Artificial general intelligence

Uncertainty in artificial intelligence: from rational beliefs to inference

Antal Péter

ComBine Lab

Artificial Intelligence group

Department of Measurement and Information Systems

Agenda

- Interpretations of probability
- Axiomatic derivations of probability theory
- Axiomatic derivations of Bayesian (decision) theory
- Axioms of structural properties of joint probability distributions

Interpretations of probability

Uncertainty

information

belief

compression

probability

complexity

learning

entropy

randomness

determinism

openness

indeterminism

emergence

Interpretations of probability

- Sources of uncertainty
  - inherent uncertainty in the physical process;
  - inherent uncertainty at macroscopic level;
  - ignorance;
  - practical omissions;
- Interpretations of probabilities:
  - combinatoric;
  - physical propensities;
  - frequentist;
  - personal/subjectivist;
  - instrumentalist;

$$\lim_{N\to\infty}\frac{N_A}{N}=\lim_{N\to\infty}\hat{p}_N(A)=p(A)\,?\,p(A|\xi)$$

Physicalist view of probabilitites

- .A.Einstein: „God does not play dice..”
  https://arxiv.org/ftp/arxiv/papers/1301/1301.1656.pdf
- Einstein-Podolski-Rosen paradox / Bell Test
- S. Hawking: „Does god play dice?”
  http://www.hawking.org.uk/does-god-play-dice.html
- The BIG Bell Test (Nov30, 2016)
  - http://bist.eu/100000-people-participated-big-bell-test-unique-worldwide-quantum-physics-experiment/

A chronology of uncertain inference

- [1713] Ars Conjectandi (The Art of Conjecture), Jacob Bernoulli
  - **Subjectivist interpretation** of probabilities
- [1718] The Doctrine of Chances, Abraham de Moivre
  - the first textbook on probability theory
  - **Forward predictions**
    - „given a specified number of white and black balls in an urn, what is the probability of drawing a black ball?”
- [1764, posthumous] Essay Towards Solving a Problem in the Doctrine of Chances, Thomas Bayes
  - **Backward questions**: „given that one or more balls has been drawn, what can be said about the number of white and black balls in the urn”
- [1812], Théorie analytique des probabilités, Pierre-Simon Laplace
  - General Bayes rule
- [1933]: A. Kolmogorov: *Foundations of the Theory of Probability*

Axiomatic derivations of probability theory

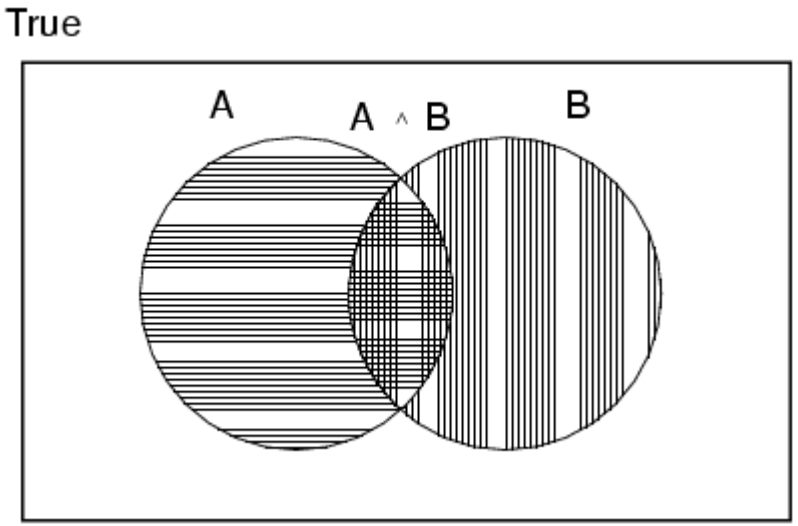
Other AI approaches to uncertain reasoning

- Certainty factors
- Fuzzy logic
  - https://en.wikipedia.org/wiki/Fuzzy_logic
- Dempster-Schafer theory (imprecise probability theories)
  - https://en.wikipedia.org/wiki/Dempster%E2%80%93Shafer_theory

Cheeseman, P.C., 1985, August. In Defense of Probability. In *IJCAI* (Vol. 2, pp. 1002-1009).
Heckerman, D.E. and Shortliffe, E.H., 1992. From certainty factors to belief networks. *Artificial Intelligence in Medicine*, *4*(1), pp.35-52.

Axioms of probability

- For any propositions *A, B*
-
  - $0 \leq P(A) \leq 1$
  - $P(true) = 1$ and $P(false) = 0$
  - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



Normative derivations of probability theory

- Cox's theorem
  - https://en.wikipedia.org/wiki/Cox%27s_theorem
  - Goal:
    - Divisibility and comparability – The plausibility of a proposition is a real number and is dependent on information we have related to the proposition.
    - Common sense – Plausibilities should vary sensibly with the assessment of plausibilities in the model.
    - Consistency – If the plausibility of a proposition can be derived in many ways, all the results must be equal.
  - Associativity: $AB|X=g(A|X,B|AX)$
  - Monotonicity (isomorphism with multiplication): $w(AB|X)=w(A|X)w(B|AX)$
  - Boundary conditions: $w(Certainty|X)=1$, $w(Impossibility|X)=0$
  - Negation function f: $f(not\ X)=1-f(X)$ (+isomorphism with multiplication)

Cox-Jaynes axioms

https://en.wikipedia.org/wiki/Cox%27s_theorem

➔ *finite additivity*

Sigma-additivity: Kolmogorov's measure-theoretic formulation

Basic concepts of probability theory

- Joint distribution
- Conditional probability
- Bayes' rule
- Chain rule
- Marginalization
- General inference
- Independence
  - Conditional independence
  - Contextual independence

Conditional probability

- Definition of conditional probability:
- $P(a \mid b) = P(a \wedge b) / P(b)$ if $P(b) > 0$
- Product rule gives an alternative formulation:
- $P(a \wedge b) = P(a \mid b)\, P(b) = P(b \mid a)\, P(a)$
- Chain rule is derived by successive application of product rule:
- $\mathbf{P}(X_1, \ldots, X_n) = \mathbf{P}(X_1,\ldots,X_{n-1})\, \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1})$
  $= \mathbf{P}(X_1,\ldots,X_{n-2})\, \mathbf{P}(X_{n-1} \mid X_1,\ldots,X_{n-2})\, \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1})$
  $= \ldots$
  $= \pi_{i=1}^{n}\, \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})$

