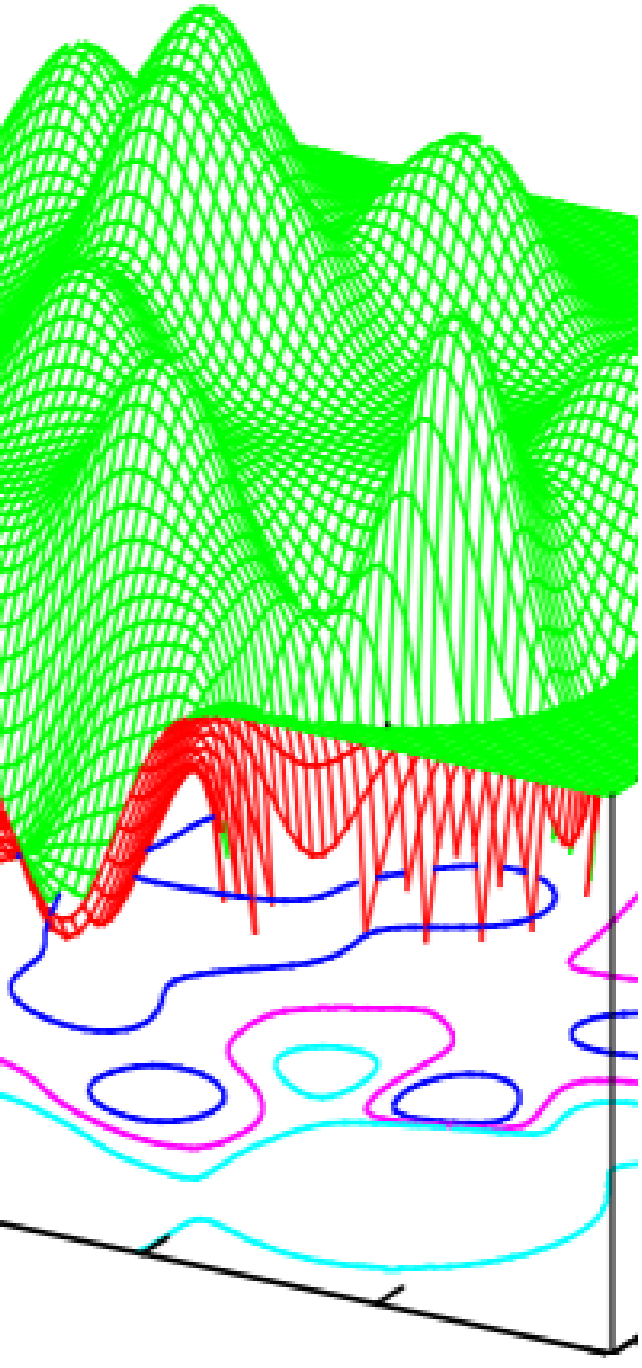


Machine Learning

Graphical Models

Bayesian Networks, examples, conditional independence, inference

Marc Toussaint
FU Berlin

The need for modelling

- Given a real world problem, translating it to a well-defined learning problem is non-trivial.
- The “framework” of plain regression/classification is rather restricted: input x, output y.
- Graphical models (probabilstic models with multiple random variables and dependencies) are a more general framework for modelling “problems”; regression & classification become a special case; Reinforcement Learning, decision making, but also language processing, image segmentation, are special cases.

Graphical Models

- The core difficulty in modelling is specifying

  What are the relevant variables?

  How do they depend on each other?

  (Or how *could* they depend on each other $\rightarrow$ learning)

Graphical Models

- The core difficulty in modelling is specifying

  What are the relevant variables?

  How do they depend on each other?

  (Or how *could* they depend on each other $\rightarrow$ learning)

- **Graphical models** are a simple, graphical notation for

  1) which random variables exist

  2) which random variables are "directly coupled"

  Thereby they *describe a joint probability distribution* $P(X_1, .., X_n)$ over n random variables.

Graphical Models

- The core difficulty in modelling is specifying

  What are the relevant variables?

  How do they depend on each other?

  (Or how *could* they depend on each other $\rightarrow$ learning)

- **Graphical models** are a simple, graphical notation for

  1) which random variables exist

  2) which random variables are "directly coupled"

  Thereby they *describe a joint probability distribution* $P(X_1, .., X_n)$ over n random variables.

- 2 basic variants:

  – Bayesian Networks (aka. directed model, belief network)

  – Factor Graphs (aka. undirected model, Markov Random Field)

Bayesian Networks

- A **Bayesian Network** is a
  – directed acyclic graph (DAG)
  – where each node represents a random variable X_i
  – for each node we have a conditional probability distribution
  $$P(X_i \mid \text{Parents}(X_i))$$

- In the simplest case (discrete RVs), the conditional distribution is represented as a conditional probability table (**CPT**)

Example

drinking red wine $\rightarrow$ longevity?

