PGM 1 — Representation

Qinfeng (Javen) Shi

ML session 09

Table of Contents I

1. Probability (simplified)
   - Dice rolling and probabilities
   - Random variables and marginals
   - Independence
2. Probabilistic Graphical Models
   - History and books
   - Representations
   - Factorisation and independences
3. Reasoning Bayesian Networks by Hand
   - Factorisation
   - Reasoning
   - A universal way
4. Appendix: Probability (advanced)

Dice rolling game

Rolling a die (with numbers $1, ..., 6$).
Chance of getting a $5 = ?$

Dice rolling game

Rolling a die (with numbers $1, ..., 6$).
Chance of getting a $5 = ?$
$1/6$
Chance of getting a 5 or $4 = ?$

Dice rolling game

Rolling a die (with numbers $1, ..., 6$).
Chance of getting a $5 = ?$
$1/6$
Chance of getting a 5 or $4 = ?$
$2/6$

Dice rolling game

Rolling a die (with numbers $1, ..., 6$).
Chance of getting a 5 $=$?
1/6
Chance of getting a 5 or 4 $=$?
2/6

Probability $\approx$ a degree of confidence that an **outcome** or a number of outcomes (called **event**) will occur.

Random Variables

Assigning probabilities to events is intuitive (defer the formal treatment to the appendix).

Assigning probabilities to attributes (of the outcome) taking various values might be more convenient.

- a patient's attributes such "Age", "Gender" and "Smoking history" ...
  "Age = 10", "Age = 50", ..., "Gender = male","Gender = female"
- a student's attributes "Grade", "Intelligence", "Gender" ...

$P(Grade = A)$ = the probability that a student gets a grade of A.

Random Variables

Random Variable[1] can take different types of values *e.g.* discrete or continuous.

- $Val(X)$: the set of values that X can take
- x: a value $x \in Val(X)$

Shorthand notation:

- $P(x)$ short for $P(X = x)$
- $\sum_x P(x)$ shorthand for $\sum_{x \in Val(X)} P(X = x)$

$$\sum_x P(x) = 1$$

[1] formal definition is omitted

Example

P(Grade, Intelligence).
Grade $\in \{A, B, C\}$
Intelligence $\in \{high, low\}$.

P(Grade = B, Intelligence = high) = ?
P(Grade = B) =?

		Intelligence		
		low	*high*	
Grade	*A*	0.07	0.18	0.25
	B	0.28	0.09	0.37
	C	0.35	0.03	0.38
		0.7	0.3	1

Marginal and Conditional distribution

Distributions:

- Marginal distribution $P(X) = \sum_{y \in Val(Y)} P(X, Y = y)$ or shorthand as $P(x) = \sum_y P(x, y)$
- Conditional distribution $P(X|Y) = \frac{P(X,Y)}{P(Y)}$

Rules for events carry over for random variables:

- Chain rule: $P(X, Y) = P(X)P(Y|X)$
- Bayes' rule: $P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$

Independence and conditional independence

Independences give factorisation.

- Independence
  $X \perp\!\!\!\perp Y \Leftrightarrow P(X, Y) = P(X)P(Y)$
  - Extension: $X \perp\!\!\!\perp Y, Z$ means $X \perp\!\!\!\perp H$ where $H = (Y, Z)$.
    $\Leftrightarrow P(X, Y, Z) = P(X)P(Y, Z)$
- Conditional Independence
  $X \perp\!\!\!\perp Y|Z \Leftrightarrow P(X, Y|Z) = P(X|Z)P(Y|Z)$
  - Independence: $X \perp\!\!\!\perp Y$ can be considered as $X \perp\!\!\!\perp Y|\emptyset$

Properties

For conditional independence:

- Symmetry: $X \perp\!\!\!\perp Y|Z \Rightarrow Y \perp\!\!\!\perp X|Z$
- Decomposition: $X \perp\!\!\!\perp Y, W|Z \Rightarrow X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp W|Z$
- Weak union: $X \perp\!\!\!\perp Y, W|Z \Rightarrow X \perp\!\!\!\perp Y|Z, W$
- Contraction: $X \perp\!\!\!\perp W|Z, Y$ and $X \perp\!\!\!\perp Y|Z \Rightarrow X \perp\!\!\!\perp Y, W|Z$
- Intersection:$X \perp\!\!\!\perp Y|W, Z$ and $X \perp\!\!\!\perp W|Y, Z \Rightarrow X \perp\!\!\!\perp Y, W|Z$

For independence: let $Z = \emptyset$ *e.g.*
$X \perp\!\!\!\perp Y \Rightarrow Y \perp\!\!\!\perp X$
$X \perp\!\!\!\perp Y, W \Rightarrow X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp W$
...

Marginal and MAP Queries

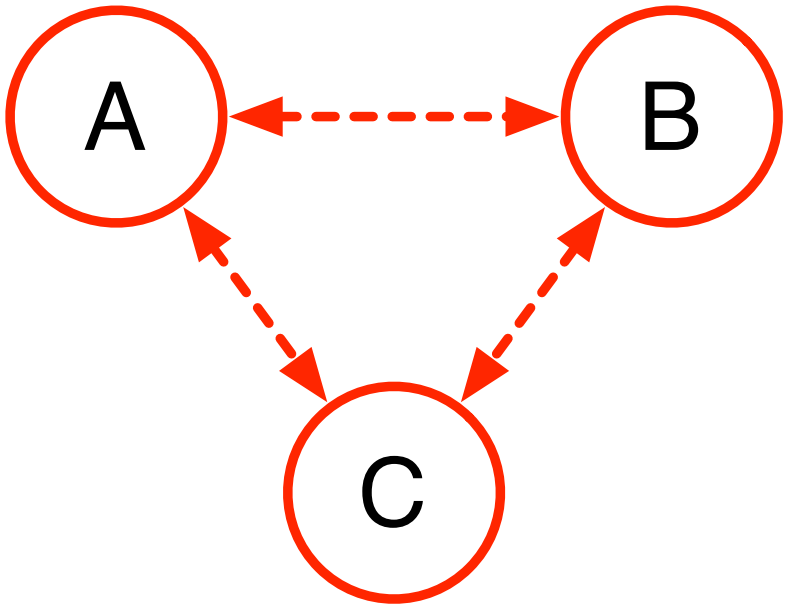
Given joint distribution $P(Y, E)$, where

- Y, query random variable(s), unknown
- E, evidence random variable(s), observed *i.e.* $\mathrm{E} = \mathrm{e}$.

