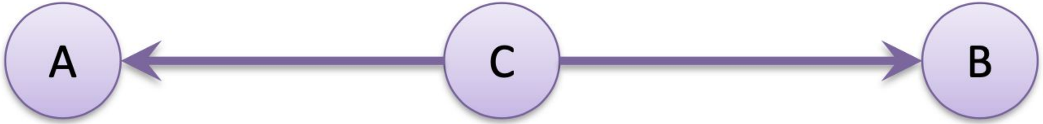
So, A and B are independent given C

Conditional Independence



$$p(A, B, C) = p(A|B, C)p(B|C)p(C)$$

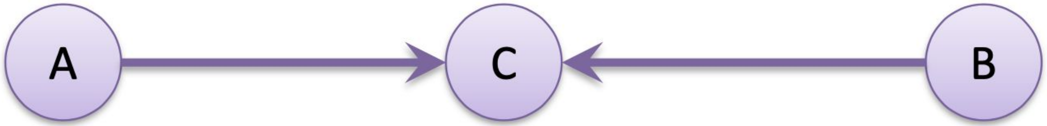
Conditional Independence



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

A

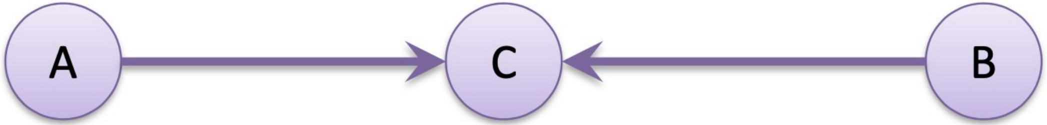
Conditional Independence



Are A and B independent ?

$$p(A, B, C) = p(A)p(B)p(C|A, B)$$

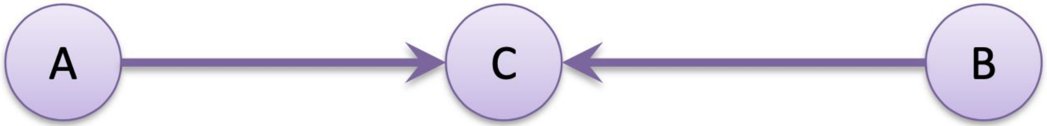
Conditional Independence



Are A and B independent ?

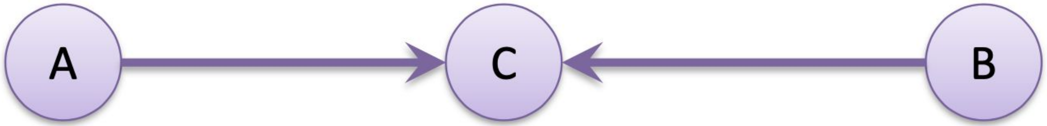
$$P(A,B) = \sum_C P(A,B,C) = P(A)P(B)\sum_C P(C|A,B) = P(A)P(B)$$

Conditional Independence



Are A and B independent given C ?

Conditional Independence



Are A and B independent given C ?

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

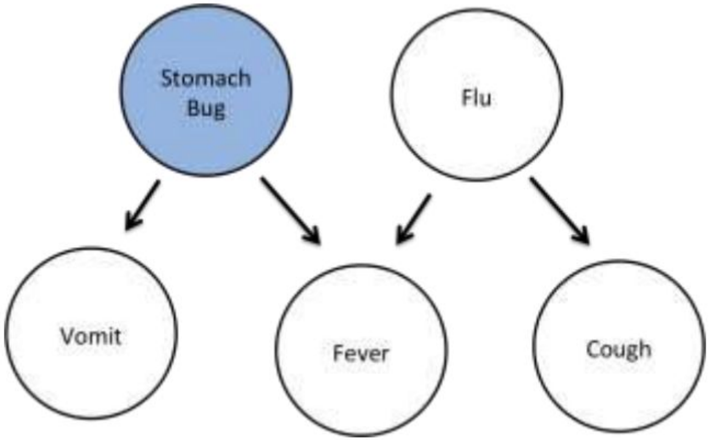
$$\neq p(A|C)p(B|C)$$

Problem 2 - From CS 221 Stanford

Stomach bug causes vomiting and fever. Flu causes fever and cough. You know that you do not have a stomach bug. If you were to vomit, would that information change the probability that you had a fever ?

Solution

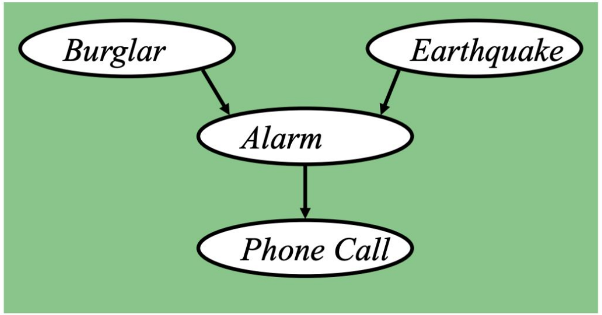
The two variables are conditionally independent given stomach bug



Problem 3 - From 10-601 CMU

The "Burglar Alarm" example

- Your house has a twitchy burglar alarm that is also sometimes triggered by earthquakes.
- Earth arguably doesn't care whether your house is currently being burgled
- While you are on vacation, one of your neighbors calls and tells you your home's burglar alarm is ringing. Uh oh!



