Bayesian Data Analysis

PHY / CSI / INF 451 / 551
Prof. Kevin H. Knuth
University at Albany, Albany NY, USA

Fall 2020

Introduction to Probability Theory

Probability as an Extension of Boolean Logic

Table 1. Boolean Algebra.

Unary Operation	
Complementation	NOT $\equiv$ $\neg$
Complementation 1	$\mathbf{A} \wedge \neg\mathbf{A} = \perp$
Complementation 2	$\mathbf{A} \vee \neg\mathbf{A} = \top$
Idempotency	$\mathbf{A} = \neg\neg\mathbf{A}$

Binary Operations	
Disjunction	OR $\equiv$ $\vee$
Conjunction	AND $\equiv$ $\wedge$
Idempotency	$\mathbf{A} \vee \mathbf{A} = \mathbf{A}$
	$\mathbf{A} \wedge \mathbf{A} = \mathbf{A}$
Commutativity	$\mathbf{A} \vee \mathbf{B} = \mathbf{B} \vee \mathbf{A}$
	$\mathbf{A} \wedge \mathbf{B} = \mathbf{B} \wedge \mathbf{A}$
Associativity	$\mathbf{A} \vee (\mathbf{B} \vee \mathbf{C}) = (\mathbf{A} \vee \mathbf{B}) \vee \mathbf{C}$
	$\mathbf{A} \wedge (\mathbf{B} \wedge \mathbf{C}) = (\mathbf{A} \wedge \mathbf{B}) \wedge \mathbf{C}$
Absorption	$\mathbf{A} \vee (\mathbf{A} \wedge \mathbf{B}) = \mathbf{A} \wedge (\mathbf{A} \vee \mathbf{B}) = \mathbf{A}$
Distributivity	$\mathbf{A} \wedge (\mathbf{B} \vee \mathbf{C}) = (\mathbf{A} \wedge \mathbf{B}) \vee (\mathbf{A} \wedge \mathbf{C})$
	$\mathbf{A} \vee (\mathbf{B} \wedge \mathbf{C}) = (\mathbf{A} \vee \mathbf{B}) \wedge (\mathbf{A} \vee \mathbf{C})$
De Morgan 1	$\neg\mathbf{A} \wedge \neg\mathbf{B} = \neg(\mathbf{A} \vee \mathbf{B})$
De Morgan 2	$\neg\mathbf{A} \vee \neg\mathbf{B} = \neg(\mathbf{A} \wedge \mathbf{B})$

Consistency
$\mathbf{A} \rightarrow \mathbf{B} \quad \Leftrightarrow \quad \mathbf{A} \wedge \mathbf{B} = \mathbf{A} \quad \Leftrightarrow \quad \mathbf{A} \vee \mathbf{B} = \mathbf{B}$

Probability

Consider the logical statements

$R = \text{"}It\ is\ raining!\text{"}$

$\neg R = \text{"}It\ is\ not\ raining!\text{"}$

While it is clear that

$$R \rightarrow (R \lor \neg R)$$

However $(R \lor \neg R) \nrightarrow R$

But we might like to quantify the degree to which

$$(R \lor \neg R) \rightarrow R$$

We will define a function called probability, denoted $p(A\,|B)$, that quantifies the **degree to which the logical statement $\boldsymbol{B}$ implies the logical statement $\boldsymbol{A}$**.

Probability

The limits of probability are defined by Boolean logic and certainty.

If the logical statement B implies the logical statement A then $P(A \mid B) = 1$.

If the logical statements A and B are disjoint, such that $A \wedge B = \emptyset$, then $P(A \mid B) = 0$.

For non-disjoint A and B ($A \wedge B \neq \emptyset$), we have that $0 < P(A \mid B) \leq 1$.

A logical statement implies itself, so that $P(A \mid A) = 1$.

Since $A \rightarrow A \vee X$, we have that $P(A \vee X \mid \mathrm{A}) = 1$.

Of course, one can rescale the probability function.
The use of percentages is a common example.

The Sum and Product Rules

Associativity of Logical OR

Knuth, K.H. and Skilling, J., 2012. Foundations of inference. *Axioms*, *1*(1), pp.38-73.
Knuth, K.H., 2019. Lattices and their consistent quantification. *Annalen der Physik*, *531*(3), p.1700370.

For disjoint X, Y, and Z

$$(X \vee Y) \vee Z = X \vee (Y \vee Z)$$

implies that any measure[1] v then obeys

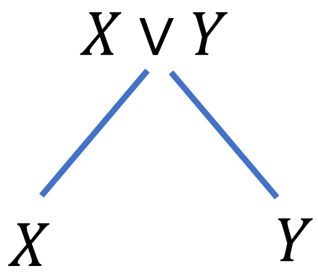
$$v(X \vee Y) = v(X) + v(Y)$$

1. v is a function that takes an element X to a real number

Sum Rule

Knuth, K.H., 2019. Lattices and their consistent quantification. *Annalen der Physik*, *531*(3), p.1700370.

