DD2380

Artificial Intelligence

Module: Taming uncertainty

Patric Jensfelt

DD2380

Artificial Intelligence

Probabilistic reasoning

Patric Jensfelt

Credits

- Based partly on material from
  - Kevin Murphy, MIT, UBC, Google
  - Danica Kragic, KTH
  - W. Burgard, C. Stachniss, M. Benewitz and K. Arras, when at Albert-Ludwigs-Universität Freiburg

Reading instructions

- Chapters 13-15
- Additional material for the HMM part

Motivation: Why use probabilities

- An artificial system (or human) does not have perfect knowledge of its environment and of the results of its actions, and it needs to deal with uncertainty at many levels

- Uncertainty plays an important role in: sensor interpretation, sensor fusion, map making, path planning, self-localization, control, etc, etc

Motivational examples

- Diagnos diseases
  - Doctor knows (Dr Watson)
    - How common a certain disease is
    - Connection with factors such as age, sex, habits, …
    - Connection with measures, e.g. temperature
  - Observe
  - Diagnose

Motivational examples

- Autonomous car: Cross intersection safely?
  - From manufacturer or learned by car
    - Sensor models
    - Statistic from different roads
    - Weather models
    - …
  - Observations from own car and others?
  - Q:
    - Can I cross if I want to be 99.99999% safe?
    - Can I cross if I want to be 99% safe?

Motivational examples

- A1: Hunt ducks!
  - Have some knowledge about how different birds move
  - Have sparse observations of birds
  - Q: How will they move next so I can shoot?
  - Q: How to tell which birds belong to the same species?

  - You get to : Design a model, learn the parameters from data, make decisions, take actions

Recap of probability theory

- Probability of event X: p(X)
- Joint probability of X and Y: p(X,Y),
  i.e. X AND Y
- Conditional probability of X given Y: p(X|Y)

Notation

- For boolean variables we will use
  - p(X) to mean the probability that X is true and
  - p(¬X) the probability that X is false.

Rules of probability

- $p(X) \in [0,1]$ (i.e. $0 \leq p(X) \leq 1$)
- $p(X) = 1 - p(\neg X)$
- Sum rule (marginalization): $p(X) = \sum_{Y} p(X,Y)$
- Product rule: $p(X,Y) = p(Y|X)p(X)$
- Prob sum to one: $1 = \sum_{\text{all } X} p(X)$

Conditioning

- Combining
  - $p(X)=\sum_{Y} p(X,Y)$ (sum rule)
  - $p(X,Y)=p(X|Y)p(Y)$ (product rule)

    gives
  - **$p(X)=\sum_{Y} p(X|Y)p(Y)$**

Bayes rule

- p(Y|X) = p(X|Y) p(Y) / p(X)
- posterior = likelihood * prior / evidence
- What is the probability of Y given some evidence X can be expressed in factors that are sometimes easier to determine.
- Exercise1: Prove it using the previous rules!

Bayes rule cont'd

- Rewrite $p(Y|X) = p(X|Y)\, p(Y) / p(X)$
- using the sum rule $p(X)=\sum_Y p(X|Y)p(Y)$

$$p(Y|X) = p(X|Y)\, p(Y) / \sum_Y p(X|Y)p(Y)$$

- $p(X)=\sum_Y p(X|Y)p(Y)$ is a normalization factor $1/\eta$

$\rightarrow$

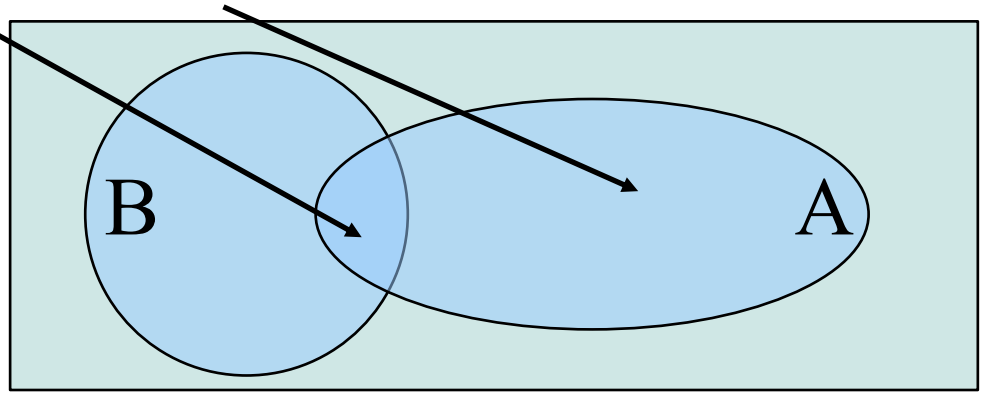
$$p(Y|X) = \eta\, p(X|Y)\, p(Y)$$

Conditional probability

- Consider two dependent events A and B

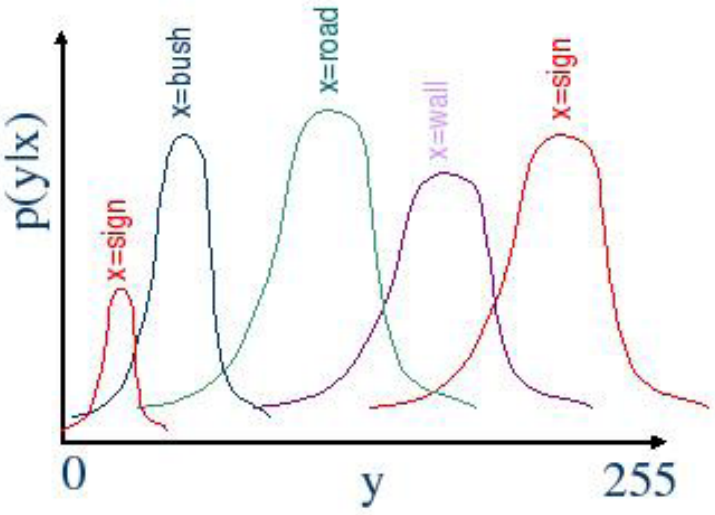
B \ A	True	False
True	0.1	0.3
False	0.4	0.2

- Probability of B given that A is true
  p(B|A) = p(A,B) / P(A)



Example conditional dependence

- Knowing what we look at gives a much better idea of what to expect to measure.
- Ex: If we know that we look for a sign we expect either bright (near 255) for the white background or dark (near 0) for pixel values for the text



So what?

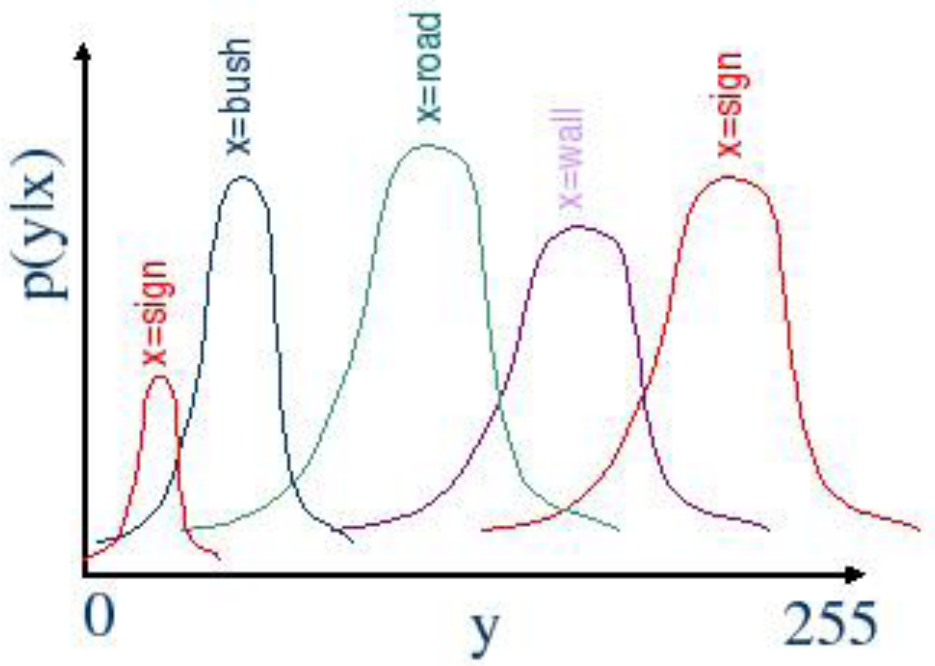
So what?

- If variables are dependent we can get information about one variable by measuring another!!

- Foundation for most of the probabilistic reasoning and statistical machine learning

Simplistic probabilistic reasoning

- If we measure 115, what is the most likely category?