Two types of queries:

- Marginal queries (a.k.a. probability queries)
  task is to compute $P(Y|E = e)$
- MAP queries (a.k.a. most probable explanation)
  task is to find $y^* = \mathrm{argmax}_{y \in Val(Y)} P(Y|E = e)$

Scenario 1



Multiple problems $(A, B, ...)$ affect each other

Joint optimal solution of all $\neq$ the solutions of individuals

Scenario 2

Two variables X, Y each taking 10 possible values.
Listing $P(X, Y)$ for each possible value of X, Y requires specifying/computing 10^2 many probabilities.

Scenario 2

Two variables X, Y each taking 10 possible values.
Listing $P(X, Y)$ for each possible value of X, Y requires specifying/computing 10^2 many probabilities.

What if we have 1000 variables each taking 10 possible values?

Scenario 2

Two variables X, Y each taking 10 possible values.
Listing $P(X, Y)$ for each possible value of X, Y requires specifying/computing 10^2 many probabilities.

What if we have 1000 variables each taking 10 possible values?
$\Rightarrow 10^{1000}$ many probabilities

$\Rightarrow$ Difficult to store, and query naively.

Remedy

Structured Learning, specially Probabilistic Graphical Models (PGMs).

PGMs

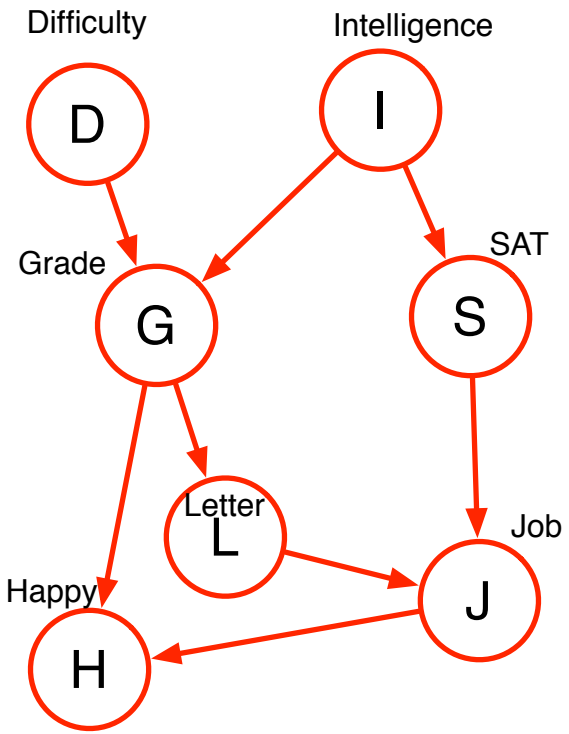
PGMs use graphs to represent the complex probabilistic relationships between random variables.

$$P(A, B, C, ...)$$

Benefits:

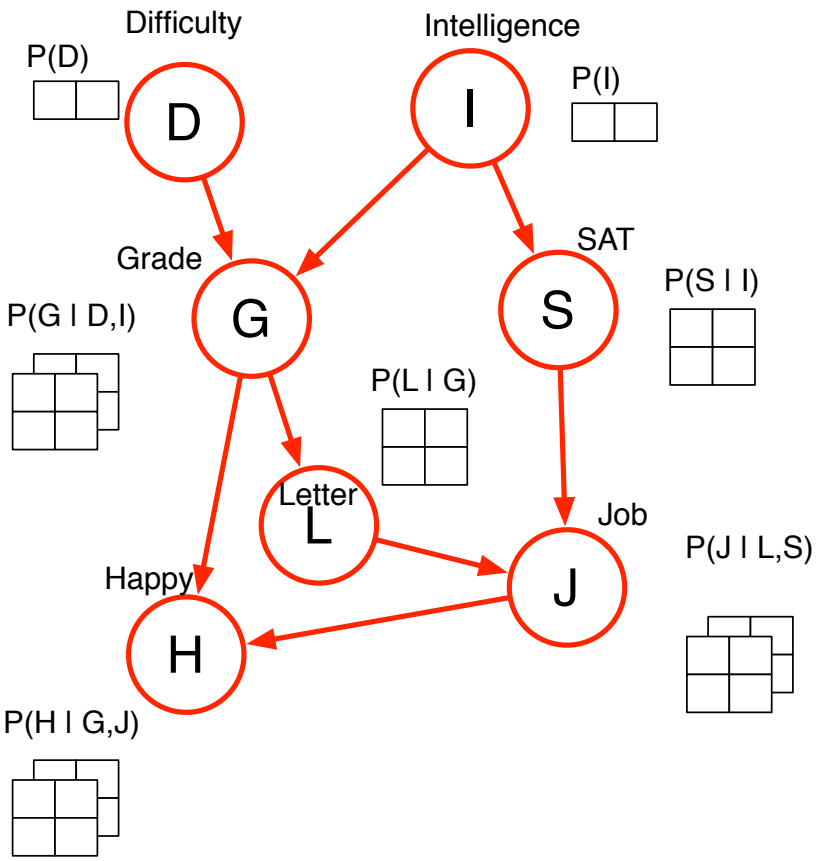
- compactly represent distributions of variables.
- Relation between variables are intuitive (such as conditional independences)
- have fast and general algorithms to query without enumeration. *e.g.* ask for $P(A|B = b, C = c)$ or $\mathbb{E}_P[f]$

An Example



Intuitive

Example



Compact

History

- Gibbs (1902) used undirected graphs in particles
- Wright (1921,1934) used directed graph in genetics
- In economists and social sci (Wold 1954, Blalock, Jr. 1971)
- In statistics (Bartlett 1935, Vorobev 1962, Goodman 1970, Haberman 1974)
- In AI, expert system (Bombal *et al.* 1972, Gorry and Barnett 1968, Warner *et al.* 1961)
- Widely accepted in late 1980s. Prob Reasoning in Intelli Sys (Pearl 1988), Pathfinder expert system (Heckerman *et al.* 1992)