For disjoint X and Y

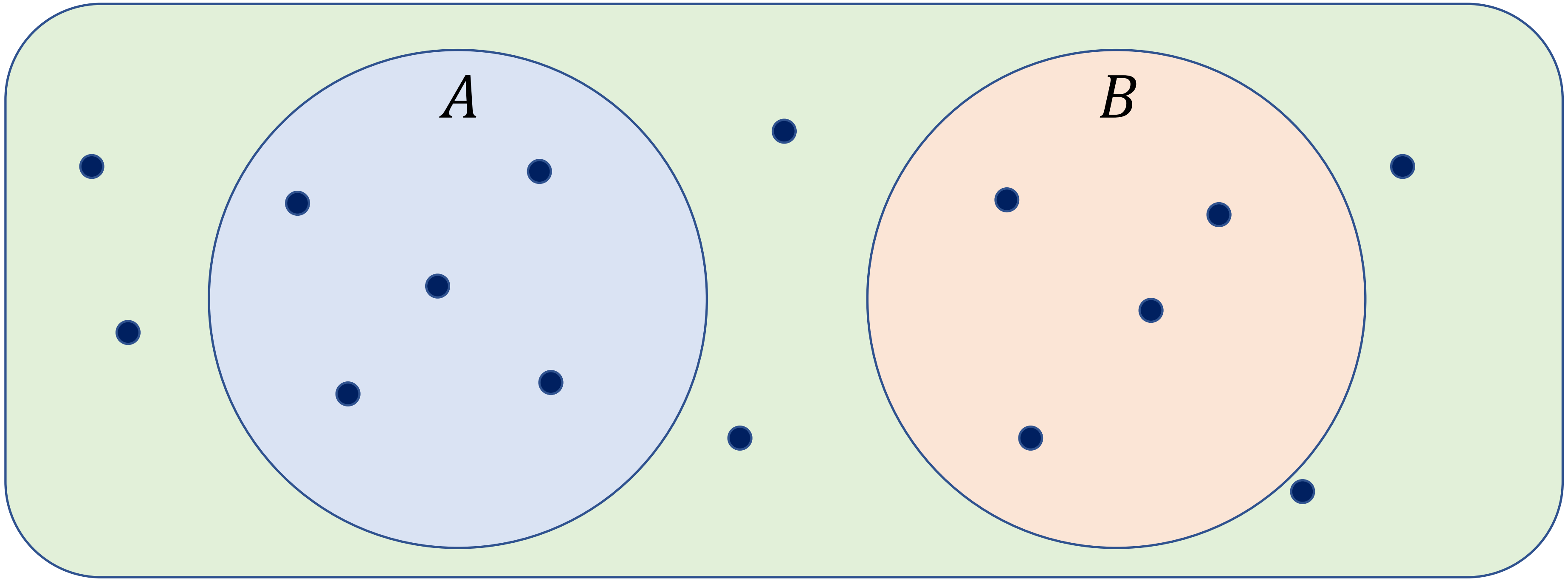


$$v(X \vee Y) = v(X) + v(Y)$$

Sum Rule

For disjoint A and B, we have that

$$p(A \vee B \mid I) = p(A \mid I) + p(B \mid I)$$



$p(A \mid I) = \frac{5}{15}$

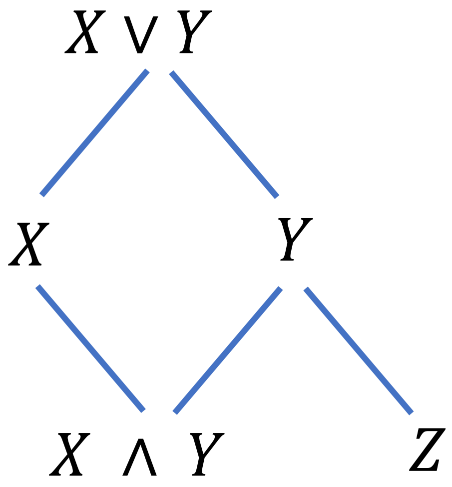
$p(B \mid I) = \frac{4}{15}$

$$p(A \vee B \mid I) = \frac{5}{15} + \frac{4}{15} = \frac{9}{15}$$

Sum Rule

Knuth, K.H., 2019. Lattices and their consistent quantification. *Annalen der Physik*, *531*(3), p.1700370.