Road classification

Bayes rule example

- Vision system for detecting zebras Z
- Prior: p(Z)=0.02 (zebra in 2% of images)
- Detector for "stripey areas" gives observations, O
- Detector gives yes/no
- Detector performance
  - p(O|Z)=0.8 (true pos)
  - p(O|¬Z)=0.1 (false pos, e.g. gate)

- Task: Calculate p(Z|O)
- What does it represent? How big roughly?

Bayes rules example solution

- p(Z|O) represents probability that there is a zebra if our detector says there is one.
- How big? Use Bayes rule!

$$p(Z|O) = \frac{p(O|Z)p(Z)}{p(O)}$$

- And then?

Bayes rules example solution

- p(Z|O) represents probability that there is a zebra if our detector says there is one.
- How big? Use Bayes rule!

$$p(Z|O) = \frac{p(O|Z)p(Z)}{p(O)}$$

Conditioning →

$$= \frac{p(O|Z)p(Z)}{p(O|Z)p(Z) + p(O|\neg Z)p(\neg Z)}$$

$$= \frac{0.8 \cdot 0.2}{0.8 \cdot 0.02 + 0.1 \cdot 0.98}$$

$$= \frac{0.016}{0.016 + 0.098}$$

$$\approx 0.1404$$

Bayes rule example discussion

- Intuition tells most people that the detector is much better than this, i.e. we would expect to see a much higher p(Z|O) since the detector is correct in 80% of the cases

- However, only 1 out of 50 images has a zebra
  → 49 out of 50 do not contain a zebra
  + detector not perfect

- Failing to account for negative evidence properly is a typical failing of human intuitive reasoning

Conditional independence

- If X is conditionally independent of Y given Z
- p(X|Y,Z) = P(X|Z)

- Which also means

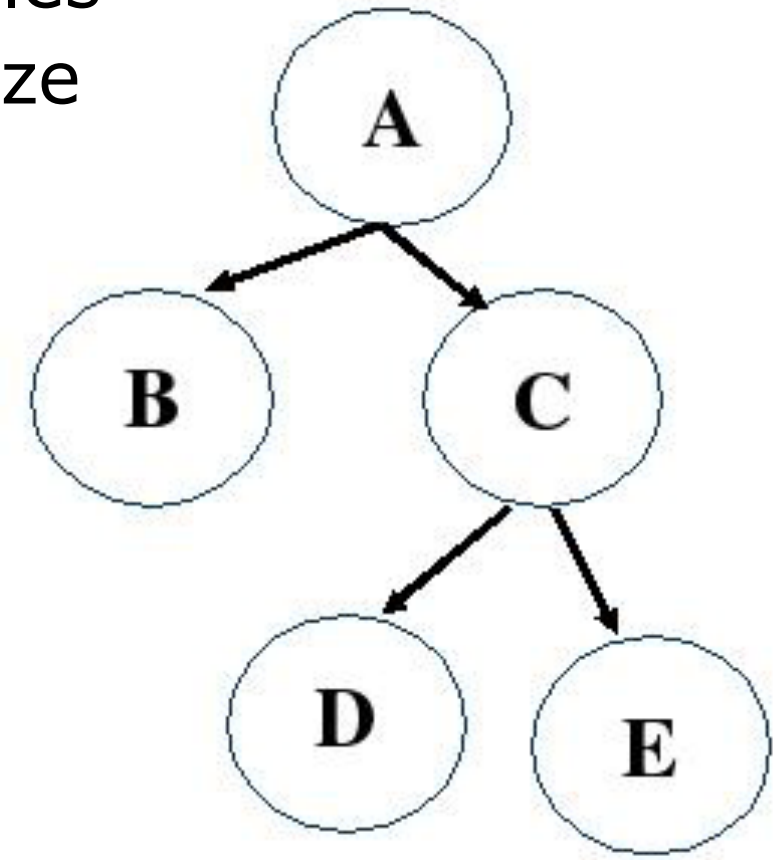
- p(X,Y|Z) = {product rules} =
  - p(X|Y,Z)p(Y|Z) = {conditional independence}
  - p(X|Z)p(Y|Z)

- NOTE: Not the same as p(X,Y) = P(X)p(Y)

Shifting gear, hang on!

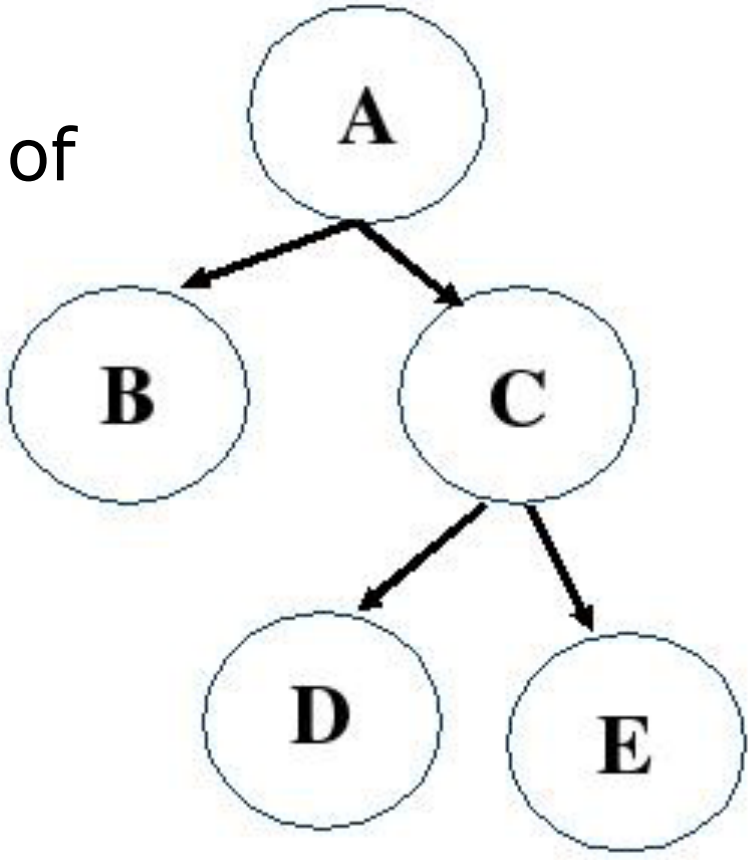
Probabilistic Graphical Models

- Compact repr. of the joint distribution over a set of variables
- Graphical repr. that helps analyze and structure prob. information
- Each variable is encoded as a node
- Variables discrete or continues
- Conditional independence assumptions coded as arcs
- Here: a Bayesian network (directed acyclic graph (DAG))



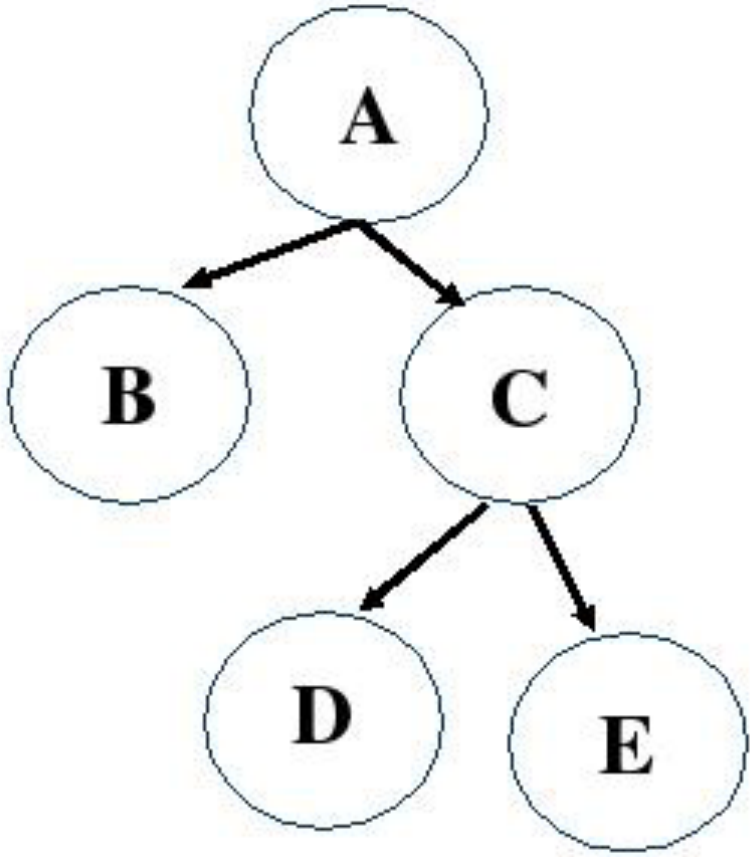
Bayesian network

- A root node
- B, D, E leaf nodes
- A "causes" B and C
  → Value of A influences value of B and C
- A parent to B and C
- B and C children of A
- A ancestor of D and E