History

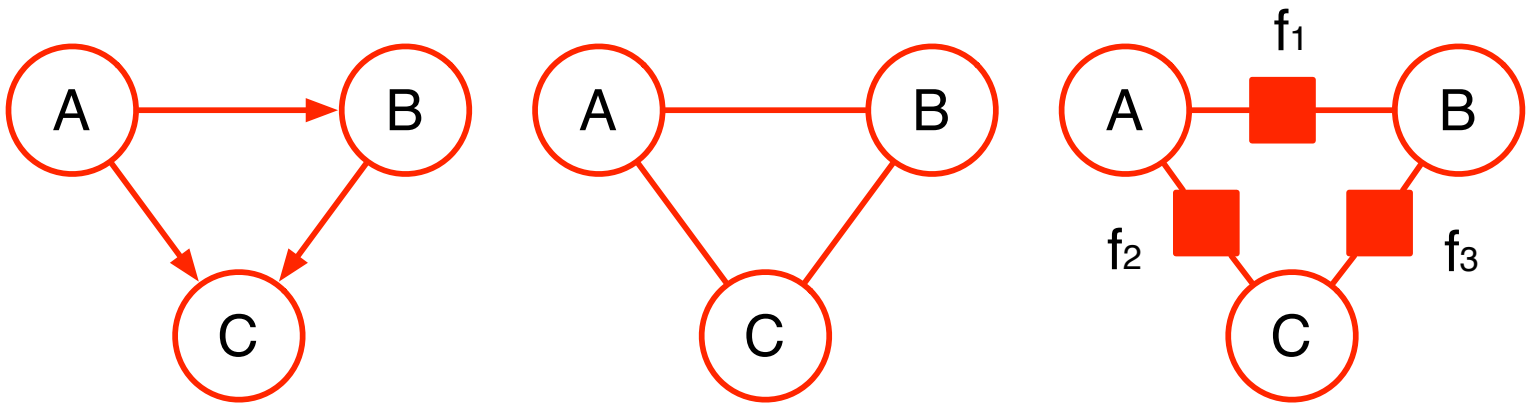
- Hot since 2001-2013. Flexible features and principled ways of learning.
  CRFs (Lafferty *et al.* 2001), SVM struct (Tsochantaridis etal 2004), M^3Net (Taskar *et al.* 2004), DeepBeliefNet (Hinton *et al.* 2006)
- Fall of graphical models and rise of deep Learning 2013-2016.
  Deep learning won a large number of challenges. Reflection at ICCV 2013 and ECCV 2014 PGM workshops.
- Marriage of graphical models and deep Learning since 2017.
  Drawbacks of deep learning become apparent, and graphical models inspired, math and reasoning driven deep learning is the new trend.

History

Good books

- Chris Bishop's book "Pattern Recognition and Machine Learning" (Graphical Models are in chapter 8, which is available from his webpage) $\approx$ 60 pages
- Koller and Friedman's "Probabilistic Graphical Models" $>$ 1000 pages
- Stephen Lauritzen's "Graphical Models"
- Michael Jordan's unpublished book "An Introduction to Probabilistic Graphical Models"
- ...

Representations



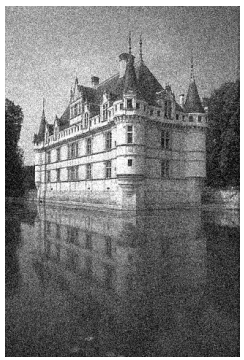
(a) Directed graph (b) Undirected graph (c) Factor graph

- Nodes represent random variables
- Edges reflect dependencies between variables
- Factors explicitly show which variables are used in each factor *i.e.* $f_1(A, B)f_2(A, C)f_3(B, C)$

Example — Image Denoising

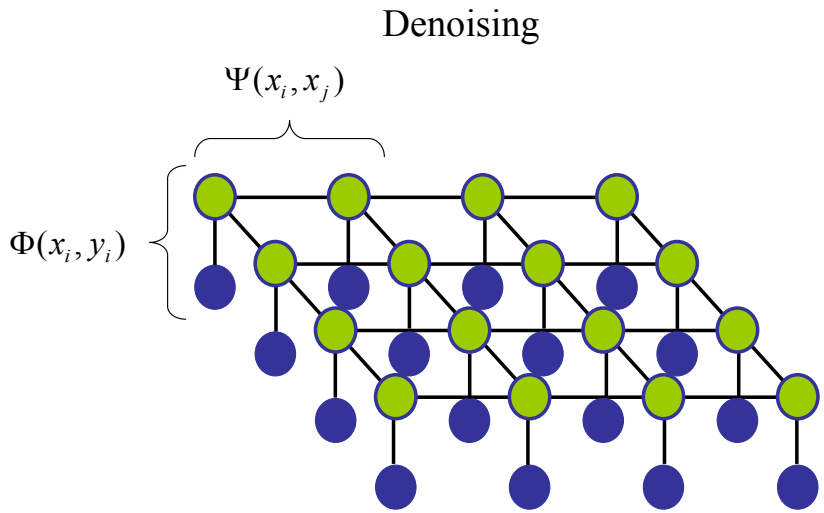
Denoising[2]