Bayesian Networks

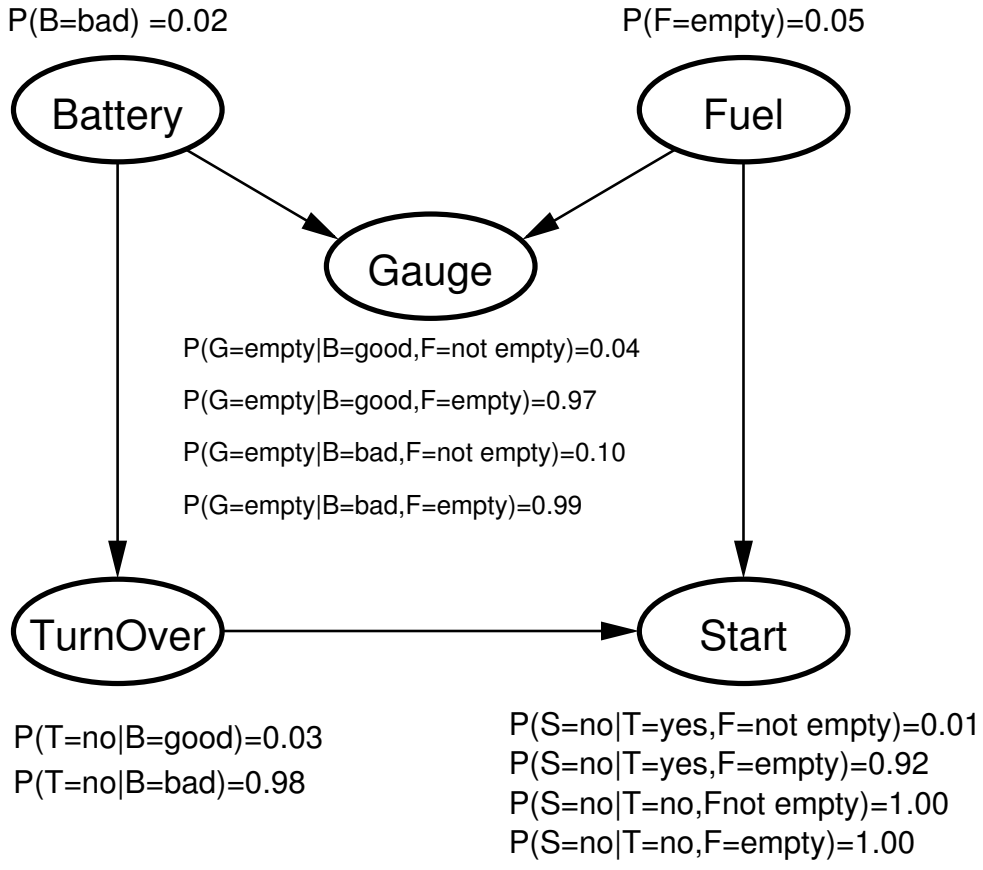
- DAG $\rightarrow$ we can sort the RVs; edges only go from lower to higher index
- **The joint distribution can be factored as**

$$P(X_{1:n}) = \prod_{i=1}^{n} P(X_i \mid \text{Parents}(X_i))$$

- Missing links imply conditional independence
- Ancestral simulation to sample from joint distribution

Example

(Heckermann 1995)



$$\iff \quad P(S,T,G,F,B) = P(B)\ P(F)\ P(G|F,B)\ P(T|B)\ P(S|T,F)$$

- table sizes: LHS = $2^5 - 1 = 31$ RHS = $1 + 1 + 4 + 2 + 4 = 12$

Constructing a Bayes Net

1. Choose a relevant set of variables X_i that describe the domain
2. Choose an ordering for the variables
3. While there are variables left
   (a) Pick a variable X_i and add it to the network
   (b) Set Parents(X_i) to some minimal set of nodes already in the net
   (c) Define the CPT for X_i

- This procedure is guaranteed to produce a DAG
- Different orderings may lead to **different Bayes nets representing the same joint distribution**:
- To ensure maximum sparsity choose a wise order (“root causes”first). Counter example: construct DAG for the car example using the ordering S, T, G, F, B
- “Wrong” ordering will give same joint distribution, but will require the specification of more numbers than otherwise necessary