Quiz: True or False?

$$Burglar \perp\!\!\!\perp Earthquake \mid PhoneCall$$

Slide from William Cohen

Solution

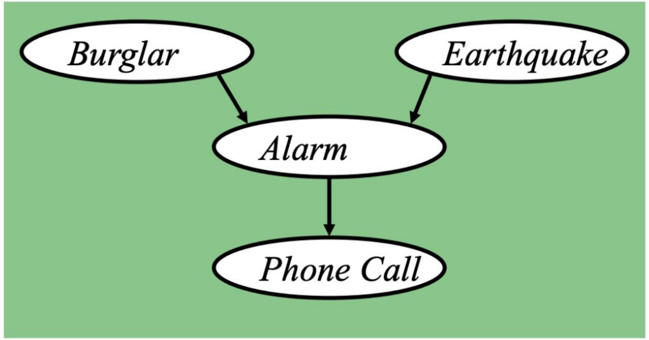
The “Burglar Alarm” example

- But now suppose you learn that there was a medium-sized earthquake in your neighborhood. Oh, whew! Probably not a burglar after all.
- Earthquake “explains away” the hypothetical burglar.
- But then it must **not** be the case that

$$Burglar \perp\!\!\!\perp Earthquake \mid PhoneCall$$

even though

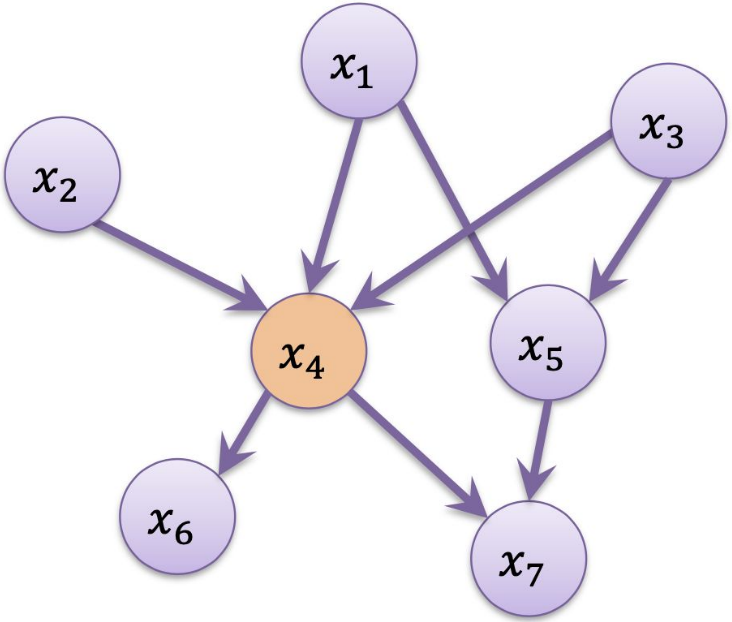
$$Burglar \perp\!\!\!\perp Earthquake$$



Slide from William Cohen

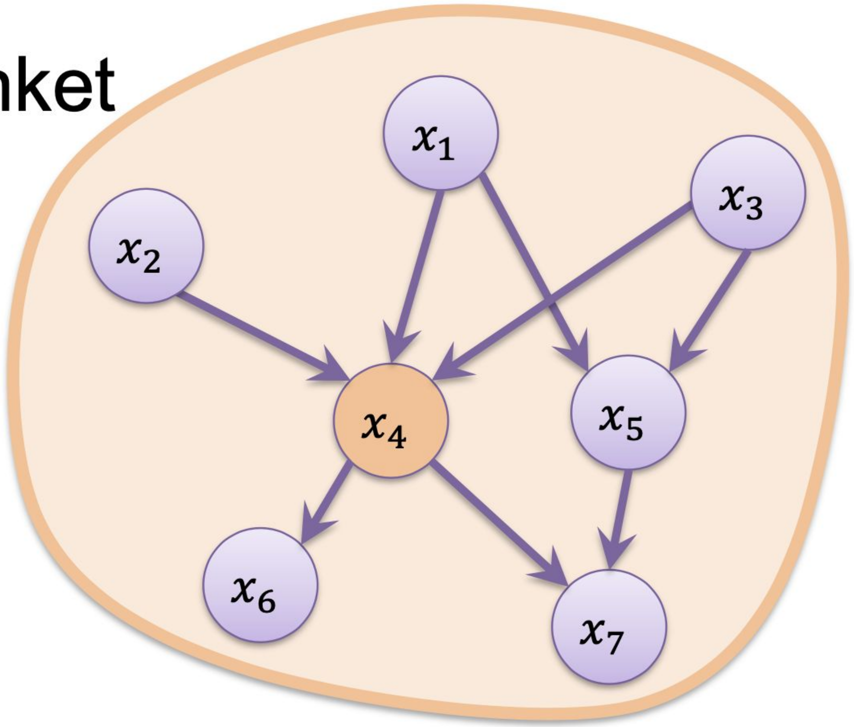
Conditional Independence

d-separated



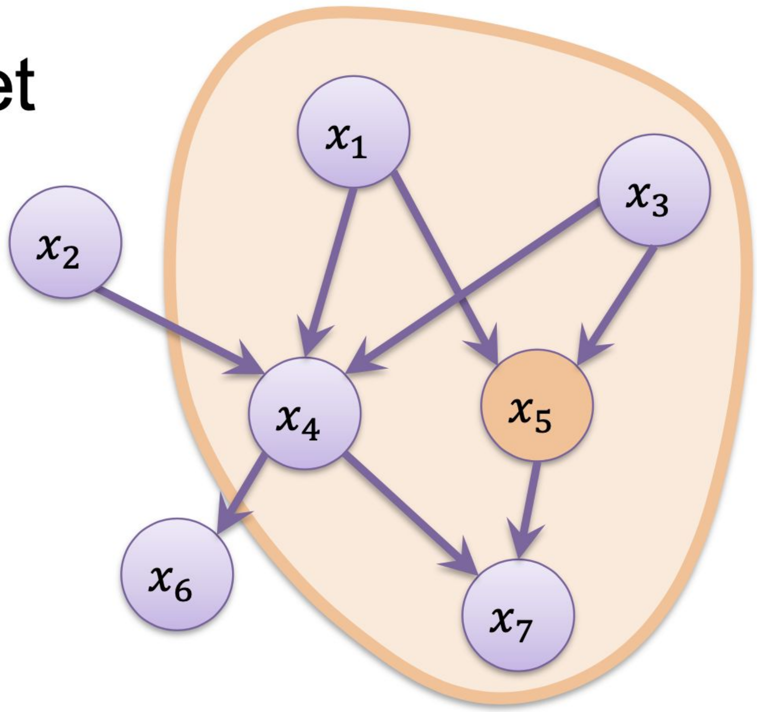