Original



Noisy

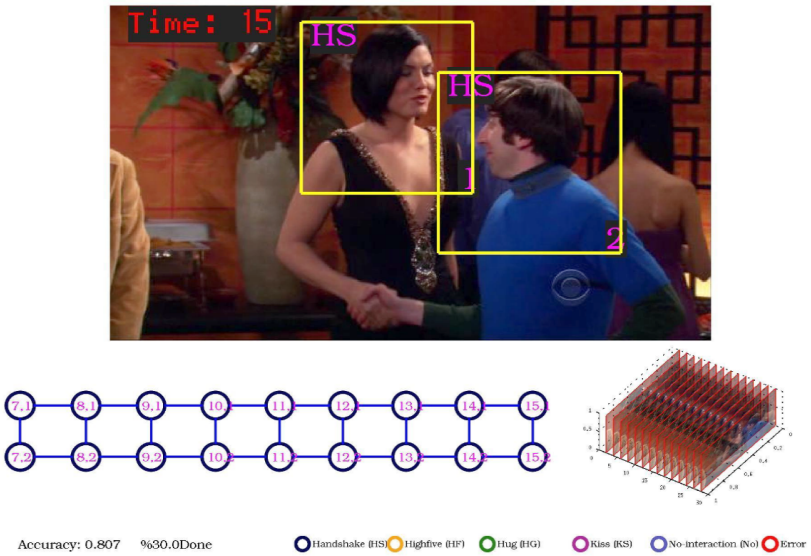
Corrected

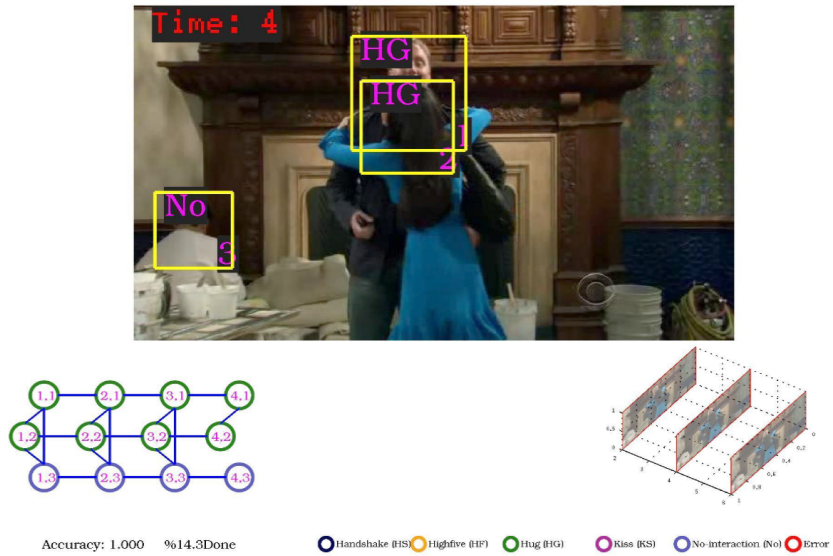


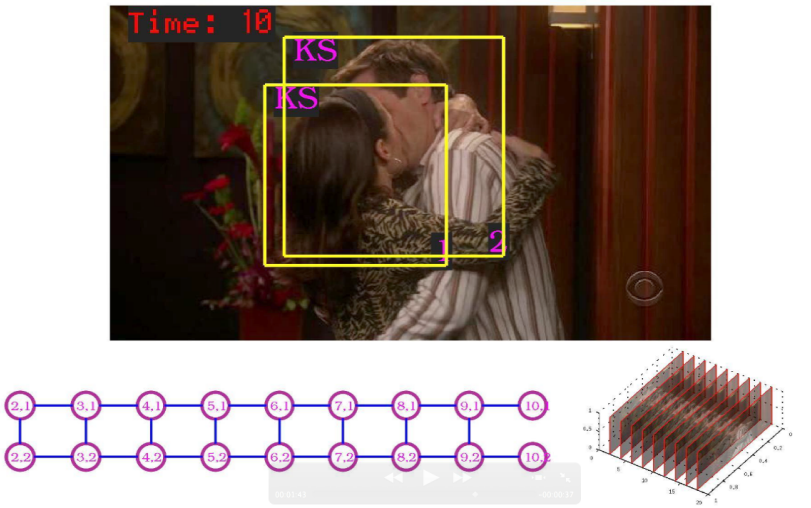
$$X^* = \text{argmax}_X \, P(X|Y)$$

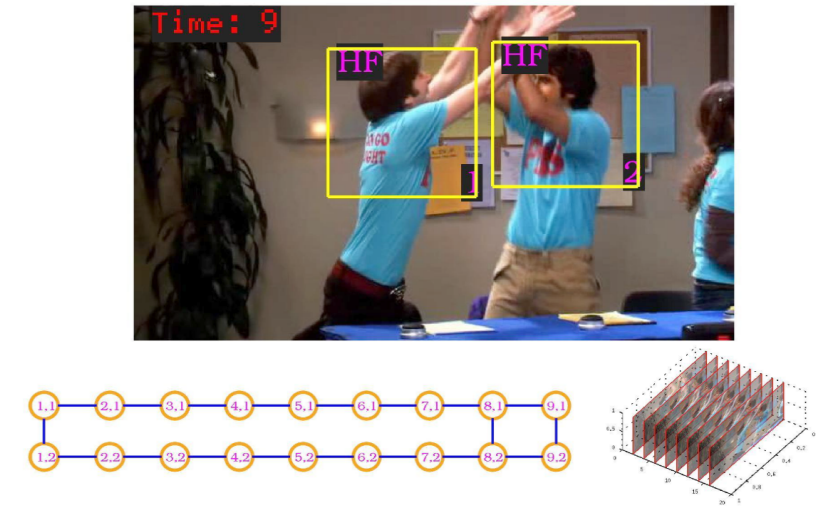
[2]This example is from Tiberio Caetano's short course: "Machine Learning using Graphical Models"

Example —Human Interaction Recognition



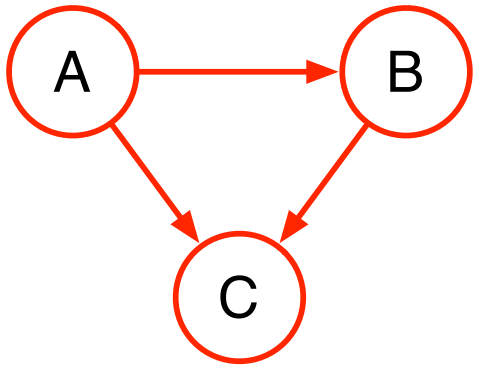






Factorisation for Bayesian networks

Directed Acyclic Graph (DAG).
Factorisation rule: $P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Pa(x_i))$
$Pa(x_i)$ denotes parent of x_i. *e.g.* $(A, B) = Pa(C)$



$\Rightarrow P(A, B, C) = P(A)P(B|A)P(C|A, B)$
Acyclic: no cycle allowed. Replacing edge $A \to C$ with $C \to A$ will form a cycle (loop *i.e.* $A \to B \to C \to A$), not allowed in DAG.

