Machine Learning I
Bayesian Networks

Marc Toussaint, Tobias Lang, K.-R. Müller
TU Berlin

Why Bayesian networks?

- Probability theory as calculus for uncertainty, information, couplings and evidence
- Bayesian networks are a generic tool for representing probability distributions and inference with coupled random variables.

$\Rightarrow$ Information Processing can be viewed as inference or message passing in Bayesian networks.

- Many concrete algorithms can be derived/explained in terms of Bayesian networks (or other graphical models) (e.g., Hidden Markov Models, Gaussian Processes, Conditional Random Fields...)
- Applications: spam filtering, structured output prediction, natural language understanding, signal processing, planning, reinforcement learning...

Recap: Probability Theory

- a random variable X assigns probabilities $P(X=x) \in \mathbb{R}$ to values $x \in \text{dom}(X)$
- probability distribution $\leftrightarrow$ table (vector) of probabilities for each value (normalization: $\sum_X P(X) = 1$)

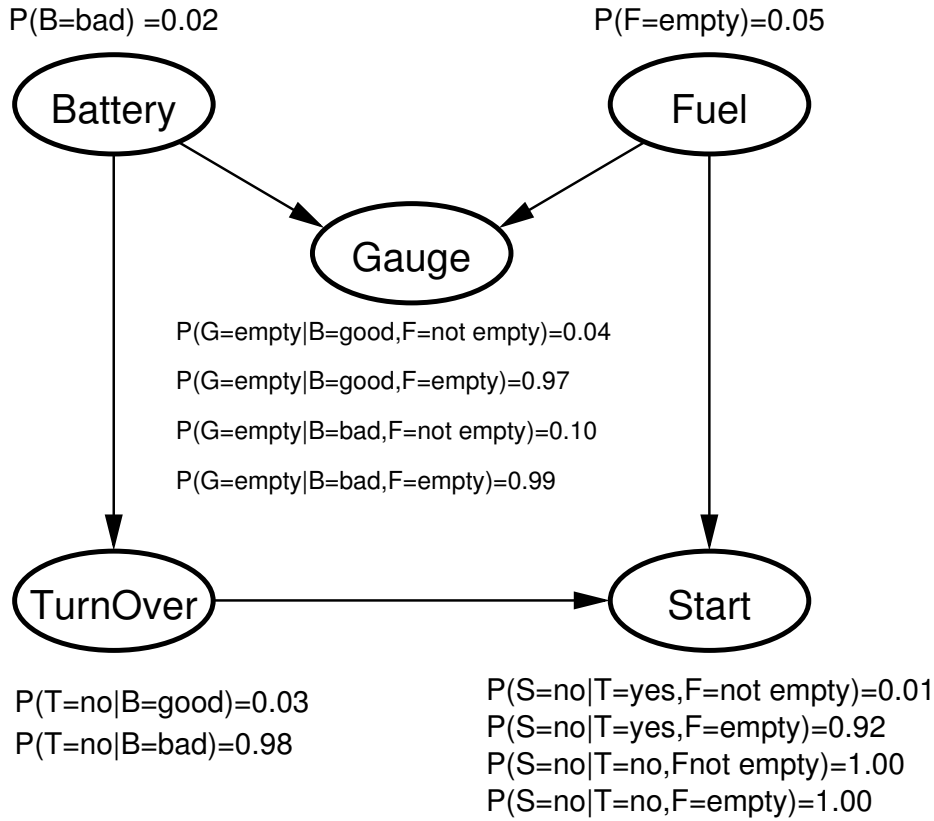
- joint distribution $P(X,Y)$ $\leftrightarrow$ table (matrix) of probabilities
- definition: marginal $P(X) = \sum_Y P(X,Y)$ (summing along columns/rows)

- definition: conditional $P(X|Y) = \frac{P(X,Y)}{P(Y)}$ (normalizing each column); implication: $P(X,Y) = P(X|Y)\ P(Y) = P(Y|X)\ P(X)$
- Chain rule: $P(X_1, .., X_n) = \prod_{i=1}^{n} P(X_i|X_1, .., X_{i\text{-}1})$
- Bayes rule:

$$P(X|Y) = \frac{P(Y|X)}{P(Y)} P(X)\ , \quad \text{posterior} = \frac{\text{likelihood}}{\text{evidence}}\ \text{prior}$$

Bayesian Network example

(Heckermann 1995)



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

as compared to the general chain rule:

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

Bayesian Network example - 2

The Bayesian Network is a graphical notation that says that the joint can be written as:

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

- table sizes: LHS = 2^5 RHS = $2^3 + 2^2 + 2^3 + 2 + 2$
- what is the probability of:
  $P(B = good, T = no, G = empty, F = notempty, S = no)$?