Bayes rule

An algebraic triviality

$$p(X \mid Y) = \frac{p(Y \mid X)\,p(X)}{p(Y)} = \frac{p(Y \mid X)\,p(X)}{\sum_X p(Y \mid X)\,p(X)}$$

A scientific research paradigm

$$p(Model \mid Data) \propto p(Data \mid Model)\,p(Model)$$

A practical method for inverting causal knowledge to diagnostic tool.

$$p(Cause \mid Effect) \propto p(Effect \mid Cause) \times p(Cause)$$

Conditional independence

$I_P(X;Y|Z)$ or $(X \perp\!\!\!\perp Y|Z)_P$ denotes that X is independent of Y given Z defined as follows

for all x,y and z with P(z)>0: $P(x;y|z)=P(x|z)\,P(y|z)$

(Almost) alternatively, $I_P(X;Y|Z)$ iff

$P(X|Z,Y)= P(X|Z)$ for all z,y with P(z,y)>0.

Other notations: $D_P(X;Y|Z)$ =def= $\neg I_P(X;Y|Z)$

Contextual independence: for not all z.

Direct dependence: $D_P(X;Y|V/\{X,Y\})$

Axiomatic derivations of „Bayesianism”

From ”rational" preferences to probabilities: ”as if" I.

1. Definition. *A decision problem is defined by the elements* $\mathcal{E}, \mathcal{C}, \mathcal{A}, \leq$, *where:*

(i) $\mathcal{E}$ *is an algebra of events,* E_j;

(ii) $\mathcal{C}$ *is a set of possible consequences,* c_j;

(iii) $\mathcal{A}$ *is a set of possible acts, which are mapping of partitions of the events to consequences;*

(iv) $\leq$ *is a binary preference relation between some of the elements of* $\mathcal{A}$.

With further ”rational" assumptions on comparability, transitivity, consistency and quantification the following suggestive result can be derived.

1. Proposition. *Given an uncertainty relation* $\leq$, *there exists a unique real number P(E) for each event E (defined as the probability of E) that they are compatible with* $\leq$ *(i.e.* $E \leq F; iff; P(E) \leq P(F)$*) and they form a finitely additive probability measure.*

Consequently, $P(A|\xi)$ denotes the subjective/personal beliefs in a given space-time-information context ξ vs. the ”frequentist" interpretation that $P(A) \triangleq \lim_{N\to\infty} N_A/N$.

Bernardo, J.M. and Smith, A.F., 2009. *Bayesian theory* (Vol. 405). John Wiley & Sons.

From preferences to utilities: ”as if" II.

The parallel result for the existence and uniqueness of utilities (or losses) is stated only for decision problems with bounded consequences.

2. Proposition. *For any decision problem* $\mathcal{E}, \mathcal{C}, \mathcal{A}, \leq$ *with bounded consequences* $c_* < c^*$,

(i) for all c, $u(c|c_*, c^*)$ *exists and unique;*

(ii) the value of $u(c|c_*, c^*)$ *is unaffected by the assumed occurrence of an event G;*

(iii) $0 = u(c_*|c_*, c^*) \leq u(c|c_*, c^*) \leq u(c^*|c_*, c^*) = 1.$

(iv) the so-called maximum expected utility principle is satisfied, i.e.

$$a_1 \leq_G a_2 \Leftrightarrow \sum_j u(c_{a_1(E_j)}) P(E_j|G) \leq \sum_j u(c_{a_2(E_j)}) P(E_j|G) \qquad (1)$$

Bernardo, J.M. and Smith, A.F., 2009. *Bayesian theory* (Vol. 405). John Wiley & Sons.

From exchangeability to parameters and ”i.i.d." : ”as if" III.

3. Proposition. *If $x_1, x_2, \ldots$ is an infinitely exchangeable sequence of 0-1 random quantities with probability measure P, that is for any n and permutation $\pi(1), \ldots, \pi(n)$ the joint mass function of P $p(x_1, \ldots, x_n) = p(x_{\pi(1)}, \ldots, x_{\pi(n)})$, there exists a distribution function $\mathcal{Q}$ such that $p(x_1, \ldots, x_n)$ has the form*

$$p(x_1, \ldots, x_n) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i} (1 - \theta^{1-x_i}) d\mathcal{Q}(\theta),$$

where,

$$\mathcal{Q}(\theta) = \lim_{n \to \infty} P[y_n/n \leq \theta],$$

with $y_n = x_1 + \cdots + x_n$, and $\theta = \lim_{n \to \infty} y_n/n$.