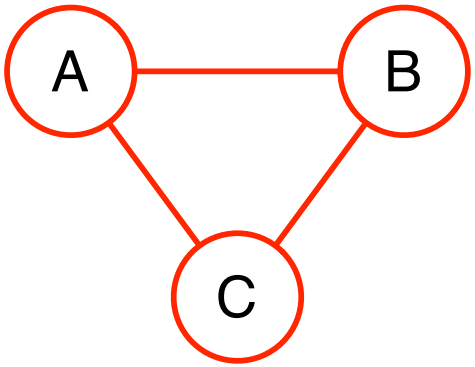
Factorisation for Markov Random Fields

Undirected Graph:

Factorisation rule: $P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{X}_c)$,

$Z = \sum_{\mathbf{X}} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{X}_c)$,

where c is an index set of a clique (fully connected subgraph), $\mathbf{X}_c$ is the set of variables indicated by c.



Consider $\mathbf{X}_{c_1} = \{A, B\}, \mathbf{X}_{c_2} = \{A, C\}, \mathbf{X}_{c_3} = \{B, C\}$

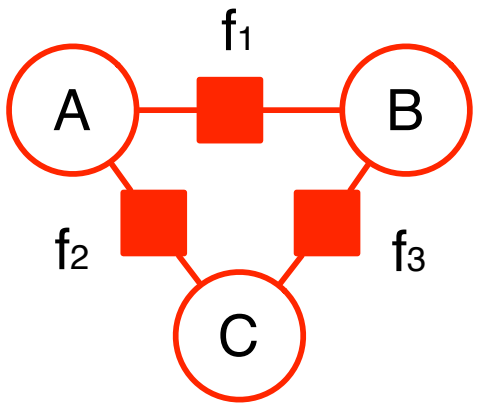
$\Rightarrow P(A, B, C) = \frac{1}{Z} \psi_{c_1}(A, B) \psi_{c_2}(A, C) \psi_{c_3}(B, C)$

Consider $\mathbf{X}_c = \{A, B, C\} \Rightarrow P(A, B, C) = \frac{1}{Z} \psi_c(A, B, C)$,

Factorisation for Markov Random Fields

Factor Graph:

Factorisation rule: $P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_i f_i(\mathbf{X}_i)$, $Z = \sum_{\mathbf{X}} \prod_i f_i(\mathbf{X}_i)$



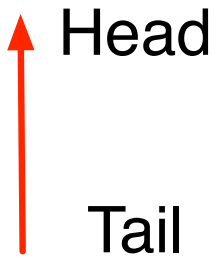
$$\Rightarrow P(A, B, C) = \frac{1}{Z} f_1(A, B) f_2(A, C) f_3(B, C)$$

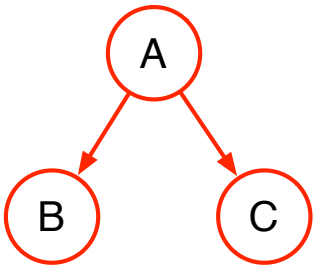
Independences

- Independence
  $A \perp\!\!\!\perp B \Leftrightarrow P(A, B) = P(A)P(B)$
- Conditional Independence
  $A \perp\!\!\!\perp B|C \Leftrightarrow P(A, B|C) = P(A|C)P(B|C)$

From Graph to Independences

Case 1:

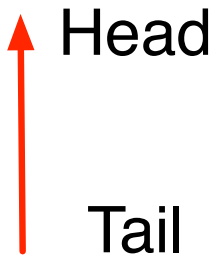


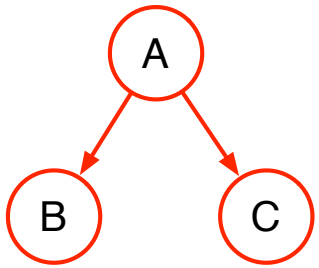


Question: $B \perp\!\!\!\perp C$?

From Graph to Independences

Case 1:



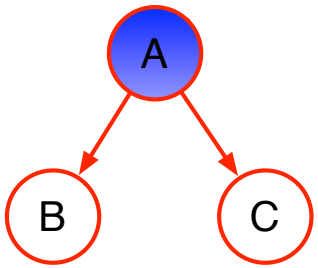


Question: $B \perp\!\!\!\perp C$?
Answer: No.

$$P(B,C) = \sum_A P(A,B,C)$$
$$= \sum_A P(B|A)P(C|A)P(A)$$
$$\neq P(B)P(C) \text{ in general}$$

From Graph to Independences

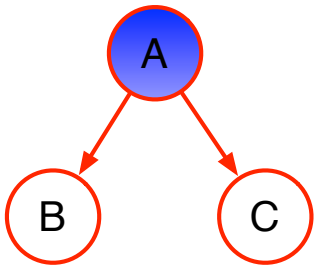
Case 2:



Question: $B \perp\!\!\!\perp C | A$?

From Graph to Independences

Case 2:

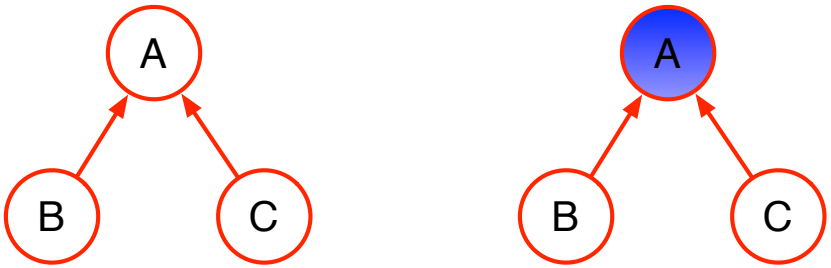


Question: $B \perp\!\!\!\perp C|A$?
Answer: Yes.

$$\begin{aligned} P(B, C|A) &= \frac{P(A, B, C)}{P(A)} \\ &= \frac{P(B|A)P(C|A)P(A)}{P(A)} \\ &= P(B|A)P(C|A) \end{aligned}$$

From graphs to independences

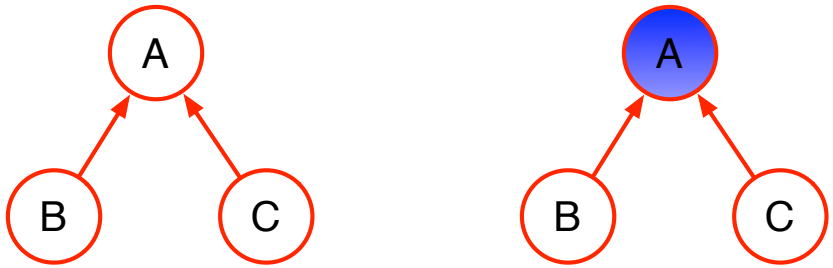
Case 3:



Question: $B \perp\!\!\!\perp C$, $B \perp\!\!\!\perp C | A$?