Bayes Nets & conditional independence

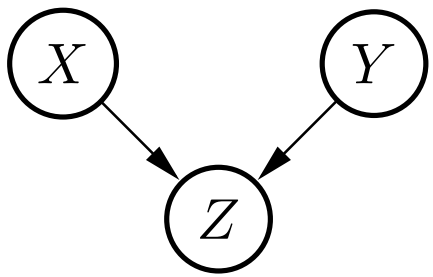
- Independence: $Indep(X, Y) \iff P(X, Y) = P(X)\ P(Y)$
- Conditional independence:

$$Indep(X, Y|Z) \iff P(X, Y|Z) = P(X|Z)\ P(Y|Z)$$

Bayes Nets & conditional independence

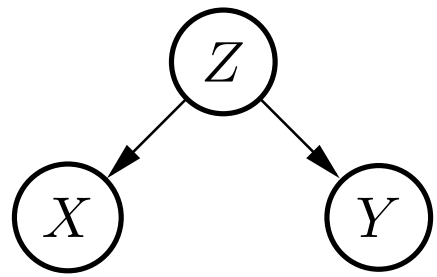
- Independence: $Indep(X, Y) \iff P(X, Y) = P(X)\ P(Y)$
- Conditional independence:

$$Indep(X, Y|Z) \iff P(X, Y|Z) = P(X|Z)\ P(Y|Z)$$



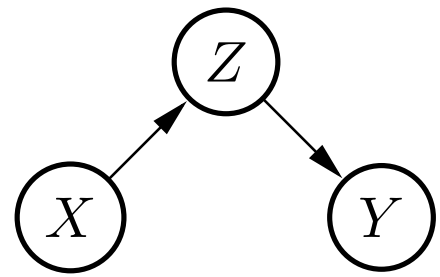
(head-to-head)

$$Indep(X, Y)$$
$$\neg Indep(X, Y|Z)$$



(tail-to-tail)

$$\neg Indep(X, Y)$$
$$Indep(X, Y|Z)$$



(head-to-tail)

$$\neg Indep(X, Y)$$
$$Indep(X, Y|Z)$$

- Head-to-head: $Indep(X, Y)$
  $P(X, Y, Z) = P(X)\ P(Y)\ P(Z|X, Y)$
  $P(X, Y) = P(X)\ P(Y)\ \sum_Z P(Z|X, Y) = P(X)\ P(Y)$

- Tail-to-tail: $Indep(X, Y|Z)$
  $P(X, Y, Z) = P(Z)\ P(X|Z)\ P(Y|Z)$
  $P(X, Y|Z) = P(X, Y, Z) = P(Z) = P(X|Z)\ P(Y|Z)$

- Head-to-tail: $Indep(X, Y|Z)$
  $P(X, Y, Z) = P(X)\ P(Z|X)\ P(Y|Z)$
  $P(X, Y|Z) = \frac{P(X,Y,Z)}{P(Z)} = \frac{P(X,Z)\ P(Y|Z)}{P(Z)} = P(X|Z)\ P(Y|Z)$

General rules for determining conditional independence in a Bayes net:

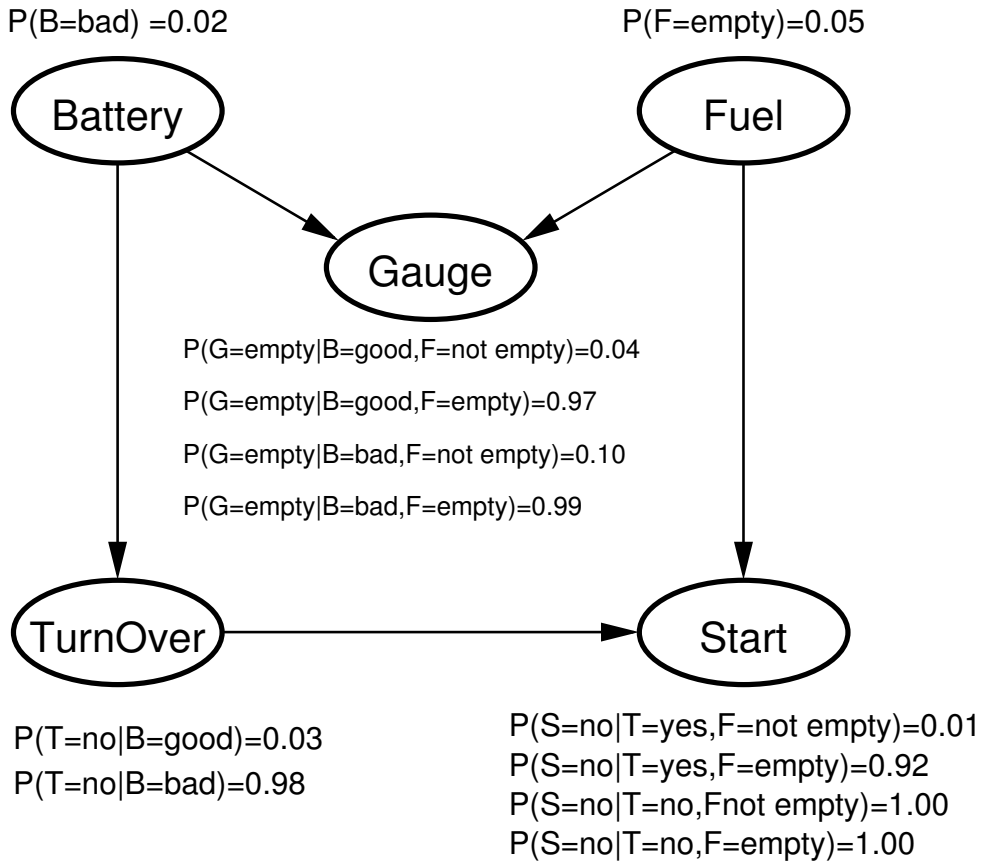
- Given three groups of random variables X, Y, Z