Bayesian Network example - 3

The Bayesian Network is a graphical notation that says that the joint can be written as:

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

- table sizes: LHS = 2^5 RHS = $2^3 + 2^2 + 2^3 + 2 + 2$
- what is the probability of:
  $P(B = good, T = no, G = empty, F = notempty, S = no)$?

$$\begin{aligned}
&= P(S = no \,|\, T = no, F = not\ empty) \cdot P(T = no \,|\, B = good) \\
&\quad \cdot P(G = empty | F = not\ empty, B = good) \\
&\quad \cdot P(F = not\ empty) \cdot P(B = good) \\
&= 1.0 \cdot 0.03 \cdot 0.04 \cdot 0.98 \cdot 0.95 \\
&\quad \approx 0.0011
\end{aligned}$$

Independence

- definition: X is independent of Y iff:

$$P(X|Y) = P(X)$$

for all possible values $x \in \text{dom}(X)$ and $y \in \text{dom}(Y)$

- in terms of the joint: X independent of Y iff:

$$P(X, Y) = P(X)\ P(Y)$$

- X independent of $Y \iff Y$ independent of X

- a set of variables $X_1, .., X_n$ is independent iff

$$P(X_1, .., X_n) = \prod_{i=1}^{n} P(X_i)$$

Independence

Example:

	Toothache = true	Toothache = false
Cavity = true	0.04	0.06
Cavity = false	0.01	0.89

- is T independent from C ?

Independence - 2

Example:

	Toothache = true	Toothache = false
Cavity = true	0.04	0.06
Cavity = false	0.01	0.89

- is T independent from C ?

  $P(C = t) = 0.04 + 0.06 = 0.1$

  $P(T = t) = 0.04 + 0.01 = 0.05$

  $P(C = t, T = t) = 0.04$

  $P(C = t) \cdot P(T = t) = 0.1 * 0.05 = 0.005$

  $\rightarrow P(C = t, T = t) \neq P(C = t) \cdot P(T = t)$

  $\rightarrow T$ and C are dependent.

Conditional Independence

- definition: X is conditionally independent of Y given Z iff

$$P(X|Y,Z) = P(X|Z)$$

for all $x \in \text{dom}(X)$, $y \in \text{dom}(Y)$, $z \in \text{dom}(Z)$
Intuition: Given Z, additional knowledge about Y does not change the probability of X.

- in terms of the joint:

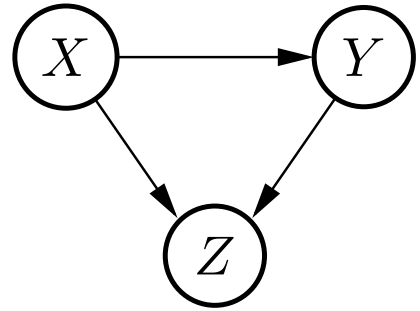
$$P(X,Y,Z) = P(X,Y|Z)\ P(Z) = P(X|Z)\ P(Y|Z)\ P(Z)$$

Nota bene: conditional independence $\not\rightarrow$ independence
independence $\not\rightarrow$ conditional independence

Bayesian Networks

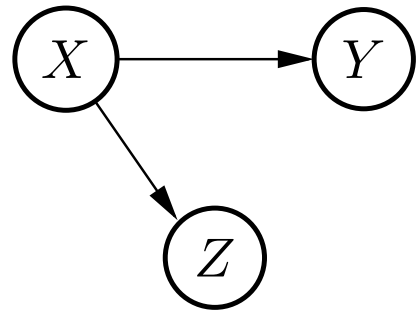
1st model:

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



2nd model:

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



- Bayesian network is a graphical notation of (in)dependence

Inference

- Assume a Bayesian net with random variables $Y_{1:k}, E_{1:m}, H_{1:m}$

- definition: *inference* is the problem to compute

$$P(Y_{1:k} \mid E_{1:m}) = \frac{P(Y_{1:k}, E_{1:m})}{P(E_{1:m})} \propto \sum_{H_{1:n}} P(Y_{1:k}, E_{1:m}, H_{1:n})$$

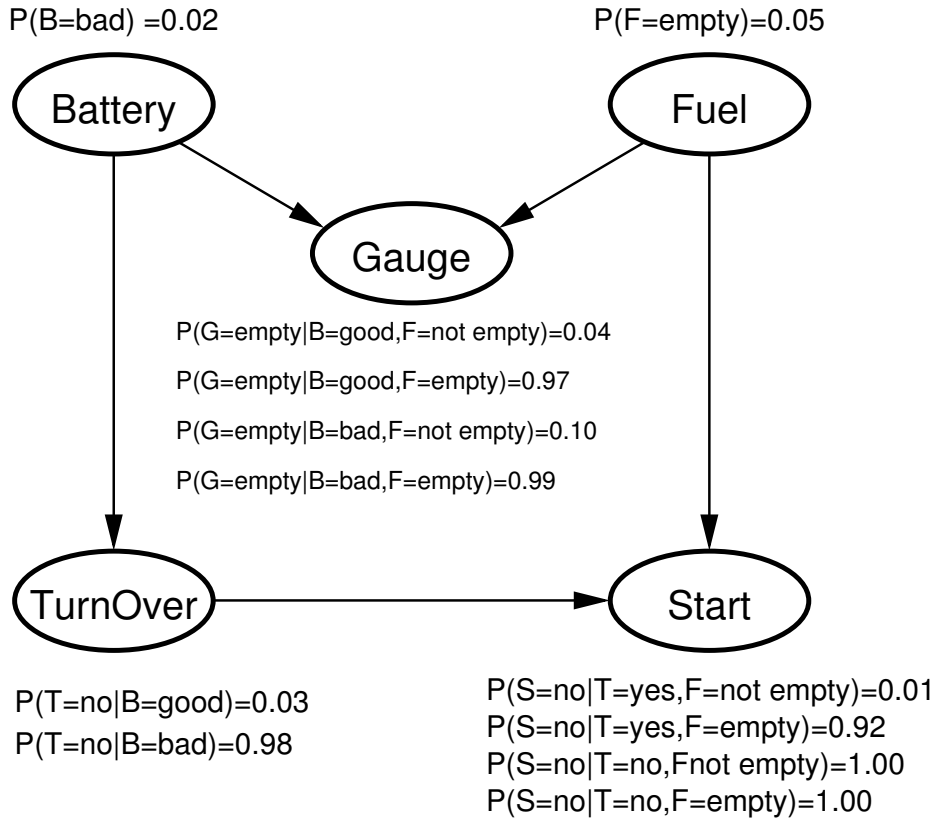
$Y_{1:n}$ = variables of interest for which to compute a-posteriori

$E_{1:m}$ = evidence variables (observed values)

$H_{1:n}$ = hidden variables (don't care about)

Inference in Bayes Nets

(Heckermann 1995)



- What is $P(F = not\ empty \mid S = no)$?

  Variable of interest F, observed variable S, hidden variables G, B, T

- What is $P(B = bad, F = empty \mid T = no)$?

Inference in Bayes Nets - 2

- 2 fundamental operations for information processing

  1) multiplication of probability distributions to fuse (independent) information

  2) summation (elimination) of variables to compute marginals
  elimination $\equiv$ "summing out variables"
  (eliminate Y from $P(X, Y|Z)$ means to compute $P(X|Z) = \sum_Y P(X, Y|Z)$)