Bayesian inference using Beta distribution

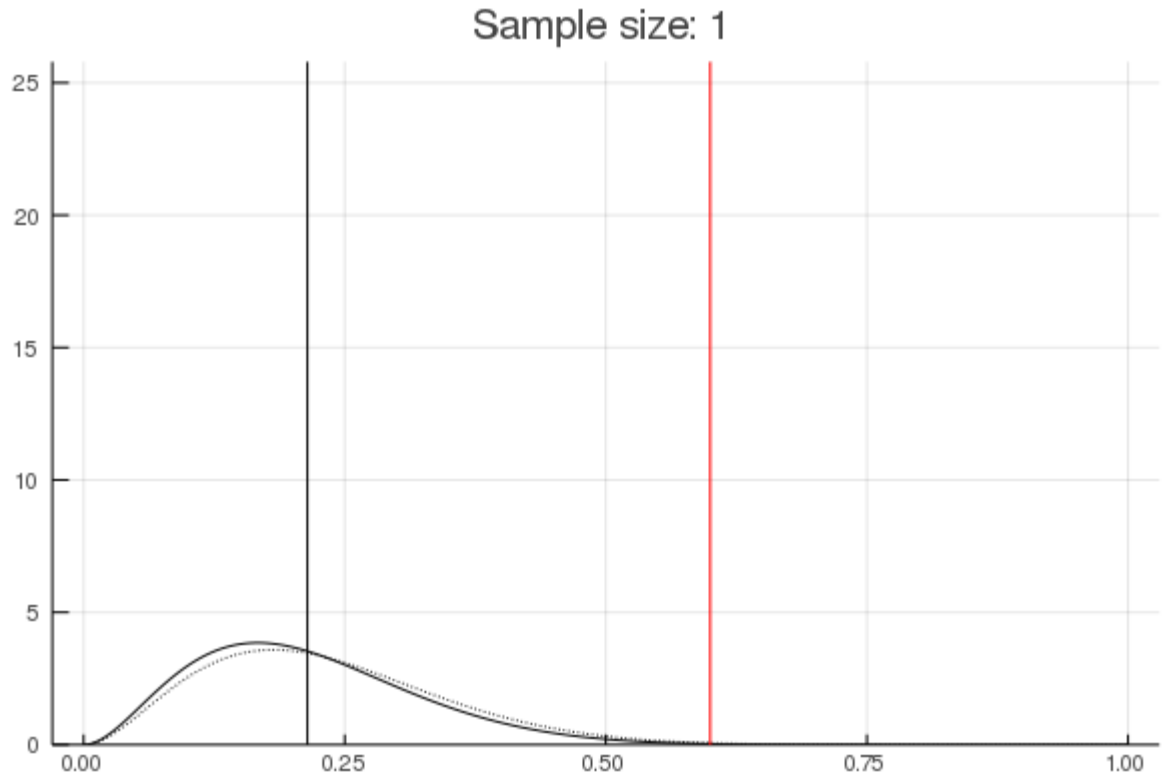
1. Example. *Assume that x denotes the sum of 1s of n independent and identically distributed (i.i.d.) Bernoulli trials, that is we assume a binomial sampling distribution. If the prior is specified using a Beta distribution, the posterior remains a Beta distribution with updated parameters.*

$$p(x|\theta) = Bin(x|n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \tag{13}$$

$$p(\theta) = Beta(\alpha,\beta) = c\theta^{\alpha-1}(1-\theta)^{\beta-1} \; where \; c = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{14}$$

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)} = c'\theta^{\alpha-1+x}(1-\theta)^{\beta-1+n-x} = Beta(\alpha+x,\beta+n-x)$$

Example



Prior: Beta(3,10)
fix: 0.6

The Bayesian statistical framework

1. Specify a joint distribution $p(x, \theta)$ over the observable quantity x and parameter θ having equal status by specifying $p(\theta)$ the prior distribution or prior, the $p(x|\theta)$ is the sampling distribution that also defines the likelihood and the likelihood function $\mathcal{L}(\theta; x)$ (the discrete model parameter is denoted with $\mathcal{M}_k$).
2. Perform a prior predictive inference

$$p(x) = \int p(x|\theta)p(\theta)d\theta \; or \; p(x) = \sum_k p(\mathcal{M}_k) \int p(x|\mathcal{M}_k) \tag{2}$$

or a posterior predictive inference after observing the data set D as

$$p(x|D) = \int p(x|\theta)p(\theta|D)d\theta \; or \; p(x|D) = \sum_k p(x|\mathcal{M}_k)p(\mathcal{M}_k|D) \tag{3}$$

3. Perform a parametric inference by the Bayes rule

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta} \propto p(x|\theta)p(\theta) \; or \; p(\mathcal{M}_k|x) = \propto p(x|\mathcal{M}_k)p(\mathcal{M}_k) \tag{4}$$

Hierarchical Bayesian modelling

A frequently occuring form in practice, that the specification usually achieved in the structured specification of the relevant model structures $\mathcal{S}_k$ or $\mathcal{M}_k$ and parameters θ_k. Correspondingly the a priori belief $p(\theta_k, \mathcal{M}_k)$ in a given model with structure k and parameters θ_k^i is expressed as a product

$$p(\theta_k, \mathcal{M}_k) = p(\mathcal{M}_k)p(\theta_k^i|\mathcal{M}_k)$$

➔ No theoretical need for „typed” probability of probabilities!

Predictive inference (”Bayesian inference")

The specification of the a priori beliefs over relevant models allows us to perform (prior) predictive inferences over the observable quantity x

$$p(x) = \sum p(\mathcal{M}_k) \int p(x|\theta_k)p(\theta_k|\mathcal{M}_k)d\theta_k$$

The operation of integration and/or summation over models and/or their parameterization implements marginalization and termed in this context as Bayesian model averaging . We can write the posterior predictive distribution conditioned on the data set D as

$$p(x|D) = \sum p(\mathcal{M}_k|D) \int p(x|\theta_k)p(\theta_k|D,\mathcal{M}_k)d\theta_k \approx p(x|D,\mathcal{M}_k^{MAP})$$

in which $\mathcal{M}_k^{MAP} = argmax_k p(\mathcal{M}_k|D)$ is called the maximum a posteriori (MAP) model.

Parametric inference (”Bayesian learning")

In the discrete case the posterior of the model $p(\mathcal{M}_k|D)$ is given by

$$p(\mathcal{M}_k|D) = \frac{p(D|\mathcal{M}_k)p(\mathcal{M}_k)}{p(D)} \tag{9}$$

where the marginal model likelihood or evidence for $\mathcal{M}_k$ is

$$p(D|\mathcal{M}_k) = \int p(D|\theta_k, \mathcal{M}_k)p(\theta_k|\mathcal{M}_k)d\theta_k \tag{10}$$

and the marginal data likelihood

$$p(D) = \sum_k p(D|\mathcal{M}_k)p(\mathcal{M}_k) \tag{11}$$

The *Bayes factor* shows the change of the ratio of prior belief to the ratio of the posteriors, i.e. the ratios of marginal likelihoods of models $\mathcal{M}_i$ and $calM_j$

2. Definition.

$$Bayesfactor(\mathcal{M}_i, \mathcal{M}_j) = \frac{p(D|\mathcal{M}_i)}{p(D|\mathcal{M}_j)} = \frac{p(\mathcal{M}_j)}{p(\mathcal{M}_i)} \frac{p(\mathcal{M}_i|D)}{p(\mathcal{M}_j|D)} \tag{12}$$

An example for full Bayesian inference

Principles for induction

- Epicurus' (342? B.C. - 270 B.C.) principle of multiple explanations which states that one should *keep all hypotheses that are consistent with the data*.
- The principle of Occam's razor (1285 - 1349, sometimes spelt Ockham). Occam's razor states that when inferring causes *entities should not be multiplied beyond necessity*. This is widely understood to mean: Among all hypotheses consistent with the observations, choose the simplest. In terms of a prior distribution over hypotheses, this is the same as giving simpler hypotheses higher a priori probability, and more complex ones lower probability.