In General



$$v(Y) = v(X \wedge Y) + v(Z)$$

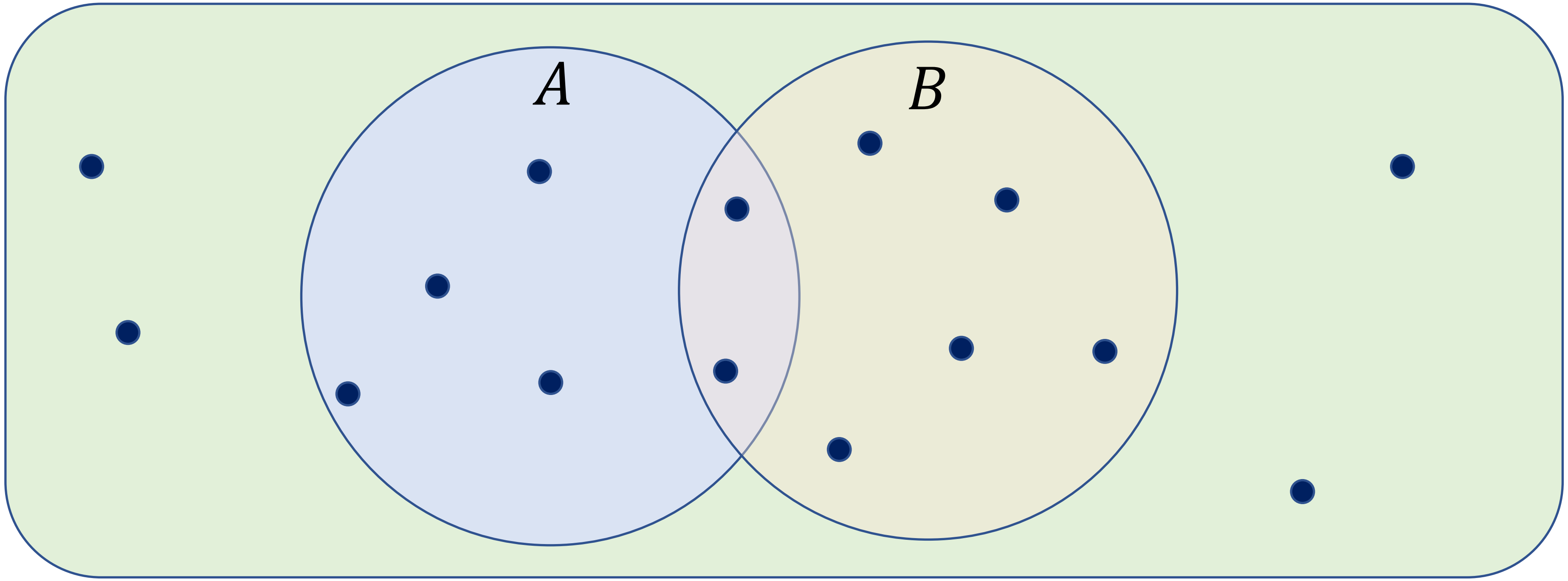
$$v(X \vee Y) = v(X) + v(Z)$$

$$v(X \vee Y) = v(X) + v(Y) - v(X \wedge Y)$$

Sum Rule

For general A and B, we have that

$$p(A \vee B | I) = p(A | I) + p(B | I) - p(A \wedge B | I)$$



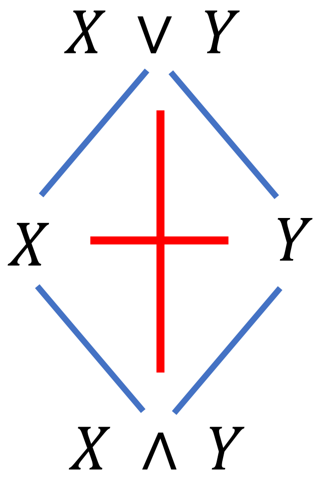
$p(A | I) = \frac{6}{15}$ $p(A \wedge B | I) = \frac{2}{15}$ $p(B | I) = \frac{7}{15}$

$$p(A \vee B | I) = \frac{6}{15} + \frac{7}{15} - \frac{2}{15} = \frac{11}{15}$$

Sum Rule

Knuth, K.H., 2019. Lattices and their consistent quantification. *Annalen der Physik*, *531*(3), p.1700370.

$$v(X \vee Y) = v(X) + v(Y) - v(X \wedge Y)$$



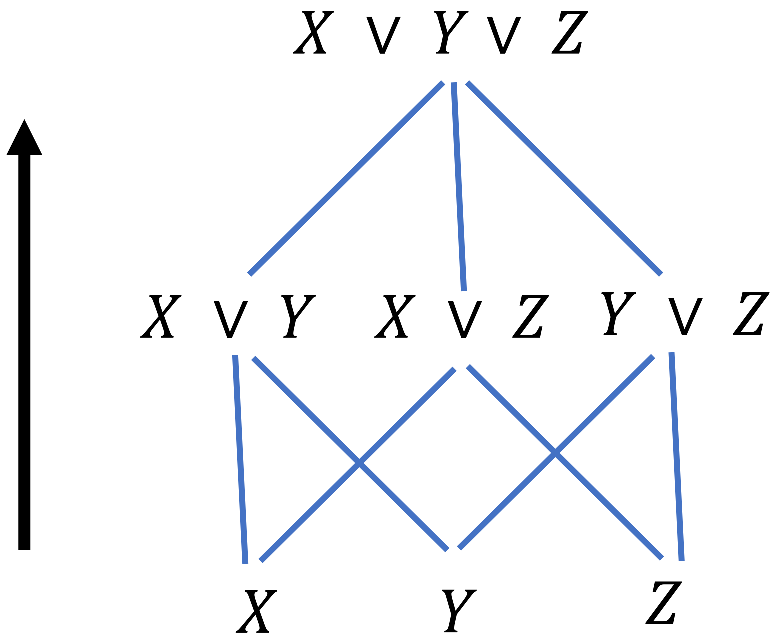
$$v(X \vee Y) + v(X \wedge Y) = v(X) + v(Y)$$

Context and Bi-Valuations

Knuth, K.H., 2009, December. Measuring on lattices. In AIP Conference Proceedings (Vol. 1193, No. 1, pp. 132-144). American Institute of Physics.

Implication is directed upward

(eg. X implies $X \lor Y$)



We would like to quantify the degree to which the statement $X \lor Y \lor Z$ implies the statement X

Define a function $w(X \mid X \lor Y \lor Z)$

The statement on the right of the solidus (|) is called the CONTEXT.

The function w is called a bi-valuation because it takes two statements to a real number.

Sum Rule

Knuth, K.H., 2009, December. Measuring on lattices. In AIP Conference Proceedings (Vol. 1193, No. 1, pp. 132-144). American Institute of Physics.