Bayesian network

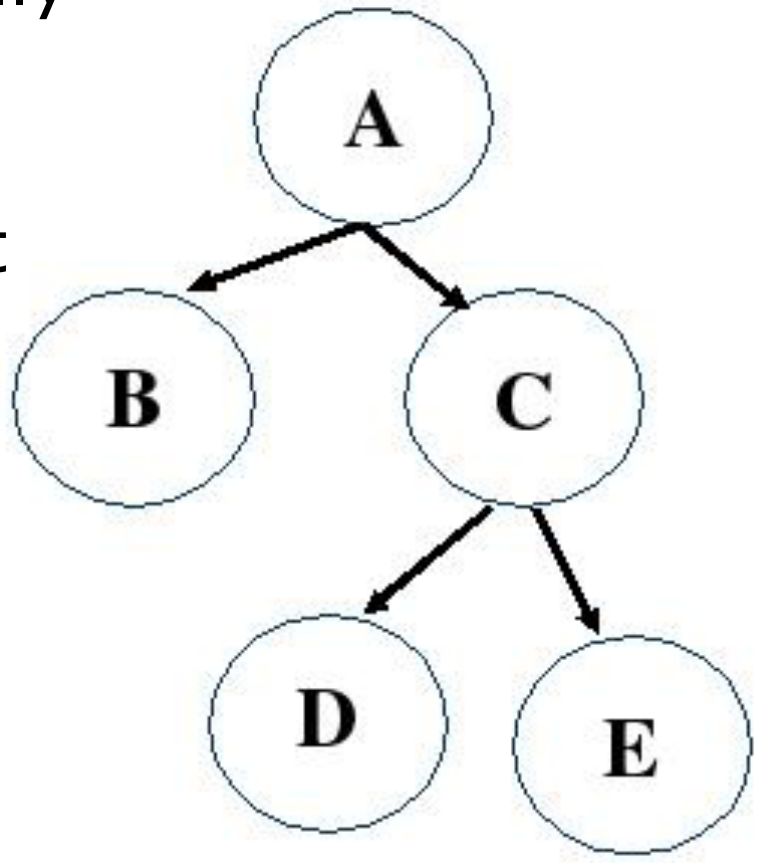
- Arrow → "direct influence over"
- A has direct influence over B



Bayesian network

- B and C are dependent
- HOWEVER, they are conditionally independent given A

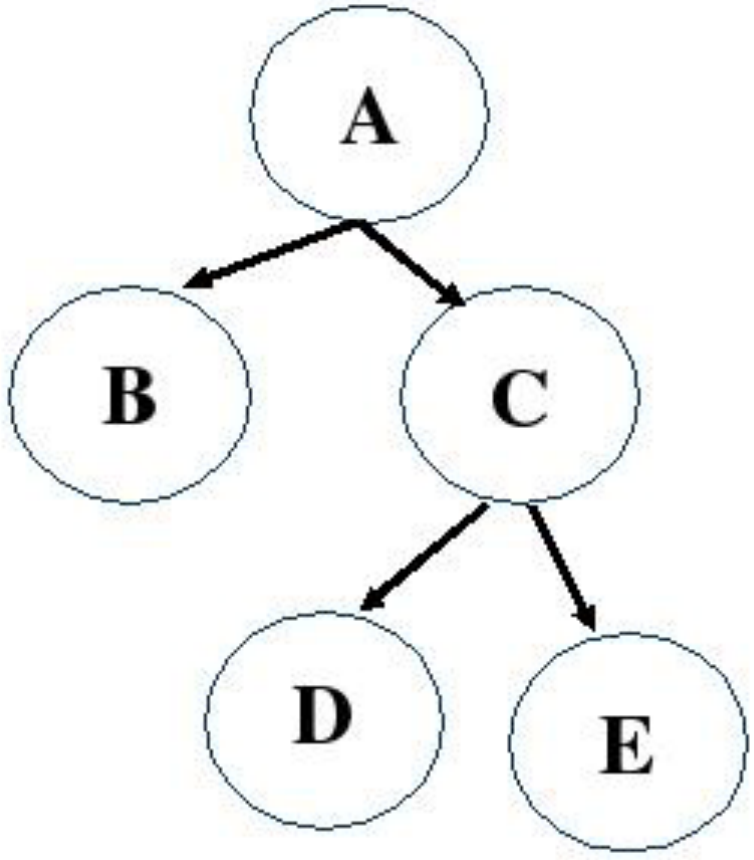
- Q: Write down the formula that relates A, B and C?



B and C independent given A

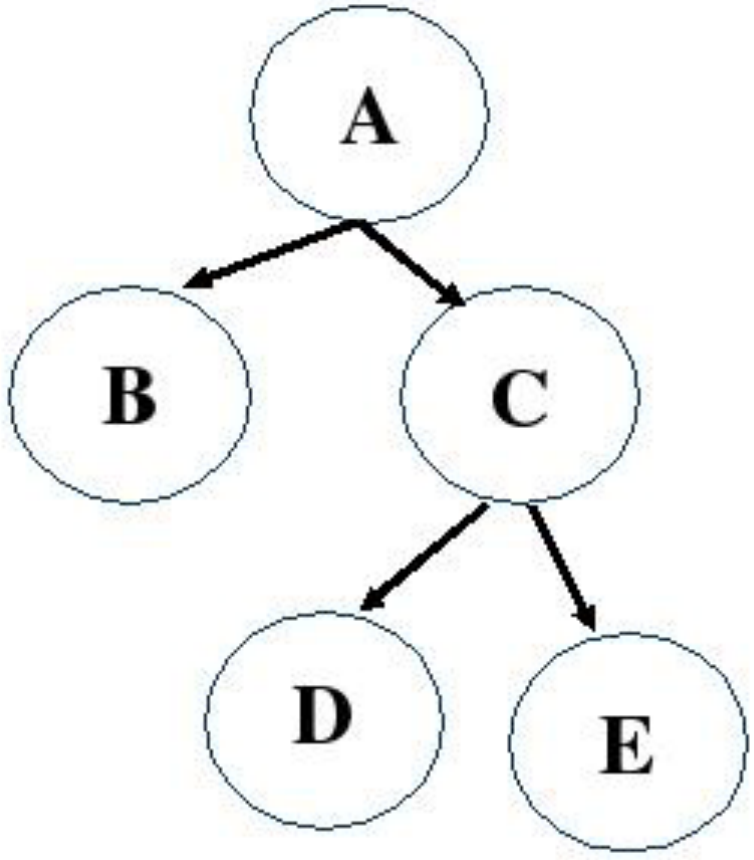
- P(B,C|A)=P(B|A)P(C|A)

- If we do not know A, knowing something about B will tell us something about C (tells us about A which tells us about C)
- but if we know A then knowing B does not tell us anything more about C than we already knew because of A.



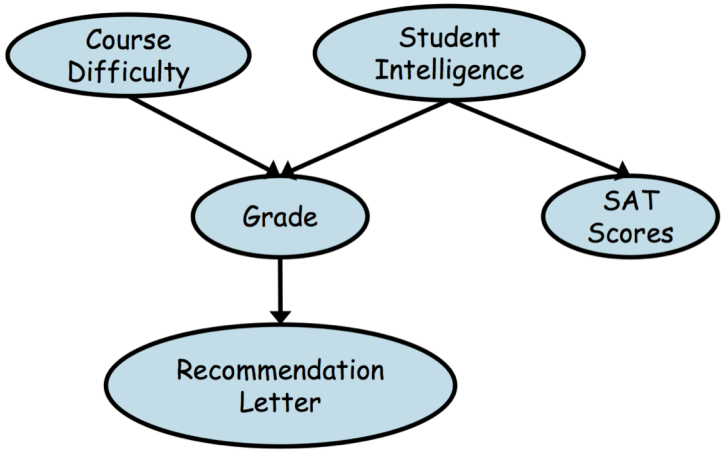
Bayesian network

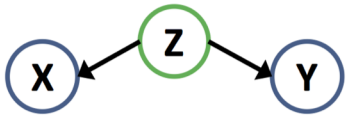
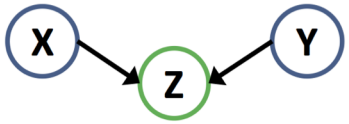
- C depends on A
- E depends on A and C. **HOWEVER**, E is conditionally independent of A given C.
- That is, C captures all the information in A relevant to determine E.