  $Indep(X, Y|Z) \iff$ every path from X to Y is "blocked by Z"

- A path is "blocked by Z" $\iff$
  – $\exists$ a node in Z that is head-to-tail w.r.t. the path, or
  – $\exists$ a node in Z that is tail-to-tail w.r.t. the path, or
  – $\exists$ another node A which is head-to-head w.r.t. the path and neither A nor any of its descendants are in Z

Example

(Heckermann 1995)



$Indep(T, F)$? $Indep(B, F|S)$? $Indep(B, S|T)$?

What can we do with Bayes nets?

- **Inference:** Given some pieces of information (prior, observed variabes) what is the implication (the implied information, the posterior) on a non-observed variable

- **Learning:**
  – Fully Bayesian Learning: Inference over parameters (e.g., β)
  – Maximum likelihood training: Optimizing parameters

- **Structure Learning** (Learning/Inferring the graph structure itself): Decide which model (which graph structure) fits the data best; thereby uncovering conditional independencies in the data.

Inference

- Inference: Given some pieces of information (prior, observed variabes) what is the implication (the implied information, the posterior) on a non-observed variable

- In a Bayes Nets: Assume there is three groups of RVs:
  – Z are observed random variables
  – X and Y are hidden random variables
  – We want to do inference about X, not Y

  Given some observed variables Z, compute the **posterior marginal** $P(X \mid Z)$ for some hidden variable X.

$$P(X \mid Z) = \frac{P(X, Z)}{P(Z)} = \frac{1}{P(Z)} \sum_{Y} P(X, Y, Z)$$

  where Y are all hidden random variables except for X

Example: Holmes & Watson

- Mr. Holmes lives in Los Angeles. One morning when Holmes leaves his house, he realizes that his grass is wet. Is it due to rain, or has he forgotten to turn off his sprinkler?

  – Calculate $P(R|H)$, $P(S|H)$ and compare these values to the prior probabilities.
  – Calculate $P(R, S|H)$.
  Note: R and S are marginally independent, but conditionally dependent

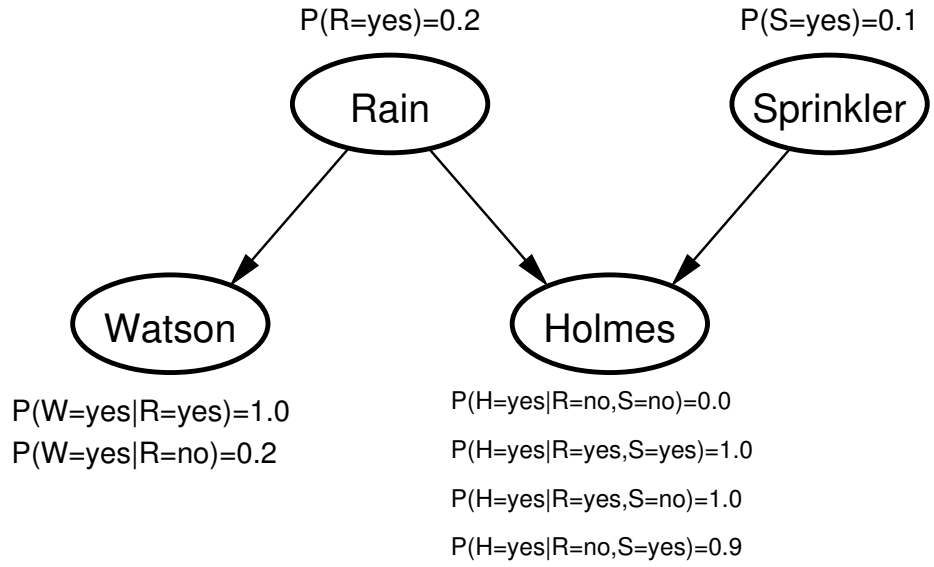
- Holmes checks Watson's grass, and finds it is also wet.