Full Bayesian learning

View learning as Bayesian updating of a probability distribution over the hypothesis space

H is the hypothesis variable, values $h_1, h_2, \ldots$, prior $\mathbf{P}(H)$ jth observation d_j gives the outcome of random variable D_j training data $\mathbf{d} = d_1, \ldots, d_N$

Given the data so far, each hypothesis has a posterior probability:

$$P(h_i|\mathbf{d}) = \alpha P(\mathbf{d}|h_i)P(h_i)$$

where $P(\mathbf{d}|h_i)$ is called the likelihood

Predictions use a likelihood-weighted average over the hypotheses:

$$\mathbf{P}(X|\mathbf{d}) = \Sigma_i \, \mathbf{P}(X|\mathbf{d}, h_i)P(h_i|\mathbf{d}) = \Sigma_i \, \mathbf{P}(X|h_i)P(h_i|\mathbf{d})$$

No need to pick one best-guess hypothesis!

Bayesian model averaging

View learning as Bayesian updating of a probability distribution over the hypothesis space

H is the hypothesis variable, values $h_1, h_2, \ldots$, prior $\mathbf{P}(H)$

jth observation d_j gives the outcome of random variable D_j
training data $\mathbf{d} = d_1, \ldots, d_N$

Given the data so far, each hypothesis has a posterior probability:

$$P(h_i|\mathbf{d}) = \alpha P(\mathbf{d}|h_i)P(h_i)$$

where $P(\mathbf{d}|h_i)$ is called the likelihood

Predictions use a likelihood-weighted average over the hypotheses:

$$\mathbf{P}(X|\mathbf{d}) = \sum_i \mathbf{P}(X|\mathbf{d}, h_i)P(h_i|\mathbf{d}) = \sum_i \mathbf{P}(X|h_i)P(h_i|\mathbf{d})$$

No need to pick one best-guess hypothesis!

Russel&Norvig: Artificial intelligence, ch.20

Bayesian Model Averaging example

Suppose there are five kinds of bags of candies:

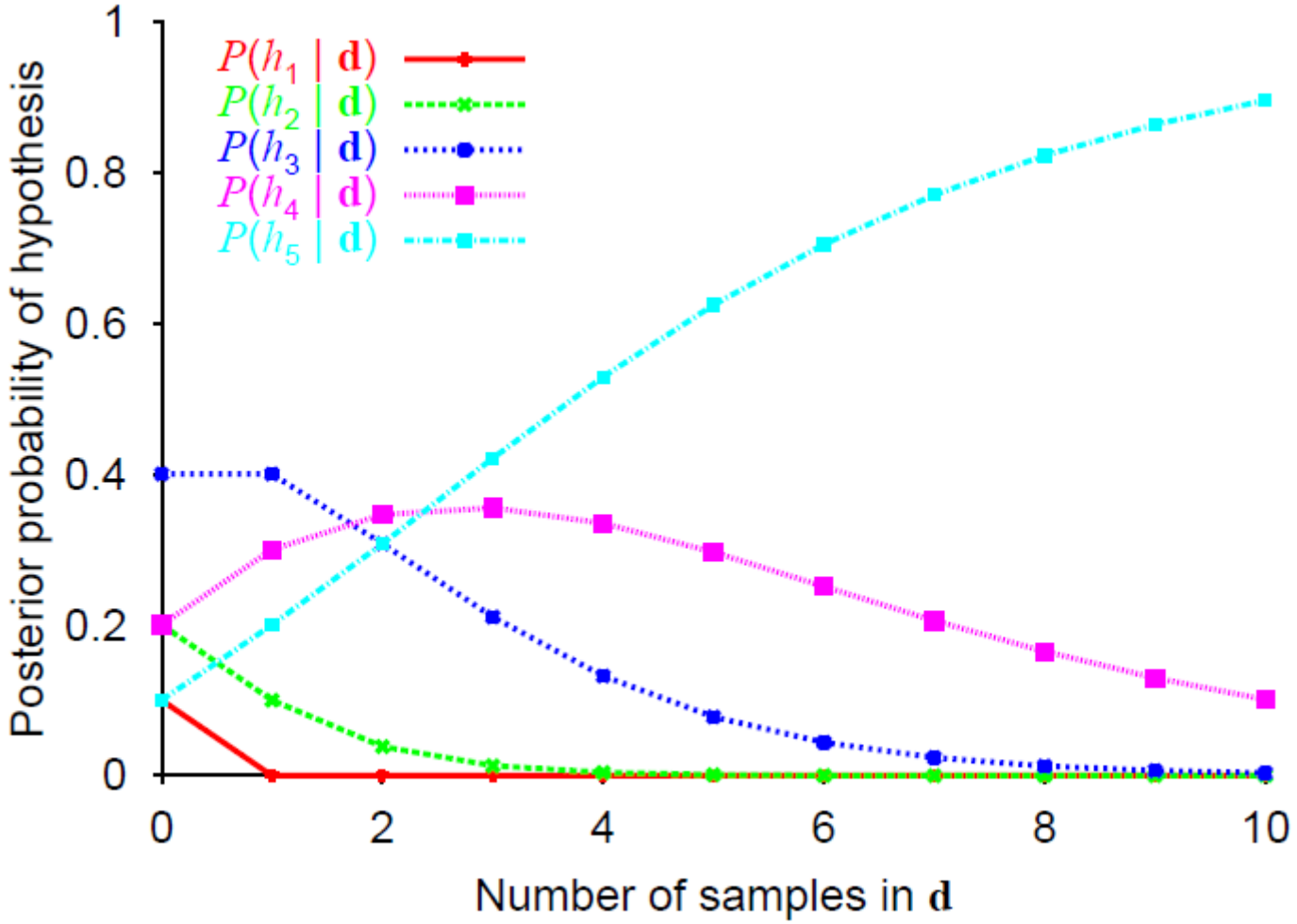
10% are h_1: 100% cherry candies
20% are h_2: 75% cherry candies + 25% lime candies
40% are h_3: 50% cherry candies + 50% lime candies
20% are h_4: 25% cherry candies + 75% lime candies
10% are h_5: 100% lime candies

Then we observe candies drawn from some bag: 

What kind of bag is it? What flavour will the next candy be?