From graphs to independences

Case 3:



Question: $B \perp\!\!\!\perp C$, $B \perp\!\!\!\perp C|A$?

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

$$\therefore P(B,C) = \sum_A P(A,B,C)$$

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

$$= P(B)P(C)$$

Bayesian networks

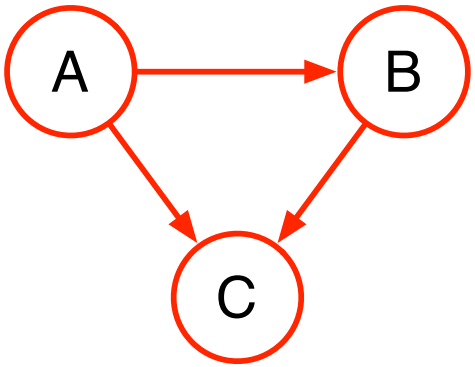
Directed Acyclic Graph (DAG).

Factorisation rule: $P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Pa(x_i))$

Bayesian networks

Directed Acyclic Graph (DAG).
Factorisation rule: $P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Pa(x_i))$

Example:



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

Reasoning with all variables

DAG tells us: $P(A, B, C) = P(A)P(B|A)P(C|A, B)$

$$P(A = a|B = b, C = c) = ?$$

All variables are involved in the query.

Reasoning with all variables

DAG tells us: $P(A, B, C) = P(A)P(B|A)P(C|A, B)$

$$P(A = a|B = b, C = c) = ?$$

All variables are involved in the query.

$$= \frac{P(A = a, B = b, C = c)}{P(B = b, C = c)}$$

$$= \frac{P(A = a)P(B = b|A = a)P(C = c|A = a, B = b)}{\sum_{A \in \{\neg a, a\}} P(A, B = b, C = c)}$$

$$= \frac{P(A = a)P(B = b|A = a)P(C = c|A = a, B = b)}{\sum_{A \in \{\neg a, a\}} P(A)P(B = b|A)P(C = c|A, B = b)}$$

$$= \frac{P(A = a)P(B = b|A = a)P(C = c|A = a, B = b)}{P(A = \neg a)P(B = b|A = \neg a)P(C = c|A = \neg a, B = b) + P(A = a)P(B = b|A = a)P(C = c|A = a, B = b)}$$

Reasoning with missing variable(s)

DAG tells us: $P(A, B, C) = P(A)P(B|A)P(C|A, B)$

$$P(A = a|B = b) = ?$$

C is missing in the query.

Reasoning with missing variable(s)

DAG tells us: $P(A, B, C) = P(A)P(B|A)P(C|A, B)$

$$P(A = a|B = b) = ?$$

C is missing in the query.

$$\begin{aligned}
&= \frac{P(A = a, B = b)}{P(B = b)} \\
&= \frac{\sum_C P(A = a, B = b, C)}{\sum_{A,C} P(A, B = b, C)} \\
&= \frac{\sum_C P(A = a)P(B = b|A = a)P(C|A = a, B = b)}{\sum_{A,C} P(A)P(B = b|A)P(C|A, B = b)}
\end{aligned}$$

Example of 4WD

Someone finds that people who drive 4WDs vehicles (S) consume large amounts of gas (G) and are involved in more accidents than the national average (A). They have constructed the Bayesian network below (here t implies "true" and f implies "false").

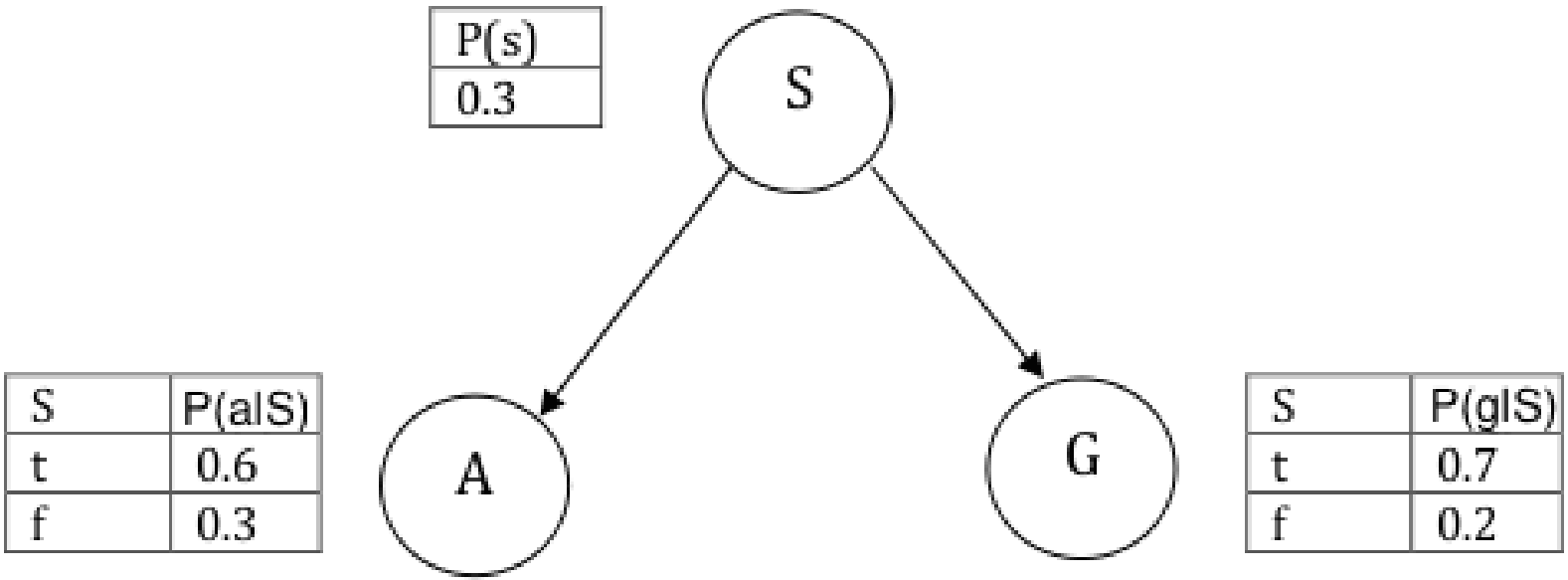
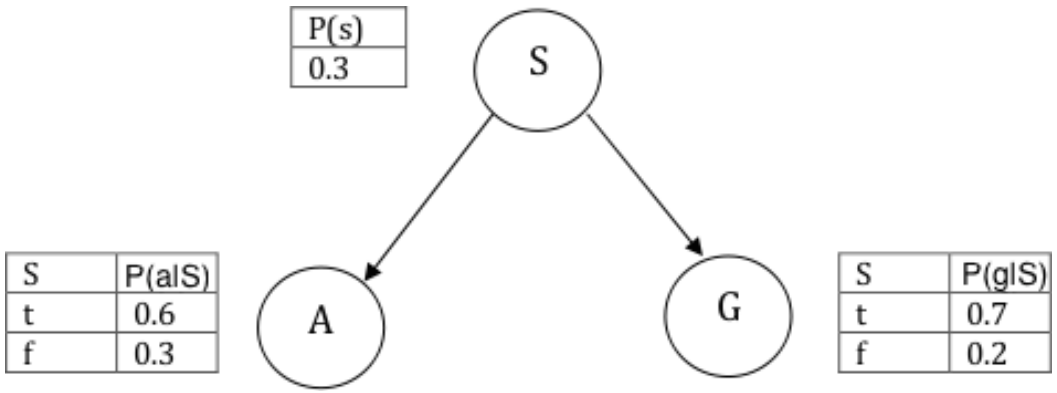