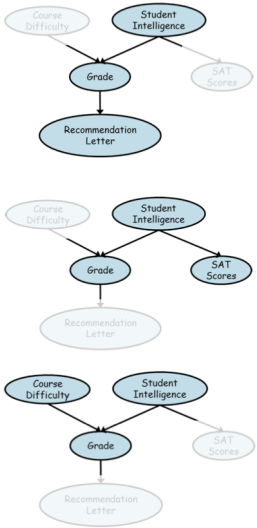


Flow of probabilistic influence

- When can X influence Y?
  - Case1: No evidence about Z
  - Case2: Evidence about Z



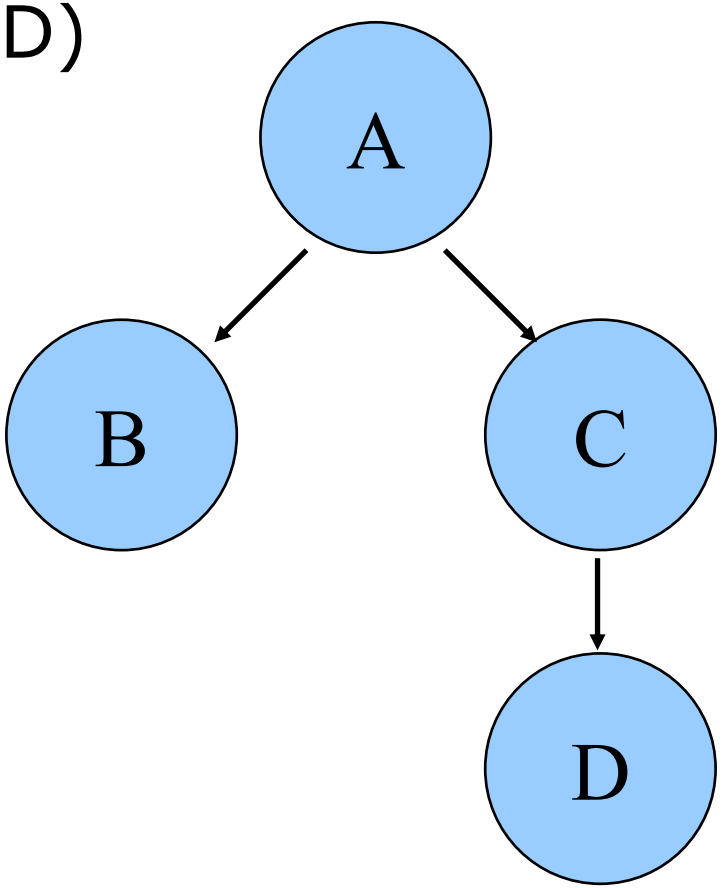
Graphical Structure	NO evidence about Z	YES evidence about Z
	✓	✗
	✓	✗
	✗	✓



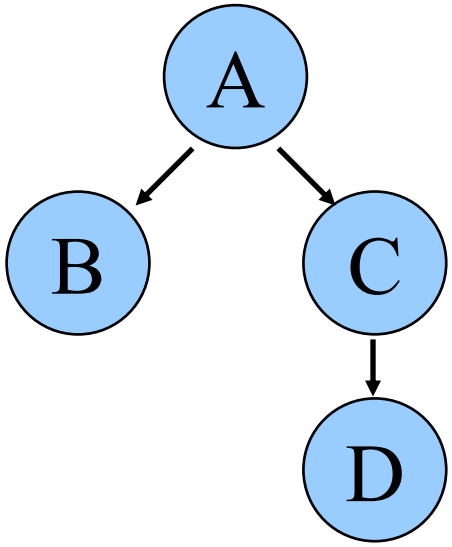
- Adapted from Daphne Koller

Joint distribution

- Exercise2: Factorize p(A,B,C,D)
- Remember product rule
  $p(X,Y) = P(Y|X)P(X)$
- Tip: Work from the top and factor out A, then B, C and D



Derivation



$$
\begin{aligned}
p(A,B,C,D) &= \{\text{product rule with X=A, Y=B,C,D}\} \\
&= p(B,C,D|A)p(A) \\
&= \{\text{product rule with X=B, Y=C,D}\} \\
&= p(C,D|A,B)p(B|A,C,D)p(A) \\
&= \{\text{B conditionally independent of C,D given A}\} \\
&= p(C,D|A,B)p(B|A)p(A) \\
&= \{\text{product rule with X=C, Y=D}\} \\
&= p(D|A,B,C)p(C|A,B)p(B|A)p(A) \\
&= \{\text{C conditionally independent of B given A}\} \\
&= p(D|A,B,C)p(C|A)p(B|A)p(A) \\
&= \{\text{D conditionally independent of A and B given C}\} \\
&= p(D|C)p(C|A)p(B|A)p(A)
\end{aligned}
$$

- In general

$$p(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{i=n} p(X_i|Parents(X_i))$$

Note

- We could have used the product rule in any order
- Looking at the graph we can make use of the conditional independencies
- Other factorization possible, but not as compact

Zebra example

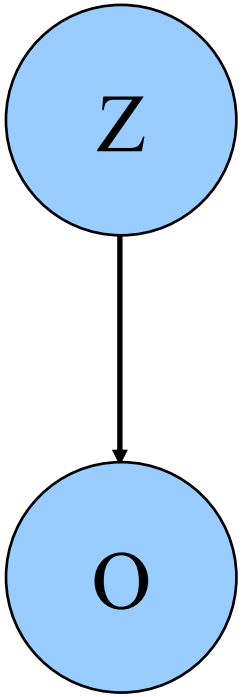
- Vision system for detecting zebras Z
- Prior: p(Z)=0.02 (zebra in 2% of images)
- Detector for "stripey areas" giving observations, O
- Detector gives true/false
- Detector performance
  - p(O|Z)=0.8 (true pos)
  - p(O|¬Z)=0.1 (false pos, e.g. gate)

- Q: Draw as Bayesian network

Zebra example

- Vision system for detecting zebras Z
- Prior: p(Z)=0.02 (zebra in 2% of images)
- Detector for "stripey areas" giving observations, O
- Detector gives true/false
- Detector performance
  - p(O|Z)=0.8 (true pos)
  - p(O|¬Z)=0.1 (false pos, e.g. gate)

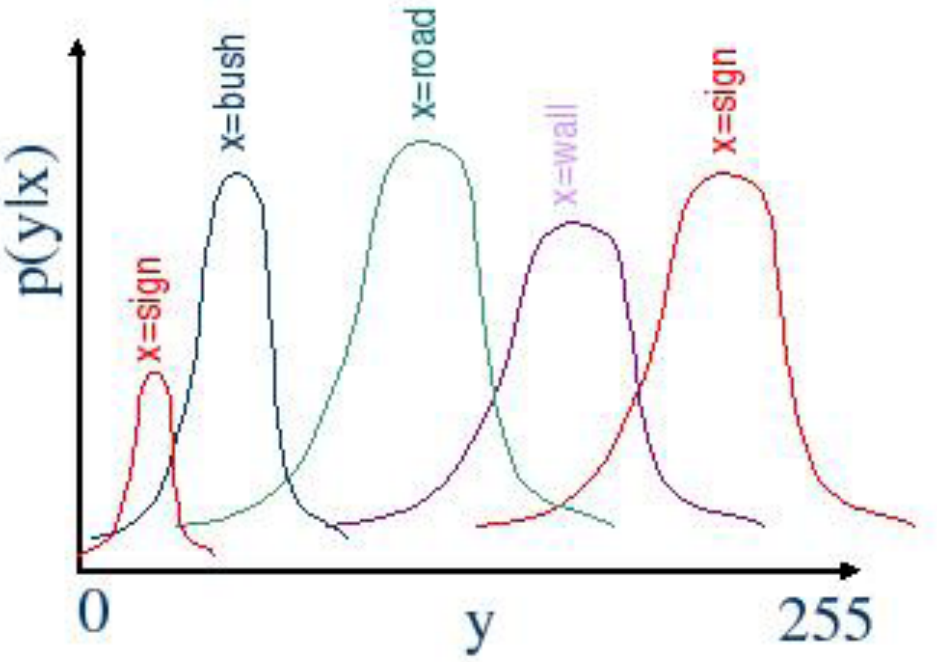
- Q: Draw as Bayesian network

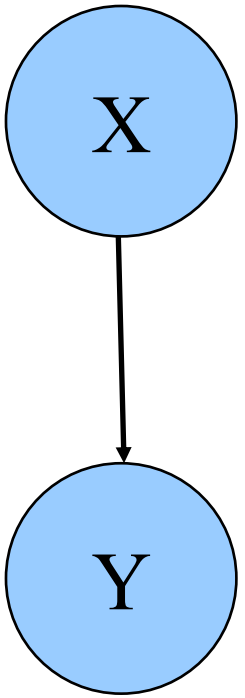


Order? : The existence of the zebra influences the observation and not vice versa