For a constant context, the Sum Rule holds for bi-valuations[2]

$$w(X \vee Y \mid Z) = w(X \mid Z) + w(Y \mid Z) - w(X \wedge Y \mid Z)$$

In the case of probability, this is

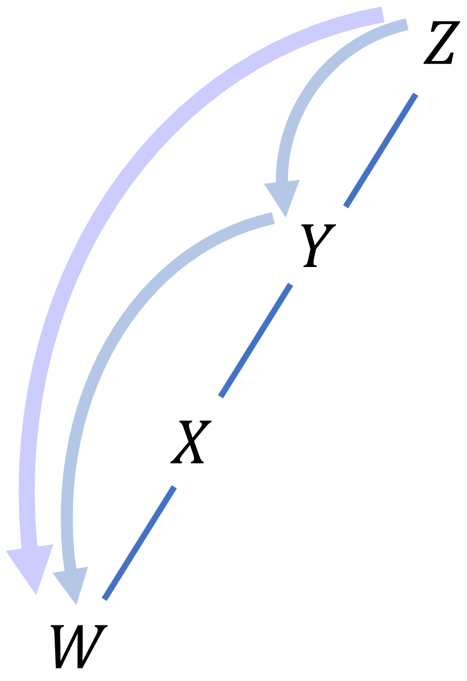
$$p(A \vee B \mid I) = p(A \mid I) + p(B \mid I) - p(A \wedge B \mid I)$$

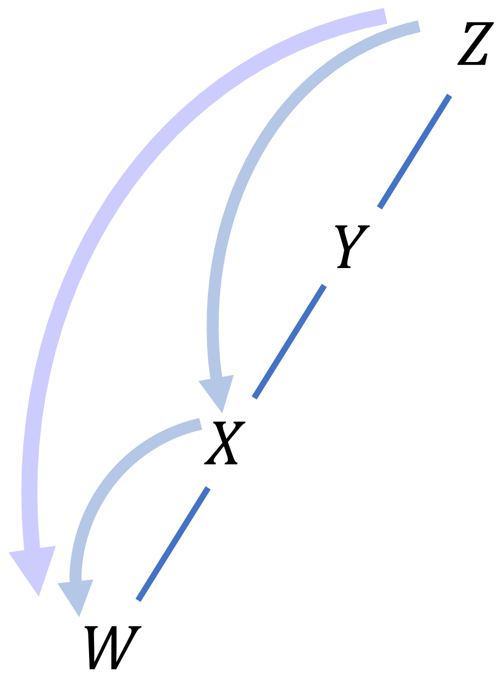
2. w is a function that takes two elements to a real number.
The second number (to the right of the solidus | is referred to as the *context*.

Associativity of Chaining of Context

Knuth, K.H. and Skilling, J., 2012. Foundations of inference. *Axioms*, *1*(1), pp.38-73.
Knuth, K.H., 2019. Lattices and their consistent quantification. *Annalen der Physik*, *531*(3), p.1700370.

associativity of changing context

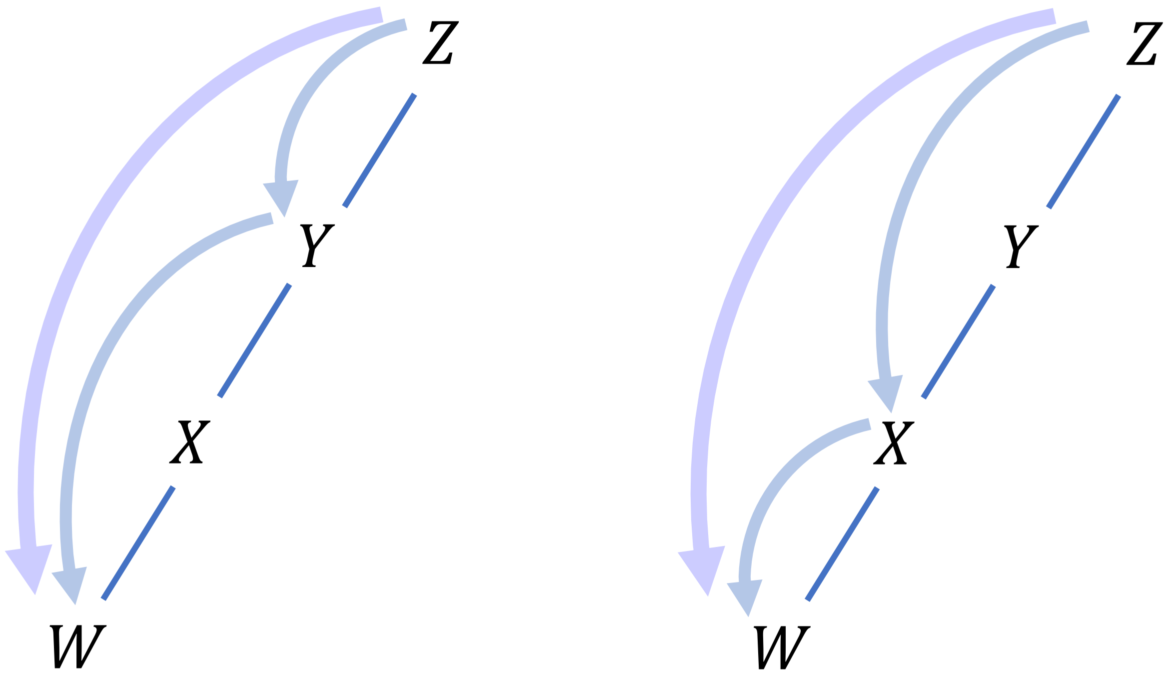




Chain Rule

Knuth, K.H. and Skilling, J., 2012. Foundations of inference. *Axioms*, *1*(1), pp.38-73.
Knuth, K.H., 2019. Lattices and their consistent quantification. *Annalen der Physik*, *531*(3), p.1700370.