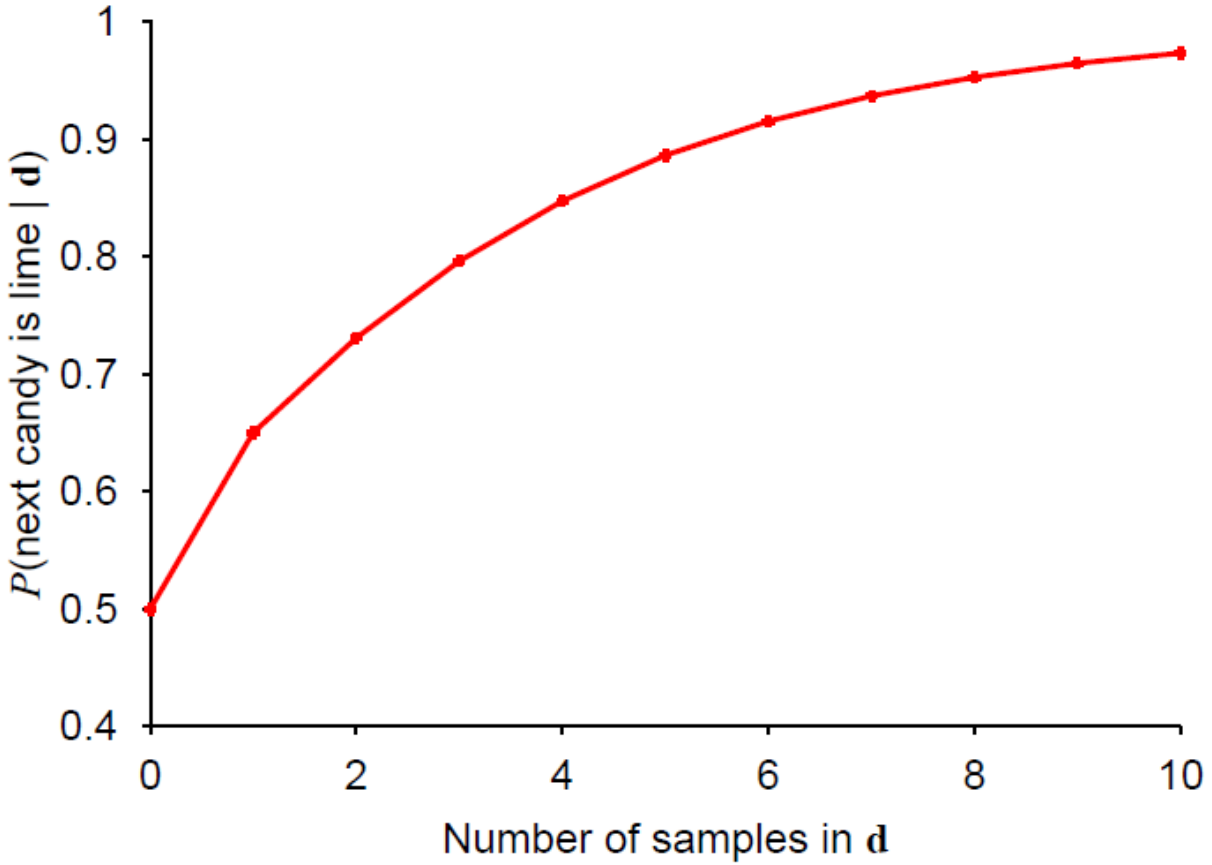
Russel&Norvig: Artificial intelligence

Learning rate for models



Russel&Norvig: Artificial intelligence

Learning rate for model predictions



Russel&Norvig: Artificial intelligence

MAP approximation

Summing over the hypothesis space is often intractable
(e.g., 18,446,744,073,709,551,616 Boolean functions of 6 attributes)

Maximum a posteriori (MAP) learning: choose h_{MAP} maximizing $P(h_i|\mathbf{d})$

I.e., maximize $P(\mathbf{d}|h_i)P(h_i)$ or $\log P(\mathbf{d}|h_i) + \log P(h_i)$

Log terms can be viewed as (negative of)
bits to encode data given hypothesis + bits to encode hypothesis
This is the basic idea of minimum description length (MDL) learning

For deterministic hypotheses, $P(\mathbf{d}|h_i)$ is 1 if consistent, 0 otherwise
$\Rightarrow$ MAP = simplest consistent hypothesis (cf. science)

ML approximation

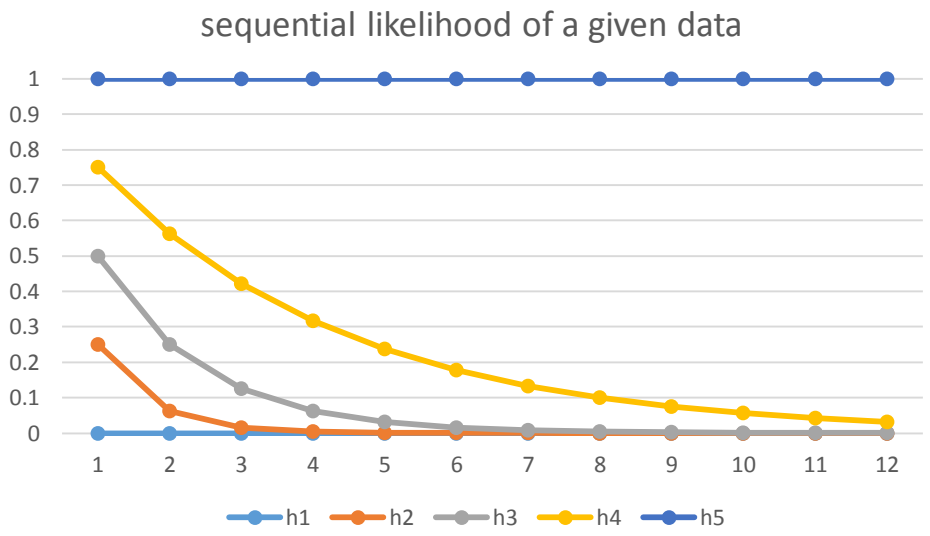
For large data sets, prior becomes irrelevant