Inference

- Measuring the intensity of a pixel (Y) tells you something about what class (X) that part of the image belongs to, sign, bush, road, etc

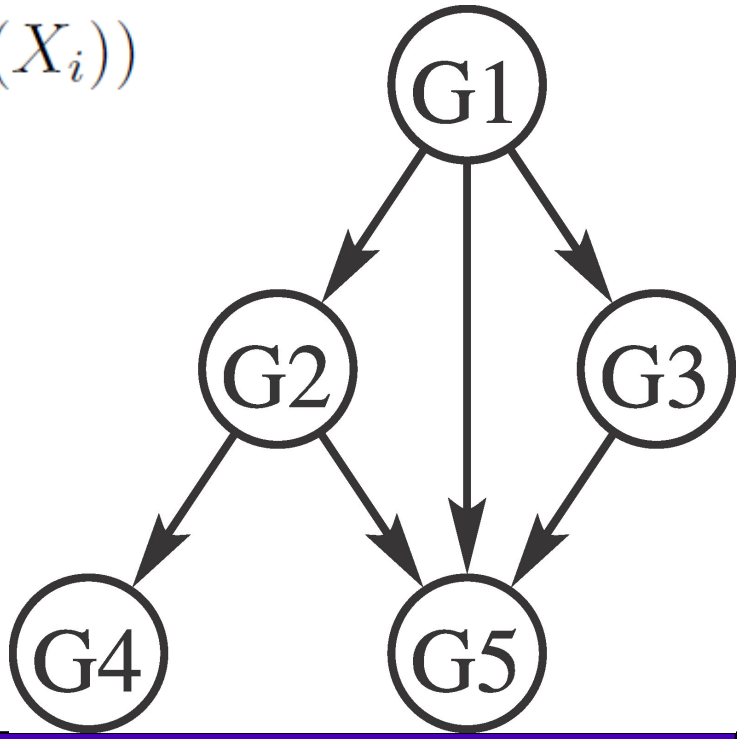




Exercise3: Factorize the graph

- p(G1,G2,G3,G4,G5) = ?
- remember

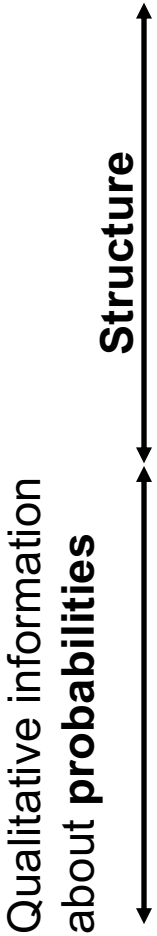
$$p(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{i=n} p(X_i | Parents(X_i))$$



Alarm example

- You have an alarm. It reacts reliably to burglaries but is sometimes triggered also by small earthquakes.
- Two neighbors John and Mary promises to call you at work when they hear the alarm
- John calls almost every time there is an alarm but sometimes confuses it with a phone ringing and calls then too
- Mary plays loud music and sometimes misses it completely but rarely mix other things with it
- Draw Bayesian network!

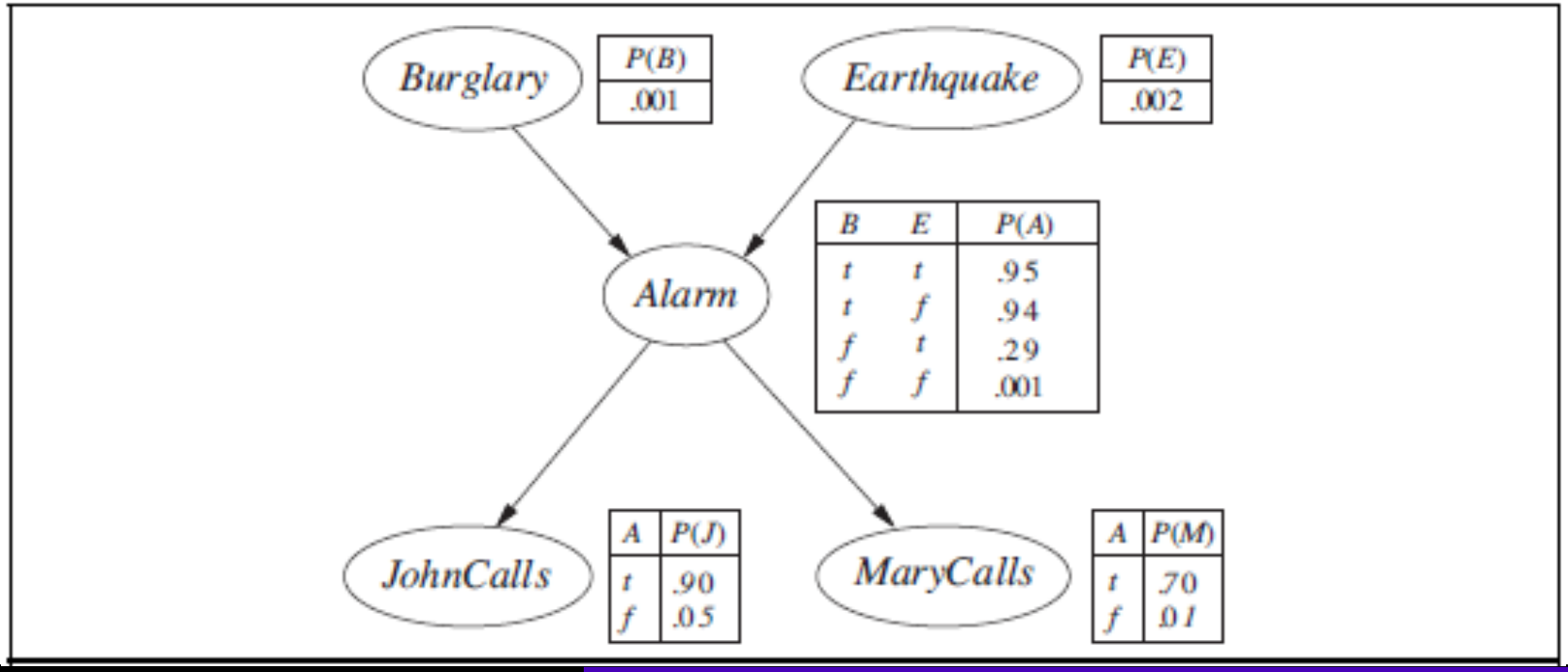
Alarm example



- You have an alarm. It reacts reliably to burglaries but is sometimes triggered also by small earthquakes.
- Two neighbors John and Mary promises to call you at work when they hear the alarm (not earthquake or burglary!)
- John calls almost every time there is an alarm but sometimes confuses it with a phone ringing and calls then too
- Mary plays loud music and sometimes misses it completely but rarely mix other things with it
- Draw Bayesian network! What variables?

Alarm example cont'd

- John and Mary calling does not depend on what triggered the alarm only the alarm itself (simplification)
- CPT – conditional probability table. (Values not specified in the text before!! Made up here)



Exercise4: Alarm example calculation

- Calculate p(J,M,A,$\neg$B,$\neg$E) Means what??