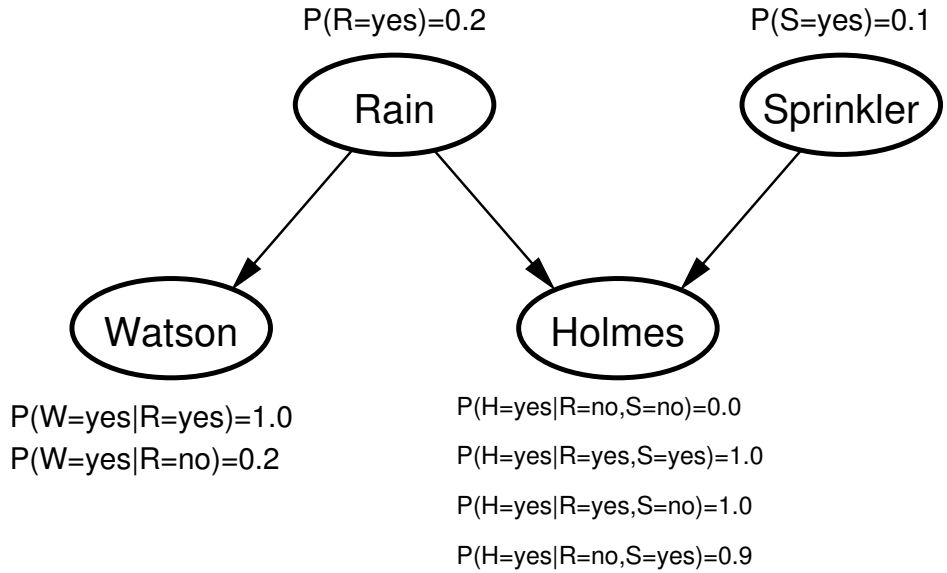
- Inference methods exploit (conditional) independencies.

Inference in Bayes Nets - 3

- If Bayesian net is a *tree*:
  – to compute a *single* marginal (single inference query like $P(B|S = no)$): *Variable Elimination*
  – to compute *all* marginals (e.g., compute $P(B|S = no)$ and $P(F|S = no)$ and $P(G|S = no)$ and $P(T|S = no)$): *Message Passing (Belief Propagation)*
  inference in time linear in the number of nodes; messages are passed up and down the tree; all the necessary computations can be carried out locally

- If Bayesian net is *not* a tree:
  – exact method: clustering (grouping) of nodes to yield a tree of cliques (junction tree)
  – approximate methods: sampling, loopy belief propagation, varational methods

Extended Example



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

Extended Example - 2

- Mr. Holmes lives in Berlin. 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 ($P(H) = 0.272$, $P(R|H) \approx 0.735$, $P(S|H) \approx 0.338$)
- Calculate $P(R, S|H)$. R and S are marginally independent, but conditionally dependent. (For instance, in case Holmes' grass is wet, the probability for rain sinks when we additionally know that the sprinkler was on.) ($P(R, S|H) = 0.0735 > 0.02 = P(R, S)$)
- Holmes checks Watson's grass, and finds it is also wet. Calculate $P(R|H, W)$, $P(S|H, W)$.
  $\rightarrow$ As Watson's grass is also wet, Holmes' grass is most probably wet due to the common cause rain – thus, the posterior for the sprinkler decreases. This effect is called *explaining away*.

Learning in Bayesian Networks

- General problem: learning probability models
  - learning *conditional probability tables (parameters)*; easier

    Especially easy if all variables are observed, otherwise can use EM

    Maximize likelihood of data (see next slide)
  - learning *structure*; harder

    Can try out a number of different structures, but there can be a huge number of structures to search through

Maximum Likelihood

- Very simple Bayesian net with two binary variables: $X \to Y$
- Given n input-output pairs $(x_1, y_1), \ldots, (x_n, y_n)$
- Goal: estimate parameters θ (specifying $P(X)$ and $P(Y \mid X)$)
- For points generated independently and identically distributed (iid), the likelihood of the data is

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p(x_i | y_i; \theta).$$

- Maximum likelihood chooses θ to maximize $\mathcal{L}$ (and thus L)
- Here (all data observed): ML estimates are relative frequencies
- Generalizes to larger Bayesian nets
- In case of hidden variables: Expectation-Maximization (EM)

Conclusions

- Bayesian nets model probability distributions using a graphical notation to specify (in)dependencies among the random variables.
- Complex generative models
- Inference methods to query a-posteriori probabilities given observed variables

- Applications: spam filtering, structured output prediction, natural language understanding, signal processing, planning, reinforcement learning...

- Later lecture: more on learning (maximum likelihood)
- Later lecture: hidden variables