associativity of changing context

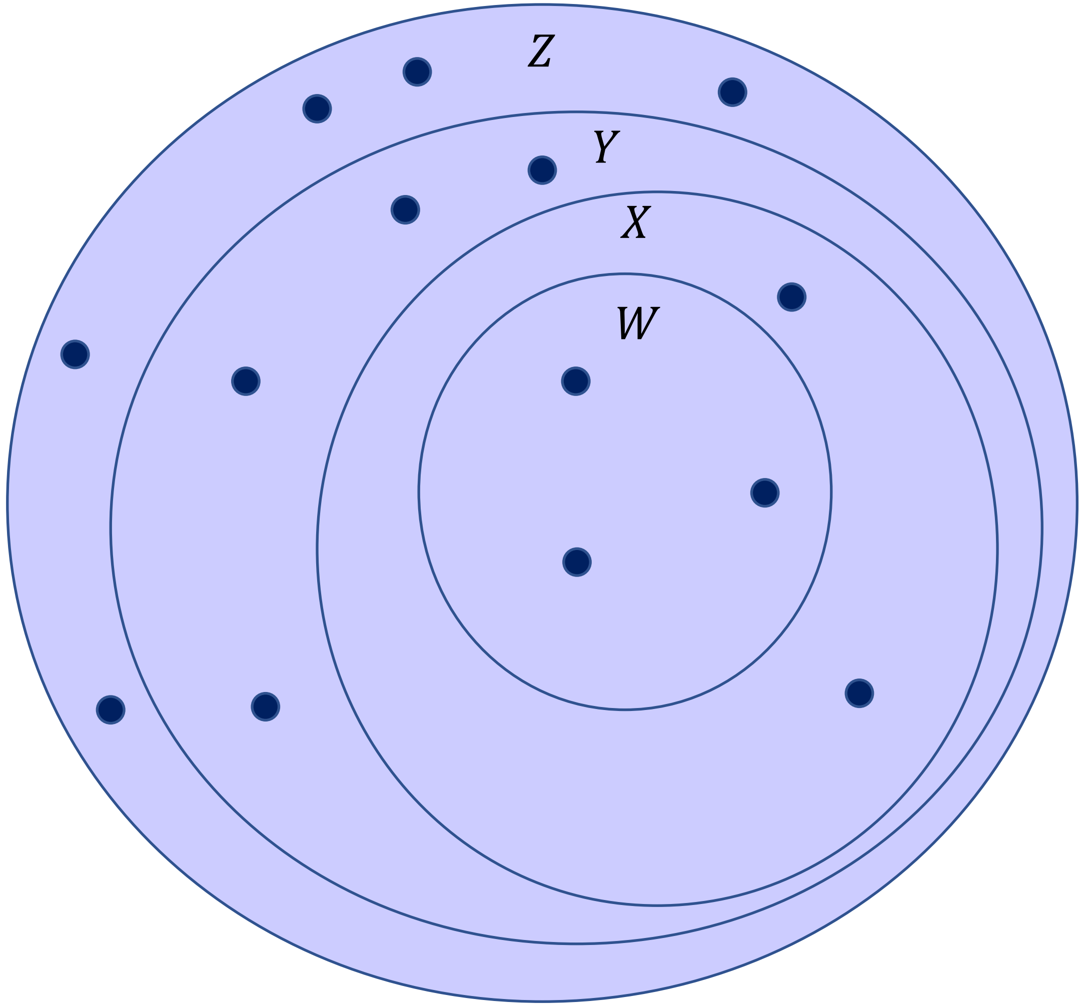


$$w(W \mid Z) = w(W \mid Y)\, w(Y \mid Z) = w(W \mid X)\, w(X \mid Z)$$

Chain Rule for Probability

Knuth, K.H. and Skilling, J., 2012. Foundations of inference. *Axioms*, *1*(1), pp.38-73.
Knuth, K.H., 2019. Lattices and their consistent quantification. *Annalen der Physik*, *531*(3), p.1700370.

$$Z \supseteq Y \supseteq X \supseteq W$$



$$p(W|Z) = p(W\,|\,Y)\,p(Y\,|\,Z)$$
$$= p(W\,|\,X)\,p(X\,|\,Z)$$

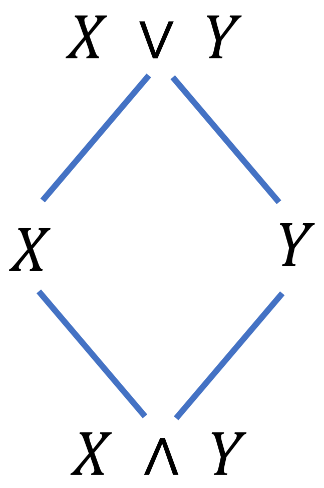
$$p(W\,|\,X) = \frac{3}{5} \qquad p(X\,|\,Z) = \frac{5}{14}$$

$$p(W\,|\,Y) = \frac{3}{9} \qquad p(Y\,|\,Z) = \frac{9}{14}$$

$$p(W\,|\,Z) = \frac{3}{14} = \frac{3}{9}\cdot\frac{9}{14} = \frac{3}{5}\cdot\frac{5}{14}$$

Knuth, K.H., 2009, December. Measuring on lattices. In AIP Conference Proceedings (Vol. 1193, No. 1, pp. 132-144). American Institute of Physics.

Consider this with context X



$$p(X \vee Y \mid X) = p(X \mid X) + p(Y \mid X) - p(X \wedge Y \mid X)$$

Since $X \rightarrow X \vee Y$, we have $p(X \vee Y \mid X) = 1$

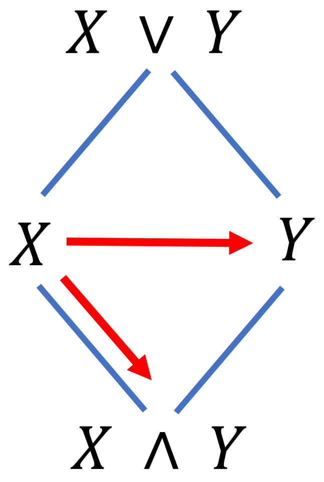
Since $X \rightarrow X$, we have $p(X \mid X) = 1$

$$1 = 1 + p(Y \mid X) - p(X \wedge Y \mid X)$$

$$p(Y \mid X) = p(X \wedge Y \mid X)$$

Knuth, K.H., 2009, December. Measuring on lattices. In AIP Conference Proceedings (Vol. 1193, No. 1, pp. 132-144). American Institute of Physics.