Maximum likelihood (ML) learning: choose h_{ML} maximizing $P(\mathbf{d}|h_i)$

I.e., simply get the best fit to the data; identical to MAP for uniform prior
(which is reasonable if all hypotheses are of the same complexity)

ML is the "standard" (non-Bayesian) statistical learning method

Maximum likelood model selection



Further examples for full Bayesian inference

Universal theory of induction

•Universal distributions

$$m(x) := \sum_{p\,:\,U(p)=x} 2^{-\ell(p)}, \qquad -\log m(x) = K(x) + O(1).$$

$$M(x) := \sum_{p\,:\,U(p)=x*} 2^{-\ell(p)}, \qquad -\log M(x) = K(x) - O(\log \ell(x))$$

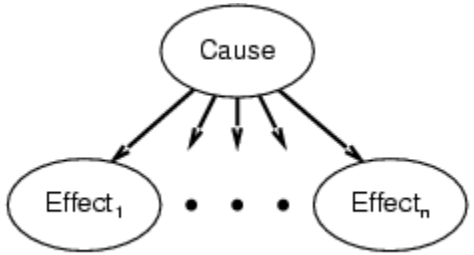
If the infinite binary sequences are distributed according to a computable measure μ , then the predictive distribution $M(x_{n+1}|x_{1:n})=M(x_{1:n+1})/M(x_{1:n})$ converges rapidly to $\mu(x_{n+1}|x_{1:n})=\mu(x_{1:n+1})/\mu(x_{1:n})$ with probability 1. Hence, M predicts almost as well as does the true distribution μ.

M. Li and P. M B. Vitanyi. *An introduction to Kolmogorov complexity and its applications*. Springer
Solomonoff, R.J., 1964. A formal theory of inductive inference. Part I. *Information and control*, *7*(1), pp.1-22.
Chaitin, G., 1974. Information-theoretic computation complexity. *IEEE Transactions on Information Theory*, *20*(1), pp.10-15.

Universal inference, universal priors, universal AI

- Solomonoff, R.J., 1960, November. A preliminary report on a general theory of inductive inference. United States Air Force, Office of Scientific Research.
- **Solomonoff, R.J., 1964. A formal theory of inductive inference. Part I. *Information and control*, *7*(1), pp.1-22.**
- **Chaitin, G., 1974. Information-theoretic computation complexity. *IEEE Transactions on Information Theory*, *20*(1), pp.10-15.**
- Chaitin, G.J., 1977. Algorithmic information theory. *IBM journal of research and development*, *21*(4), pp.350-359.
- Solomonoff, R., 1978. Complexity-based induction systems: comparisons and convergence theorems. *IEEE transactions on Information Theory*, *24*(4), pp.422-432.
- Cover, T.M., 1985. Kolmogorov complexity, data compression, and inference. In *The Impact of Processing Techniques on Communications* (pp. 23-33). Springer, Dordrecht.
- **Li, M. and Vitányi, P., 1993. *An introduction to Kolmogorov complexity and its applications*. New York: Springer.**
- Solomonoff, R.J., 1997. The discovery of algorithmic probability. *Journal of Computer and System Sciences*, *55*(1), pp.73-88.
- **Dawid, A.P. and Vovk, V.G., 1999. Prequential probability: Principles and properties. *Bernoulli*, *5*(1), pp.125-162.**
- Chater, N. and Vitányi, P.M., 2003. The generalized universal law of generalization. *Journal of Mathematical Psychology*, *47*(3), pp.346-369.
- Hutter, M., 2007. On universal prediction and Bayesian confirmation. *Theoretical Computer Science*, *384*(1), pp.33-48.
- Solomonoff, R.J., 2008. Three kinds of probabilistic induction: Universal distributions and convergence theorems. *The Computer Journal*, *51*(5), pp.566-570.
- Gács, P. and Vitányi, P.M., 2011. Raymond J. Solomonoff 1926–2009. *IEEE Information Theory Society Newsletter*, *61*(1), pp.11-16.
- **Hutter, M., 2004. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media.**
- Rathmanner, S. and Hutter, M., 2011. A philosophical treatise of universal induction. *Entropy*, *13*(6), pp.1076-1136.

Naive Bayesian network



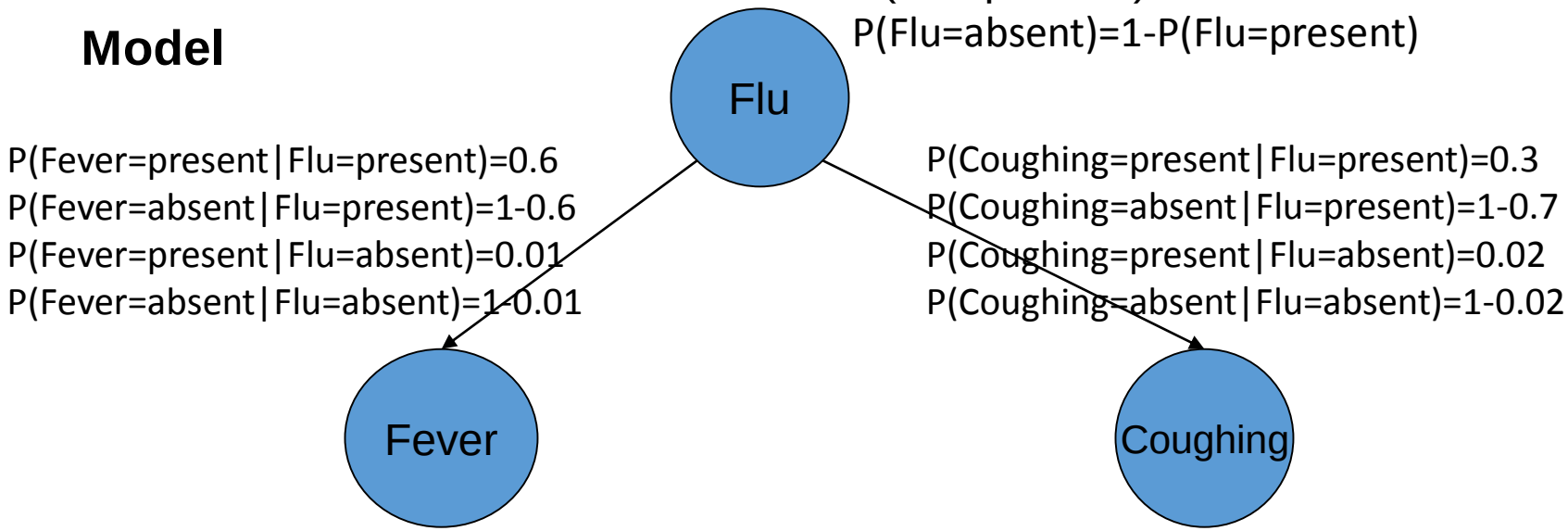
Assumptions:
1, Two types of nodes: a cause and effects.
2, Effects are conditionally independent of each other given their cause.