- J : JohnCalls
- M : MaryCalls
- A : Alarm
- B : Burglary
- E : Earthquake

- Remember that p($\neg$X) = 1-p(X) and

$$p(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{i=n} p(X_i | Parents(X_i))$$

Alarm example calculation

- $p(J,M,A,\neg B,\neg E) = 0.00062$

- So what does this tell us?
- Very unlikely that both John and Mary calls and the alarm has gone off when there is no burglary or earthquake

Two views on Bayesian networks

- Representation of the joint probability distribution
  - Helps to understand how to construct it
- Encoding of a collection of independence statements
  - Helpful in designing inference procedure

Tip

- When constructing the network try to use ordering based on
- cause → symptom (causal) rather than
- symptom → cause (diagnostic)

- Need to specify fewer numbers and numbers are easier to get

- Ex: Alarm → MaryCalls means have to specify p(MaryCalls|Alarm) which is a lot easier than p(Alarm|MaryCalls) for MaryCalls → Alarm

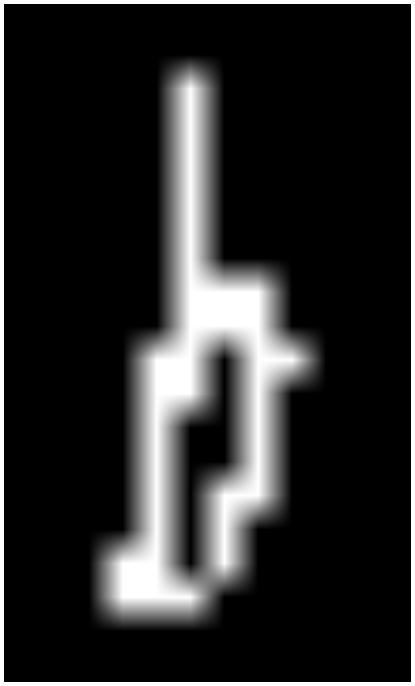
Sequential data

- Often we have data that is sampled in time or space

Sequential data

- Measurement of time series
- Example: Sign recognition

- Measure: drawn path
- Want: characters



Sequential data

- Measurement of time series
- Example: Activity recognition

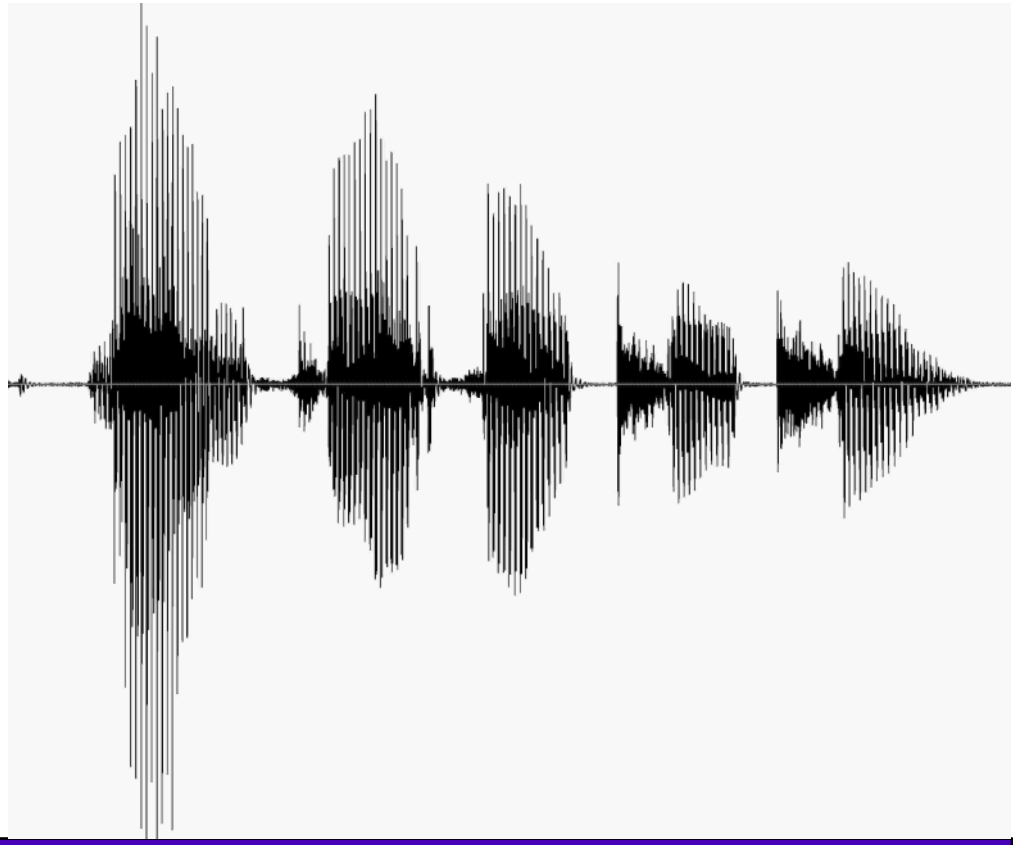
- Measure: images
- Want: what happens?

Sequential data

- Measurement of time series
- Example: Speech recognition

- Measure: audio signal
- Want: Words/sentences

- **DD2118 Speech and speaker recognition**



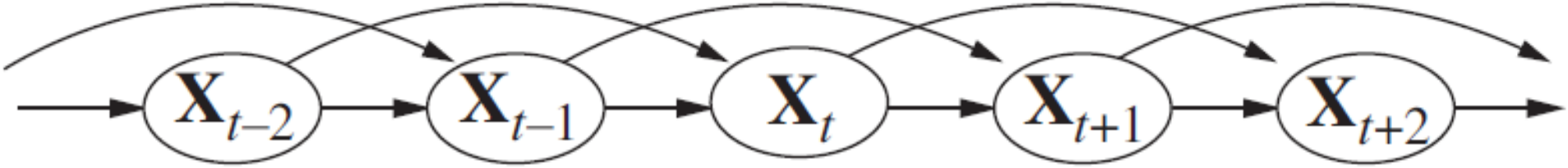
Let's get cracking with the theory again

Sequential data
Markov model

- The (first order) Markov assumption
  - The present (current state) can be predicted using local knowledge of the past (state at the previous step)
  - X_t is conditionally independent of all X_k ,k=t-2, …, given X_{t-1} ,i.e. $p(X_t|X_{t-1},X_{t-2},X_{t-3},\ldots) = p(X_t|X_{t-1})$

Second-order Markov Model

- State at time k depends on the states at times k-1 and k-2



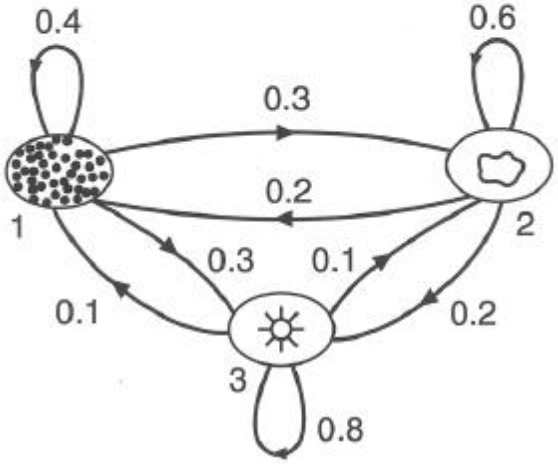
Example: Weather prediction

Once each day (e.g., at noon), the weather is observed and classified as being one of the following:

- State 1—Rain (or Snow; e.g. precipitation)
- State 2—Cloudy
- State 3—Sunny

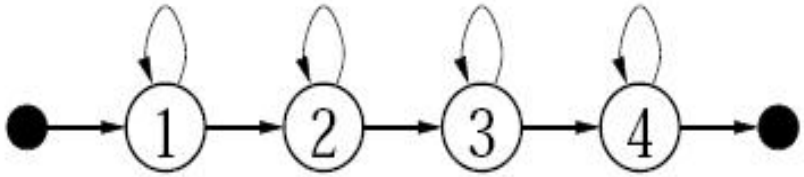
with state transition probabilities:

$$A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$



- Ex: 30% chance to transition from state Rain to state Cloudy (and Sunny).

Calculation example



Transition probabilities:

$$A = \{a_{ij}\} = \begin{bmatrix} 0.6 & 0.4 & 0 & 0 \\ 0 & 0.9 & 0.1 & 0 \\ 0 & 0 & 0.2 & 0.8 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}$$

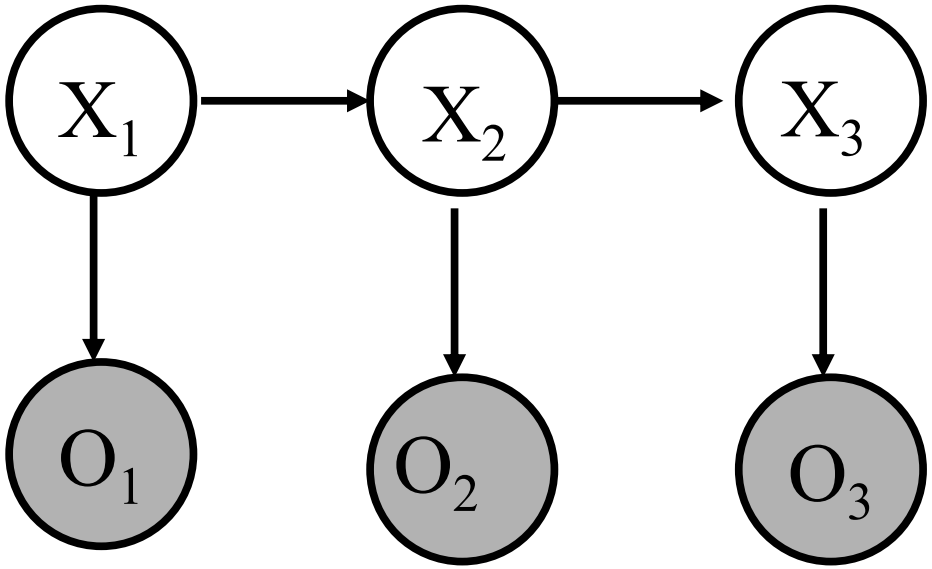
We start in state 1 at t=1

- p(X(t=2)=3) = ?
- p(X(t=2)=2) = ?
- p(X(t=3)=2) = ?