  – Calculate $P(R|H, W)$, $P(S|H, W)$
  – This effect is called explaining away

  JavaBayes: run it from the html page

  `http://www.cs.cmu.edu/~javabayes/Home/applet.html`

Example: Holmes & Watson



$$P(H, W, S, R) = P(H|S, R)\ P(W|R)\ P(S)\ P(R)$$

$$P(R|H) = \sum_{W,S} \frac{P(R, W, S, H)}{P(H)} = \frac{1}{P(H)} \sum_{W,S} P(H|S, R)\ P(W|R)\ P(S)\ P(R)$$

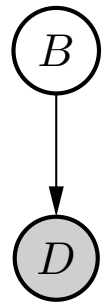
$$= \frac{1}{P(H)} \sum_{S} P(H|S, R)\ P(S)\ P(R)$$

$$P(R{=}1 \,|\, H{=}1) = \frac{1}{P(H{=}1)}(1.0 \cdot 0.2 \cdot 0.1 + 1.0 \cdot 0.2 \cdot 0.9) = \frac{1}{P(H{=}1)} 0.2$$

$$P(R{=}0 \,|\, H{=}1) = \frac{1}{P(H{=}1)}(0.9 \cdot 0.8 \cdot 0.1 + 0.0 \cdot 0.8 \cdot 0.9) = \frac{1}{P(H{=}1)} 0.072$$

- These types of calculations can be automated
  $\rightarrow$ Variable Elimination Algorithm

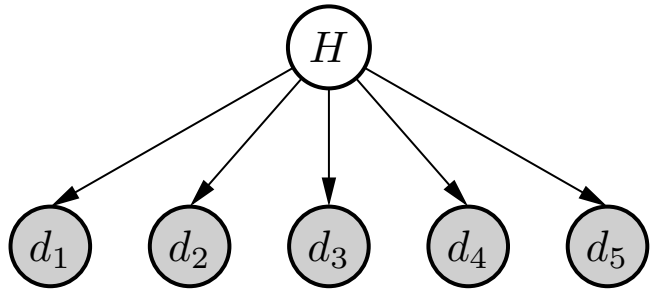
Bavarian dialect example



- Two binary random variables(RVs): B (bavarian) and D (dialect)
- Given:
  $P(D, B) = P(D \mid B)\ P(B)$
  $P(D{=}1 \mid B{=}1) = 0.4$, $P(D{=}1 \mid B{=}0) = 0.01$, $P(B{=}1) = 0.15$

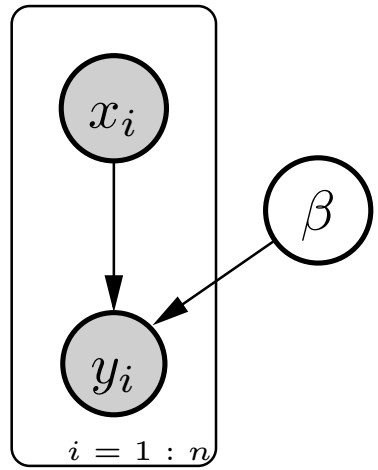
- **Notation:** Grey shading usually indicates “observed”

Coin flipping example



- One binary RV H (hypothesis), 5 RVs for the coin tosses $d_1, .., d_5$
- Given:
  $P(D, H) = \prod_i P(d_i \mid H)\ P(H)$
  $P(H{=}1) = \frac{999}{1000}$, $P(d_i{=}\texttt{H} \mid H{=}1) = \frac{1}{2}$, $P(d_i{=}\texttt{H} \mid H{=}2) = 1$

Ridge regression



- One multi-variate RV β, $2n$ RVs $x_{1:n}$, $y_{1:n}$ (observed data)
- Given:
  $P(D, \beta) = \prod_i \Big[P(y_i \,|\, x_i, \beta)\; P(x_i)\Big]\; P(\beta)$
  $P(\beta) = \mathcal{N}(\beta \,|\, 0, \frac{\sigma^2}{\lambda})$, $P(y_i \,|\, x_i, \beta) = \mathcal{N}(y_i \,|\, x_i^\top \beta, \sigma^2)$

- **Plate notation:** Plates (boxes with index ranges) mean "copy n-times"