Consider this with context X



$$p(X \vee Y \mid X) = p(X \mid X) + p(Y \mid X) - p(X \wedge Y \mid X)$$

Since $X \to X \vee Y$, we have $p(X \vee Y \mid X) = 1$

Since $X \to X$, we have $p(X \mid X) = 1$

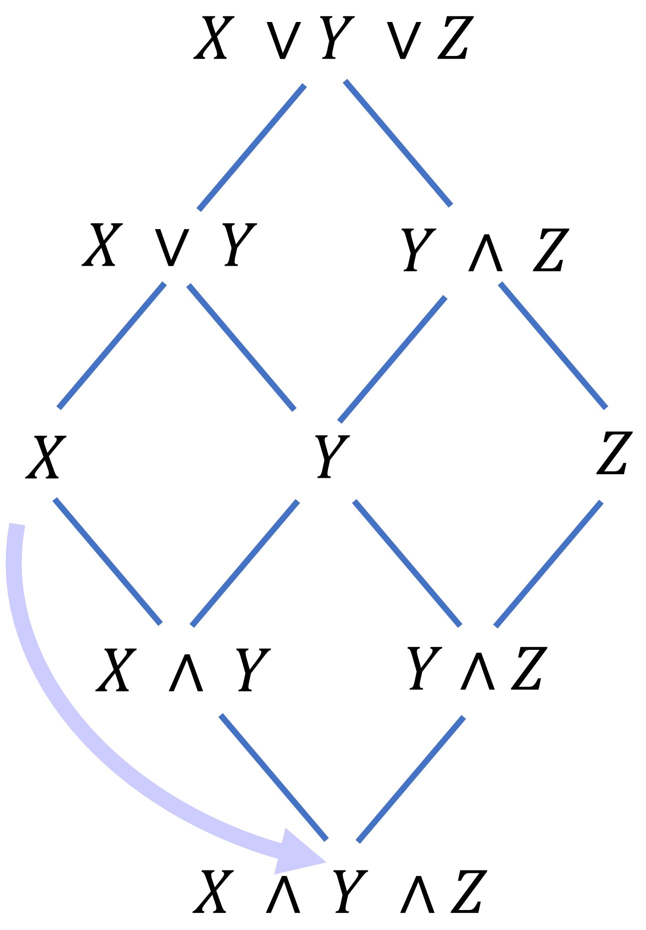
$$1 = 1 + p(Y \mid X) - p(X \wedge Y \mid X)$$

$$\boxed{p(Y \mid X) = p(X \wedge Y \mid X)}$$

We will be using this result in the derivation that follows.

The Product Rule

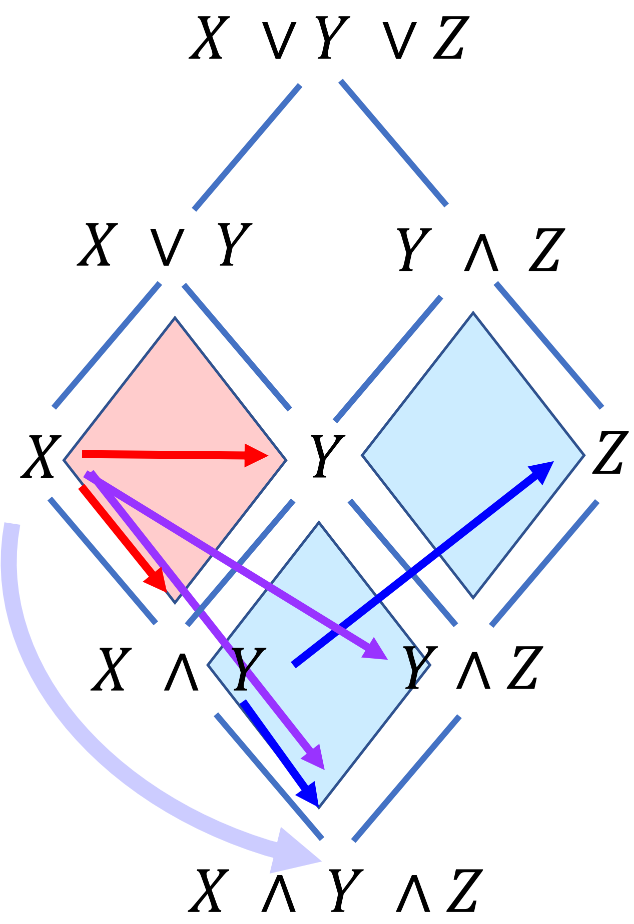
Knuth, K.H., 2009, December. Measuring on lattices. In AIP Conference Proceedings (Vol. 1193, No. 1, pp. 132-144). American Institute of Physics.



$$p(X \wedge Y \wedge Z \mid X) = p(X \wedge Y \wedge Z \mid X \wedge Y)\, p(X \wedge Y \mid X)$$

The Product Rule

Knuth, K.H., 2009, December. Measuring on lattices. In AIP Conference Proceedings (Vol. 1193, No. 1, pp. 132-144). American Institute of Physics.