Conditional Independence

Markov blanket

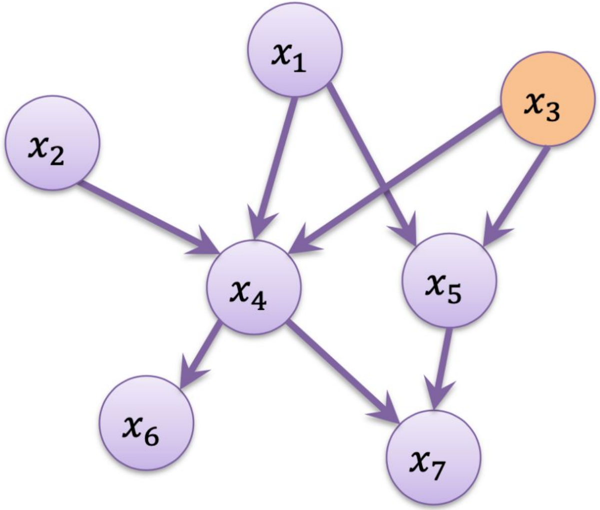


Conditional Independence

Markov blanket

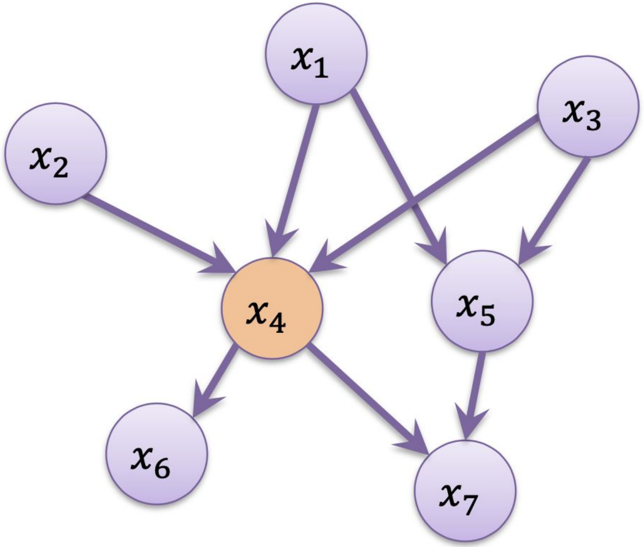


Conditional Independence



$$p(x_1, x_2, x_4, x_5, x_6, x_7 | x_3) =$$
$$p(x_1)p(x_2)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$

Conditional Independence



$$p(x_1, x_2, x_3, x_5, x_6, x_7 | x_4) =$$
$$p(x_1, x_2, x_3 | x_4)\, p(x_5 | x_1, x_3) p(x_6 | x_4) p(x_7 | x_4, x_5)$$

Resources

- Bishop, C., Pattern Recognition and Machine Learning (2006)
- MacKay, D., Information Theory, Inference, and Learning Algorithms (2005)
- Lecture 23 CMU Intro to ML slides
- Practice problems 1 Stanford CS 221
- notes by Michael Jordan
- notes by Zoubin Ghahramani