Variables (nodes)

Flu: present/absent
FeverAbove38C: present/absent
Coughing: present/absent

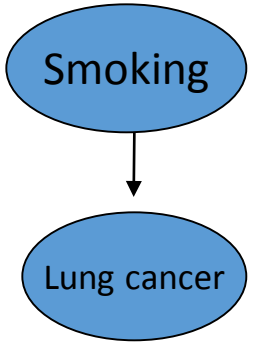
Model

P(Flu=present)=0.001
P(Flu=absent)=1-P(Flu=present)



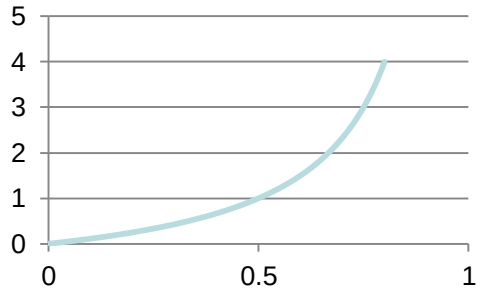
Domingos, Pedro, and Michael Pazzani. "On the optimality of the simple Bayesian classifier under zero-one loss." *Machine learning* 29.2-3 (1997): 103-130.
Friedman, Jerome H. "On bias, variance, 0/1—loss, and the curse-of-dimensionality." *Data mining and knowledge discovery* 1.1 (1997): 55-77.
Hand, David J., and Keming Yu. "Idiot's Bayes—not so stupid after all?." *International statistical review* 69.3 (2001): 385-398.

Conditional probabilities, odds, odds ratios



	¬S	S	
¬LC	P(¬S, ¬LC)	P(S, ¬LC)	P(¬LC)
LC	P(¬S, LC)	P(S, LC)	P(LC)
	P(¬S)	P(S)	

Probability:
P(LC)
Conditional probabilities (e.g., probability of LC given S)**:**
P(LC| ¬S)= ??? P(LC| S)= ??? P(LC)
Odds:
[0,1] →[0,∞]: Odds(p)=p/(1-p)
O(LC| ¬S)= ??? O(LC| S)
Odds Ratio (OR) Independent of prevalence!
OR(LC,S)=O(LC| S)/O(LC| ¬S)



Naive Bayesian network (NBN)

Decomposition of the joint:

$P(Y,X_1,..,X_n) = P(Y)\prod_i P(X_i,|Y, X_1,..,X_{i-1})$ //by the chain rule

$= P(Y)\prod_i P(X_i,|Y)$ // by the N-BN assumption

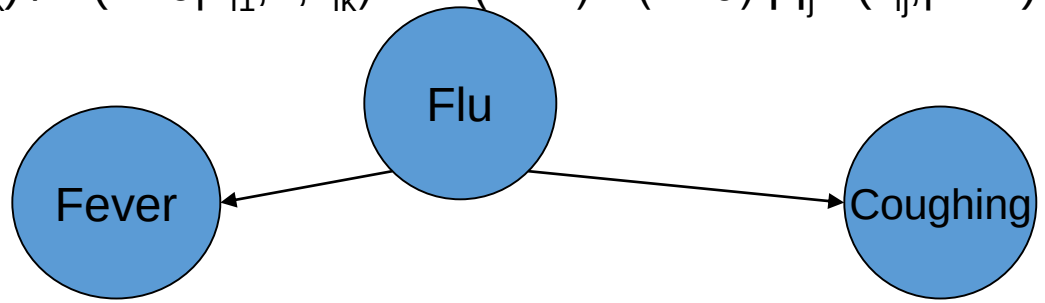
2n+1 parameteres!

Diagnostic inference:

$P(Y|x_{i1},..,x_{ik}) = P(Y)\prod_j P(x_{ij},|Y) / P(x_{i1},..,x_{ik})$

If Y is binary, then the odds

$P(Y=1|x_{i1},..,x_{ik}) / P(Y=0|x_{i1},..,x_{ik}) = P(Y=1)/P(Y=0) \prod_j P(x_{ij},|Y=1) / P(x_{ij},|Y=0)$



$$p(Flu = present \mid Fever = absent, Coughing = present)$$

$$\propto p(Flu = present)\, p(Fever = absent \mid Flu = present)\, p(Coughing = present \mid Flu = present)$$

Full Bayesian naive-BN

- Structure prior: p(G)
  - Specify priors for edges in G
  - Penalize deviation from a prior structure G_0
- Parameter prior: $p(\Theta|G)$
  - θ denotes the complete parametrization for G
  - Specify $p(\Theta|G)$ independently for each variable?
  - Specify $p(\Theta|G)$ using a „convenient” (~conjugate) prior?
- Inference
  - Tractable?

Full Bayesian inference with N-BNs using complete data

- Integration over parameters?
  - Analytical solution under parameter independence!
  - Hyperparameter update.
- Bayesian model averaging over exponential number of structures?
  - Analytical solution!
  - Existence of equivalent „super”-parametrization!!

Dash, Denver, and Gregory F. Cooper. "Exact model averaging with naive Bayesian classifiers." *ICML*. 2002.