$$p(X \wedge Y \wedge Z \mid X) = p(X \wedge Y \wedge Z \mid X \wedge Y)\, p(X \wedge Y \mid X)$$

$$p(X \wedge Y \wedge Z \mid X) = p(X \wedge Y \wedge Z \mid X \wedge Y)\, p(Y \mid X)$$

$$p(X \wedge Y \wedge Z \mid X) = p(Z \mid X \wedge Y)\, p(Y \mid X)$$

$$p(X \wedge Y \wedge Z \mid X) = p(Z \mid X \wedge Y)\, p(Y \mid X)$$

$$p(Y \wedge Z \mid X) = p(Z \mid X \wedge Y)\, p(Y \mid X)$$

Changing notation and rearranging

$$p(Y, Z \mid X) = p(Y \mid X)\, p(Z \mid X, Y)$$

Sum and Product Rules

$$p(A \vee B \mid I) = p(A \mid I) + p(B \mid I) - p(A \wedge B \mid I)$$ Sum Rule

$$p(A, B \mid I) = p(B \mid I)\, p(A \mid B, I)$$ Product Rule

$$= p(A \mid I)\, p(B \mid A, I)$$

Sum and Product Rules

$$p(A \vee B \mid I) = p(A \mid I) + p(B \mid I) - p(A \wedge B \mid I)$$ Sum Rule

$$\begin{aligned} p(A, B \mid I) &= p(B \mid I)\, p(A \mid B, I) \\ &= p(A \mid I)\, p(B \mid A, I) \end{aligned}$$ Product Rule

$$p(B \mid I)\, p(A \mid B, I) = p(A \mid I)\, p(B \mid A, I)$$

Sum and Product Rules

$$p(A \lor B \mid I) = p(A \mid I) + p(B \mid I) - p(A \land B \mid I)$$ Sum Rule

$$p(A, B \mid I) = p(B \mid I)\, p(A \mid B, I) = p(A \mid I)\, p(B \mid A, I)$$ Product Rule

$$p(B \mid I)\, p(A \mid B, I) = p(A \mid I)\, p(B \mid A, I)$$

$$p(A \mid B, I) = \frac{p(A \mid I)\, p(B \mid A, I)}{p(B \mid I)}$$

Sum and Product Rules + Bayes Theorem

$$p(A \lor B \mid I) = p(A \mid I) + p(B \mid I) - p(A \land B \mid I)$$ Sum Rule

$$p(A, B \mid I) = p(B \mid I)\, p(A \mid B, I)$$
$$= p(A \mid I)\, p(B \mid A, I)$$ Product Rule

Bayes Theorem

$$p(A \mid B, I) = \frac{p(A \mid I)\, p(B \mid A, I)}{p(B \mid I)}$$

Bayes Theorem

Bayes Theorem

$$p(A \mid B, I) = \frac{p(A \mid I)\, p(B \mid A, I)}{p(B \mid I)}$$

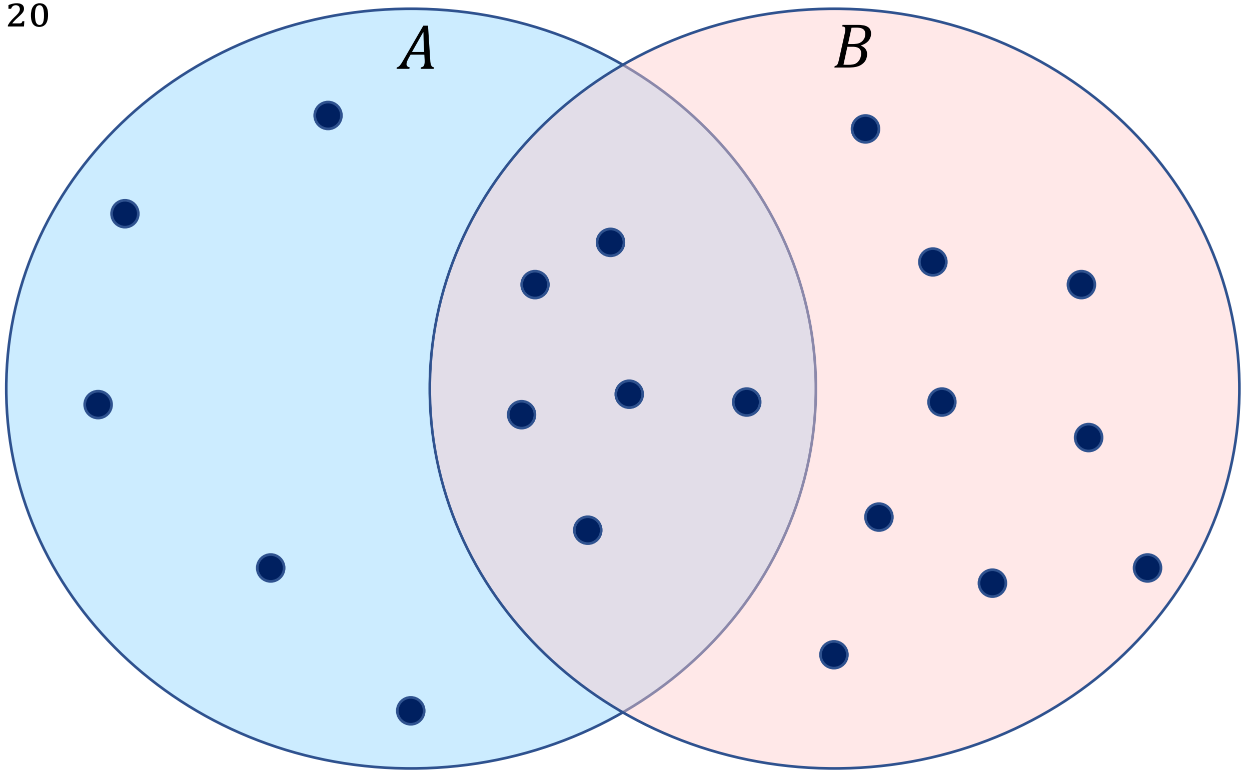
$$p(A \mid B, I) = \frac{6}{15} = \frac{\frac{6}{11} \cdot \frac{11}{20}}{\frac{15}{20}}$$