Figure : 4WD Bayesian network

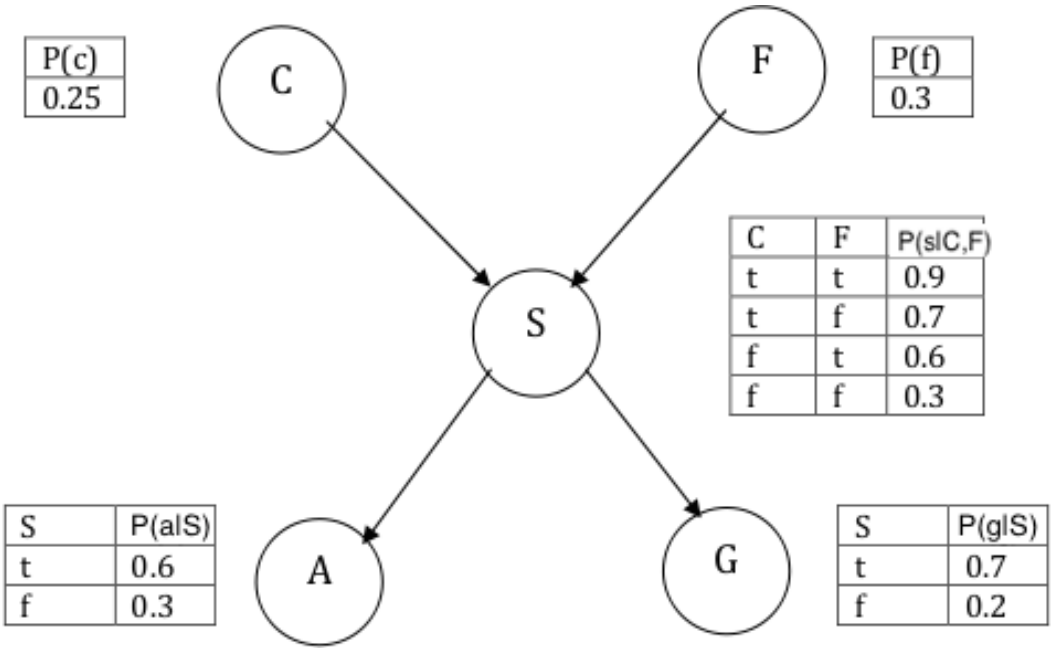
Example of 4WD



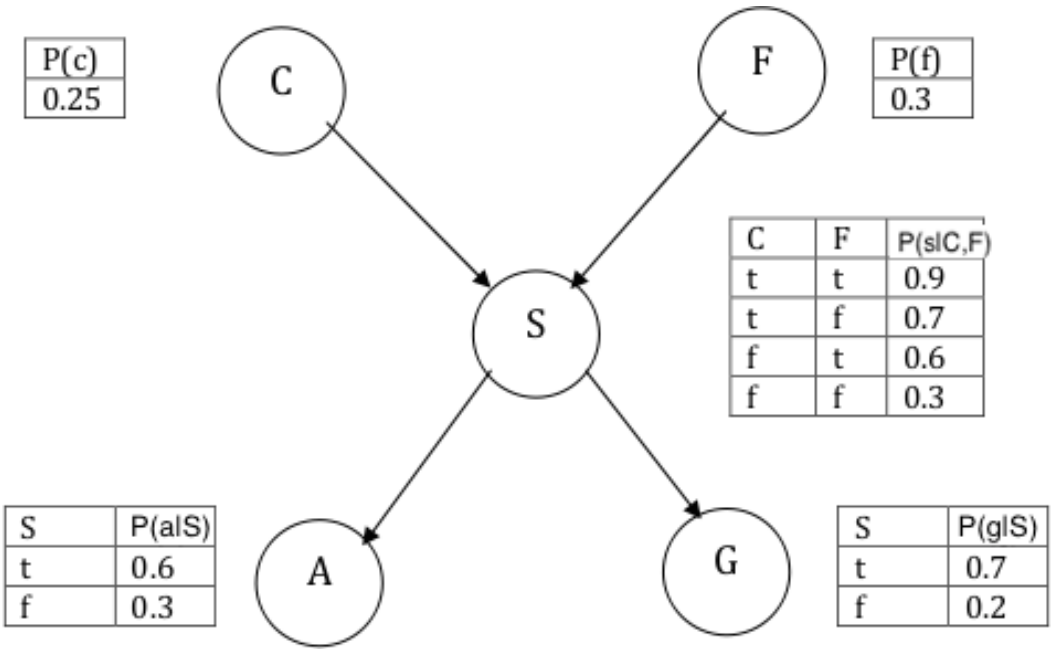
- $P(\neg g, a|s)$? (*i.e.* $P(G = \neg g, A = a|S = s)$)
- $P(a|s)$?
- $P(A|s)$?