What if we cannot observe the state?

- Examples:
  - Cannot observe the weather, only the temperature
    - State=weather, observation=temperature
  - Cannot observe the words spoken, only the sound uttered
    - State=word, observation=sound
  - Cannot observe the position of the car only the laser scanner readings
    - State=position, observation=laser data

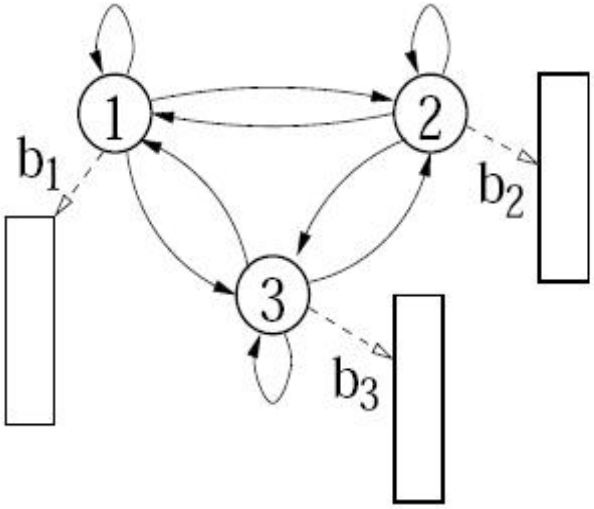
Hidden Markov Models (HMMs)



- Hidden states
- Observations

- State transition model: $p(X_t=j|X_{t-1}=i)=A(i,j)=a_{ij}$
- Observation model: $p(O_t=j|X_t=i)=b_{ij}$

Hidden Markov Models



Probability of state i producing a *discrete* observation o_t, which can have one of a finite set of values, is:

Elements of HMM

1. Number of states N, $x \in \{1, \ldots, N\}$;

2. Number of events K, $k \in \{1, \ldots, K\}$;

3. Initial-state probabilities,
   $\pi = \{\pi_i\} = \{P(x_1 = i)\}$ for $1 \leq i \leq N$;

4. State-transition probabilities,
   $A = \{a_{ij}\} = \{P(x_t = j | x_{t-1} = i)\}$ for $1 \leq i, j \leq N$;

5. Discrete output probabilities,
   $B = \{b_i(k)\} = \{P(o_t = k | x_t = i)\}$ for $1 \leq i \leq N$ and $1 \leq k \leq K$.

Elements of HMM

1. Number of states N, $x \in \{1, \ldots, N\}$;
2. Number of events K, $k \in \{1, \ldots, K\}$;

The number of possible observations types

3. Initial-state probabilities,
   $\pi = \{\pi_i\} = \{P(x_1 = i)\}$ for $1 \leq i \leq N$;
4. State-transition probabilities,
   $A = \{a_{ij}\} = \{P(x_t = j | x_{t-1} = i)\}$ for $1 \leq i, j \leq N$;
5. Discrete output probabilities,
   $B = \{b_i(k)\} = \{P(o_t = k | x_t = i)\}$ for $1 \leq i \leq N$ and $1 \leq k \leq K$.

Elements of HMM

1. Number of states N, $x \in \{1, \ldots, N\}$;

2. Number of events K, $k \in \{1, \ldots, K\}$;

3. Initial-state probabilities,
   $\pi = \{\pi_i\} = \{P(x_1 = i)\}$ for $1 \leq i \leq N$;

Note that we can start with the probability spread over several states

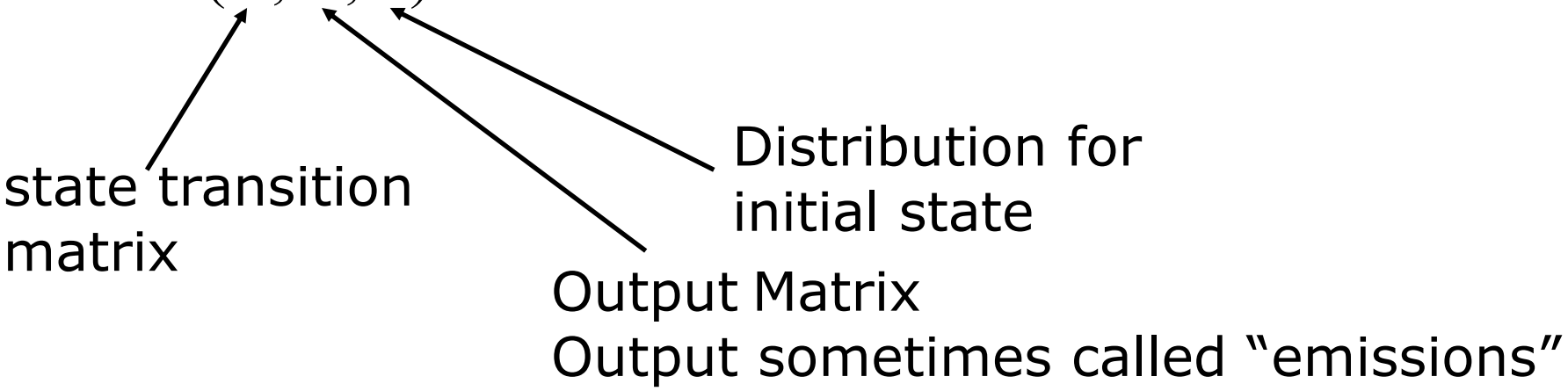
4. State-transition probabilities,
   $A = \{a_{ij}\} = \{P(x_t = j | x_{t-1} = i)\}$ for $1 \leq i, j \leq N$;

5. Discrete output probabilities,
   $B = \{b_i(k)\} = \{P(o_t = k | x_t = i)\}$ for $1 \leq i \leq N$ and $1 \leq k \leq K$.

The model if often called λ

- λ is the model, i.e.,

$$\lambda=(A, B, \pi)$$



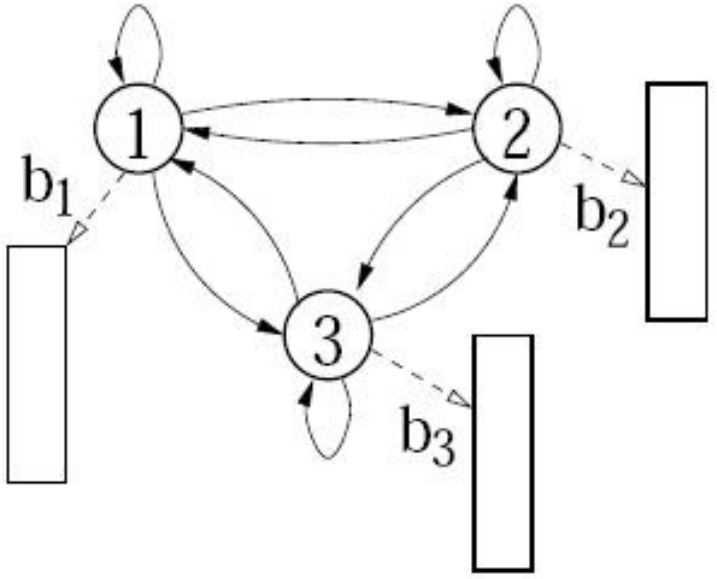
- λ is sometimes called *M*
- A, B and π are row-stochastic matrices (their rows sum to 1)

Elements of HMM, λ

- Initial Distribution : contains the probability of the (hidden) model being in a particular hidden state at time t = 1 (sometimes t=0).
- Often referred to as π

- Ex: π=[0.5 0.2 0.3], i.e.,
  - p(X =1)=0.5
  - p(X^1=2)=0.2
  - p(X^1=3)=0.3

1



Elements of HMM, λ

- State transition matrix : holding the probability of transitioning from one hidden state to another hidden state.
- Ex: a_{12} gives $p(X_{t+1}=1|X=2)$, i.e. probability to transition from state 2 to state 1