Extensions of N-BNs

- Tree-augmented BNs
- BN-augmented BNs
- Hierarchical BNs
- Multiple parents
  - Explaining away
- „Context-sensitive” N-BNs

Langseth, Helge, and Thomas D. Nielsen. "Classification using hierarchical naive bayes models." *Machine learning* 63.2 (2006): 135-159.

On the subjectivity of priors and losses

Optimal decision/estimation:

$$x^* = argmin_{\hat{x}} \int L(x, \hat{x}) p(x|D) dx$$

Axioms of structural properties of probability distributions

The independence model of a distribution

The independence map (model) M of a distribution P is the set of the valid independence triplets:

$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots, I_{P,K}(X_K;Y_K|Z_K)\}$

If P(X,Y,Z) is a Markov chain, then

$M_P=\{D(X;Y), D(Y;Z), I(X;Z|Y)\}$

Normally/almost always: D(X;Z)

Exceptionally: I(X;Z)

The semi-graphoid axioms

1. Symmetry: The observational probabilistic conditional independence is symmetric.

$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \; iff \; I_p(\boldsymbol{Y};\boldsymbol{X}|\boldsymbol{Z})$$

2. Decomposition: Any part of an irrelevant information is irrelevant.

$$I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z}) \Rightarrow \; I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \; and \; I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z})$$

3. Weak union: Irrelevant information remains irrelevant after learning (other) irrelevant information.

$$I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z}) \Rightarrow \; I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}\cup\boldsymbol{W})$$

4. Contraction: Irrelevant information remains irrelevant after forgetting (other) irrelevant information.

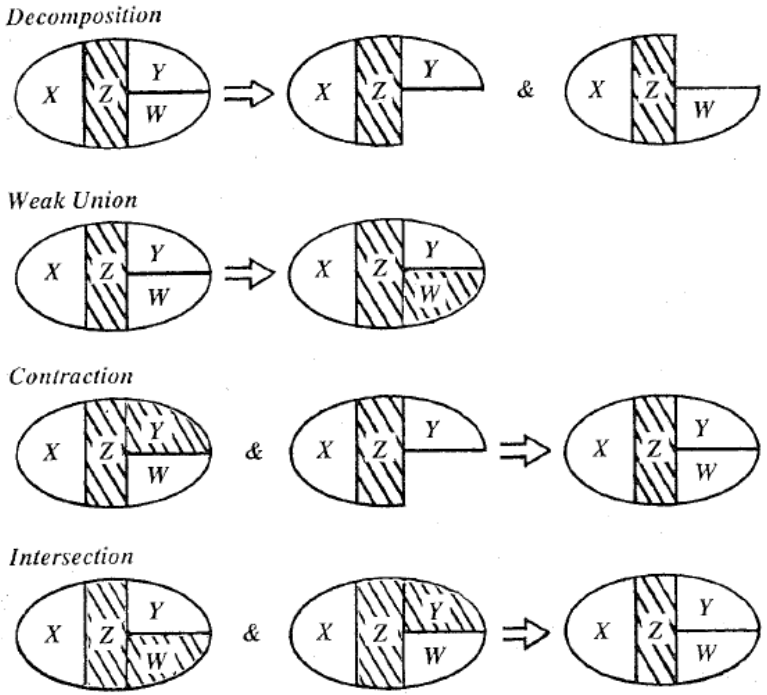
$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \; and \; I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z}\cup\boldsymbol{Y}) \Rightarrow \; I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z})$$

Graphoids

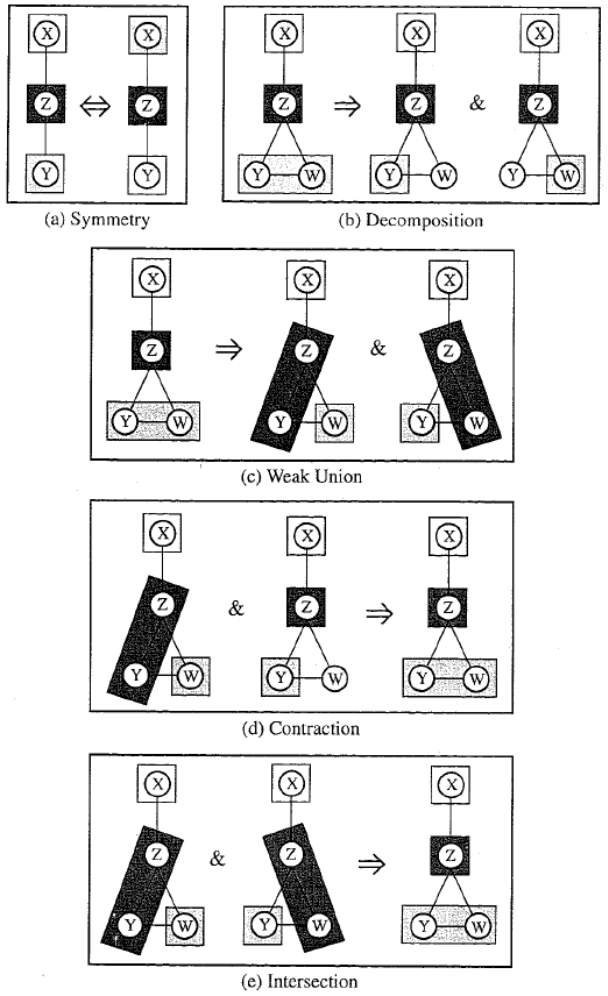
Graphoids: Semi-graphoids+Intersection (holds only in strictly positive distribution)

Intersection: Symmetric irrelevance implies joint irrelevance if there are no dependencies.

$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}\cup\boldsymbol{W}) \text{ and } I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z}\cup\boldsymbol{Y}) \Rightarrow I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z})$$



J.Pearl: Probabilistic Reasoning in intelligent systems, 1998



Summary

- Probability theory is a unified theory for uncertainty
- Normative derivation of uncertain reasoning
  - Bayes' rule as automation of rational inference with uncertainty
- Axiomatic derivations of „Bayesianism"
  - „As if" representation of beliefs over models
- Axioms of structural properties of probability distribution
  - Independence models
- Next: human biases, causality, the value alignment problem