Example of 4WD

Someone else finds that there are two types of people that drive 4WDs, people from the country (C) and people with large families (F). After collecting some statistics, here is the new Bayesian network.



Example of 4WD



- $P(g, \neg a, s, c)$?
- $P(g, \neg a | c, f)$?

How to reason by hand?

[3] the joint distribution over all variables of your Bayesian network

How to reason by hand?

A universal way: express the query probability in terms of the **full distribution**[3], and then **factorise** it.

[3] the joint distribution over all variables of your Bayesian network

How to reason by hand?

A universal way: express the query probability in terms of the **full distribution**[3], and then **factorise** it.

Step by step:

1. when you see a conditional distribution, break it into the nominator and the denominator.

[3] the joint distribution over all variables of your Bayesian network

How to reason by hand?

A universal way: express the query probability in terms of the **full distribution**[3], and then **factorise** it.

Step by step:

1. when you see a conditional distribution, break it into the nominator and the denominator.
2. when you see a distribution (may be from the nominator and/or the denominator) with missing variable(s), rewrite it as a sum of the full distribution w.r.t. the missing variable(s).

[3] the joint distribution over all variables of your Bayesian network

How to reason by hand?

A universal way: express the query probability in terms of the **full distribution**[3], and then **factorise** it.

Step by step:

1. when you see a conditional distribution, break it into the nominator and the denominator.
2. when you see a distribution (may be from the nominator and/or the denominator) with missing variable(s), rewrite it as a sum of the full distribution w.r.t. the missing variable(s).
3. After everything is expressed by the full distribution, factorise the full distribution into local distributions (which are known).

[3] the joint distribution over all variables of your Bayesian network

Probability Space

Probability space (a.k.a Probability triple) $(\Omega, \mathcal{F}, P)$:

- **Outcome space** (or sample space), denoted Ω (read "Omega") : the set of all possible outcomes[4].
  - roll a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$. flip a coin: $\Omega = \{Head, Tail\}$.
- **σ-Field** (read "sigma-field", a set of events), denoted $\mathcal{F}$: An event ($\alpha \in \mathcal{F}$) is a set of outcomes.
  - roll a die to get 1: $\alpha = \{1\}$;
  - to get 1 or 3: $\alpha = \{1, 3\}$
  - roll a die to get an even number: $\alpha = \{2, 4, 6\}$
- **Probability measure** P: the assignment of probabilities to the events; *i.e.* a function returning an event's probability.

[4] of the problem that you are considering

Probability measure

Probability measure (or distribution) P over $(\Omega, \mathcal{F})$: a function from $\mathcal{F}$ (events) to $[0, 1]$ (range of probabilities), such that,

- $P(\alpha) \geq 0$ for all $\alpha \in \mathcal{F}$
- $P(\Omega) = 1$, $P(\emptyset) = 0$.
- For $\alpha, \beta \in \mathcal{F}$, $P(\alpha \cup \beta) = P(\alpha) + P(\beta) - P(\alpha \cap \beta)$

Interpretations of Probability

- Frequentist Probability: $P(\alpha) =$ frequencies of the event. *i.e.* fraction of times the event occurs if we repeat the experiment indefinitely.
  - A die roll: $P(\alpha) = 0.5$, for $\alpha = \{2, 4, 6\}$ means if we repeatedly roll this die and record the outcome, then the fraction of times the outcomes in α will occur is 0.5.

Interpretations of Probability

- Frequentist Probability: $P(\alpha) =$ frequencies of the event. *i.e.* fraction of times the event occurs if we repeat the experiment indefinitely.
  - A die roll: $P(\alpha) = 0.5$, for $\alpha = \{2, 4, 6\}$ means if we repeatedly roll this die and record the outcome, then the fraction of times the outcomes in α will occur is 0.5.
  - Problem: non-repeatable event *e.g.* "it will rain tomorrow morning" (tmr morning happens exactly once, can't repeat).
- Subjective Probability: $P(\alpha) =$ one's own degree of belief that the event α will occur.

Conditional probability

Event α: "students with grade A"
Event β: "students with high intelligence"
Event $\alpha \cap \beta$: "students with grade A and high intelligence"

Conditional probability

Event α: "students with grade A"
Event β: "students with high intelligence"
Event $\alpha \cap \beta$: "students with grade A and high intelligence"

Question: how do we update the our beliefs given new evidence?

Conditional probability

Event α: "students with grade A"
Event β: "students with high intelligence"
Event $\alpha \cap \beta$: "students with grade A and high intelligence"

Question: how do we update the our beliefs given new evidence?
e.g. suppose we learn that a student has received the grade A, what does that tell us about the person's intelligence?

Conditional probability

Event α: "students with grade A"
Event β: "students with high intelligence"
Event $\alpha \cap \beta$: "students with grade A and high intelligence"

Question: how do we update the our beliefs given new evidence?
e.g. suppose we learn that a student has received the grade A, what does that tell us about the person's intelligence?

Answer: Conditional probability.
Conditional probability of β given α is defined as

$$P(\beta|\alpha) = \frac{P(\alpha \cap \beta)}{P(\alpha)}$$

Chain rule and Bayes' rule

- Chain rule: $P(\alpha \cap \beta) = P(\alpha)P(\beta|\alpha)$
  More generally,
  $P(\alpha_1 \cap ... \cap \alpha_k) = P(\alpha_1)P(\alpha_2|\alpha_1) \cdots P(\alpha_k|\alpha_1 \cap ... \cap \alpha_{k-1})$
- Bayes' rule:

$$P(\alpha|\beta) = \frac{P(\beta|\alpha)P(\alpha)}{P(\beta)}$$

That's all

Thanks!