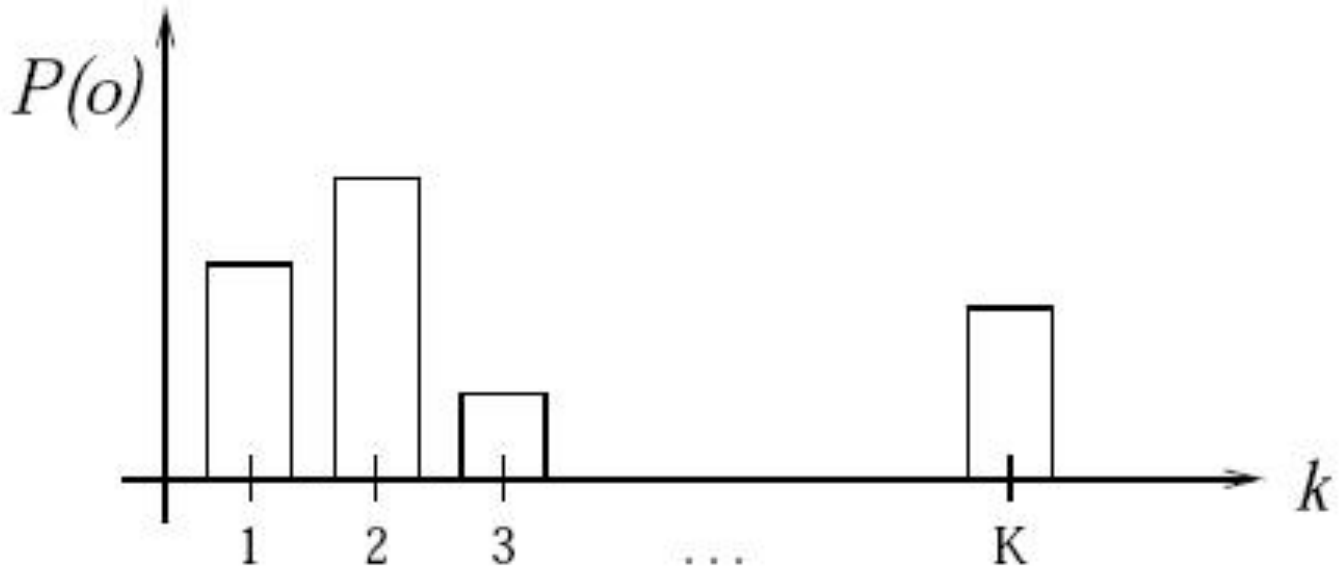
	$X_{t+1}=1$	$X_{t+1}=2$	...	$X_{t+1}=N$
$X_t=1$	a_{11}	a_{12}	...	a_{1N}
$X_t=2$	a_{21}	a_{22}	...	a_{2N}
...	...	...	...	...
$X_t=N$	a_{N1}	a_{N2}	...	a_{NN}

Elements of HMM, λ

- Output matrix : Contains the probability of observing a particular measurement given that the hidden model is in a particular hidden state.
- $b_i(O_t=j)$ is the probability to observe j in state i

	$O_t=1$	$O_t=2$	...	$O_t=K$
$X_t=1$	$b_1(1)$	$b_1(2)$	...	$b_1(K)$
$X_t=2$	$b_2(1)$	$b_2(2)$	...	$b_2(K)$
...	...	...	...	...
$X_t=N$	$b_N(1)$	$b_N()$	...	$b_N(K)$

Discrete output probabilities



- NOTE: The probability for a certain observation/event typically depends on the state we are in

What is an HMM, reeeallly?!?!?

- It is a **model** (not necessarily a perfect one) for system
- It can be used to (e.g.)
  - Generate predictions about how the system will behave
  - Analyze if a sequence of measurements match a certain model, e.g., did the person say "Bayesian" (i.e. match our model for that word) or "Bearnaise" (i.e. match that model)?
  - Learn something about a system by learning the model parameters.

A1

- At A-B level in A1 you will **learn models** for the motion of birds to be able to **predict** where they will be in the next time step. You will also identify what type of bird you see by comparing the observations to the models you built.

- Design a model, learn the parameters from data, make decisions, take actions

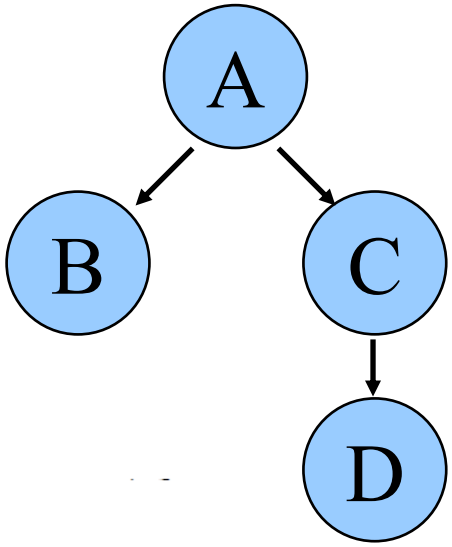
End of lecture

See exercises after this...

Exercise 1: Deriving Bayes rules

- Product rule
  - p(X,Y)=p(Y|X)p(X)
  - p(Y,X)=p(X|Y)p(Y)
- p(X,Y)=p(Y,X) (symmetry)
- p(Y|X)p(X) = p(X|Y)p(Y)
- → p(Y|X) = p(X|Y) p(Y) / p(X)

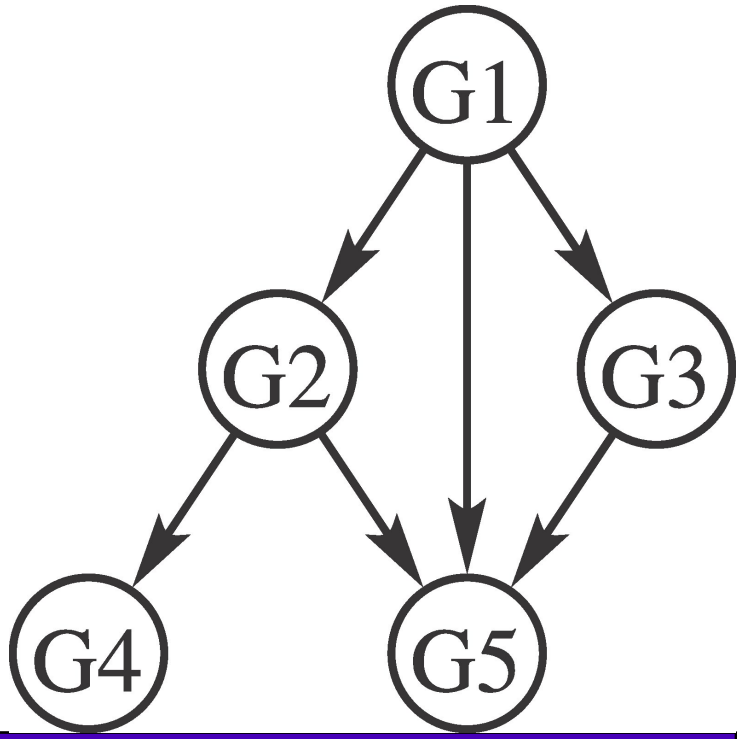
Exercise2: Factorization



$$
\begin{aligned}
p(A,B,C,D) &= \{\text{product rule with X=A, Y=B,C,D}\} \\
&= p(B,C,D|A)p(A) \\
&= \{\text{product rule with X=B, Y=C,D}\} \\
&= p(C,D|A,B)p(B|A,C,D)p(A) \\
&= \{\text{B conditionally independent of C,D given A}\} \\
&= p(C,D|A,B)p(B|A)p(A) \\
&= \{\text{product rule with X=C, Y=D}\} \\
&= p(D|A,B,C)p(C|A,B)p(B|A)p(A) \\
&= \{\text{C conditionally independent of B given A}\} \\
&= p(D|A,B,C)p(C|A)p(B|A)p(A) \\
&= \{\text{D conditionally independent of A and B given C}\} \\
&= p(D|C)p(C|A)p(B|A)p(A)
\end{aligned}
$$

Exercise3: Factorize the graph

- p(G1,G2,G3,G4,G5)
- =p(G1) p(G2|G1) p(G3|G1) p(G4|G2) p(G5|G1,G2,G3)



Exercie4: Alarm example calculation

- $p(J,M,A,\neg B,\neg E)$ = {use graph!!}
- $= p(\neg B)p(\neg E)p(A|\neg B,\neg E)p(J|A)p(M|A)$
- = {read from CPT and use $p(\neg X) = 1-p(X)$}
- $= (1-0.001) * (1-0.002) * 0.001 * 0.9 * 0.7$
- $= 0.00062$