$p(A \mid I) = \frac{11}{20}$

$p(B \mid I) = \frac{15}{20}$

$p(B \mid A, I) = \frac{6}{11}$

$p(A \mid B, I) = \frac{6}{15}$



Bayes Theorem

Bayes Theorem

$$p(A \mid B, I) = \frac{p(A \mid I)\, p(B \mid A, I)}{p(B \mid I)}$$

$A \rightarrow model$
$B \rightarrow data$

$$p(model \mid data, I) = \frac{p(model \mid I)\, p(data \mid model, I)}{p(data \mid I)}$$

Bayes Theorem

$$p(model \mid data, I) = \frac{p(model \mid I)\, p(data \mid model, I)}{p(data \mid I)}$$

Prior Probability: The probability of the model based only on one's prior information

Likelihood: The probability that the data could have been observed given the model

Evidence: The probability that the data could have been observed based only on the prior information
This term often serves as a normalization factor

Posterior Probability: The probability of the model based both on the prior information and the data

Bayes’ Theorem as a Learning Rule

Bayes Theorem as a Learning Rule

$$p(model \mid data, I) = p(model \mid I)\frac{p(data \mid model, I)}{p(data \mid I)}$$

Data-dependent Term

One's prior belief about a model (prior probability) is modified by a data-dependent term resulting in the posterior probability, which describes one's state of belief considering both prior information and data

Parallel versus Sequential Learning

Consider that we have N pieces of independent data: $d_1, d_2, \cdots, d_N$

We can consider the data as a compound logical statement $D = d_1 \wedge d_2 \wedge \cdots \wedge d_N$ and use Bayes' Theorem

$$p(model \mid D, I) = p(model \mid I) \frac{p(D \mid model, I)}{p(D \mid I)}$$

Data are considered in parallel

$$p(model \mid D, I) = p(model \mid I) \frac{p(d_1 \wedge d_2 \wedge \cdots \wedge d_N \mid model, I)}{p(d_1 \wedge d_2 \wedge \cdots \wedge d_N \mid I)}$$

apply the product rule

$$p(model \mid D, I) = p(model \mid I) \frac{p(d_1 \mid model, I)\, p(d_2 \wedge \cdots \wedge d_N \mid model, I)}{p(d_1 \mid I)\, p(d_2 \wedge \cdots \wedge d_N \mid I)}$$

$$p(model \mid D, I) = p(model \mid I) \frac{p(d_1 \mid model, I)}{p(d_1 \mid I)} \frac{p(d_2 \wedge \cdots \wedge d_N \mid model, I)}{p(d_2 \wedge \cdots \wedge d_N \mid I)}$$

The posterior for d_1 can serve as the prior for the remaining data

$$p(model \mid D, I) = p(model \mid d_1, I) \frac{p(d_2 \wedge \cdots \wedge d_N \mid model, I)}{p(d_2 \wedge \cdots \wedge d_N \mid I)}$$

Parallel versus Sequential Learning

$$p(model\,|D,I) = p(model\,|d_1,I)\,\frac{p(d_2 \wedge \cdots \wedge d_N|\,model,I)}{p(d_2 \wedge \cdots \wedge d_N|\,I)}$$

$$p(model\,|D,I) = p(model\,|d_1,I)\,\frac{p(d_2\,|model,I)}{p(d_2\,|\,I)}\frac{p(d_3 \wedge \cdots \wedge d_N|\,model,I)}{p(d_3 \wedge \cdots \wedge d_N|\,I)}$$

$$p(model\,|D,I) = p(model\,|d_1,I)\,\frac{p(d_2\,|model,I)}{p(d_2\,|\,I)}\frac{p(d_3 \wedge \cdots \wedge d_N|\,model,I)}{p(d_3 \wedge \cdots \wedge d_N|\,I)}$$

$$\vdots$$

$$p(model\,|D,I) = p(model\,|\,I)\,\frac{p(d_1\,|model,I)}{p(d_1\,|\,I)}\frac{p(d_2\,|model,I)}{p(d_2\,|\,I)}\cdots\frac{p(d_N\,|model,I)}{p(d_N\,|\,I)}$$

where the **data are considered sequentially**.

The posterior at each step is then used as the prior for the next step.

Normalization and Marginalization

Normalization

Recall that probability is normalized so that the sum of the probability over all possibilities is equal to 1.

Let $\{a_1, a_2, \cdots, a_N\}$ be an exhaustive set of mutually exclusive logical statements

Since the set is exhaustive, the statement $a_1 \vee a_2 \vee \cdots \vee a_N$ is known to be TRUE.

$$p(a_1 \vee a_2 \vee \cdots \vee a_N \mid I) = 1$$

Applying the sum rule

$$p(a_1 \mid I) + p(a_2 \mid I) + \cdots + p(a_N \mid I) = 1$$

$$\sum_{i=1}^{N} p(a_i \mid I) = 1$$

Summation with Multiple Parameters

Let $\{b_1, b_2, \cdots, b_M\}$ be an exhaustive set of mutually exclusive logical statements.

Look at

$$
\begin{aligned}
\sum_{k=1}^{M} p(a, b_k | I) &= \sum_{k=1}^{M} p(a \mid I)\, p(b_k | a, I) \\
&= p(a | I) \sum_{k=1}^{M} p(b_k | a, I) \\
&= p(a | I) \cdot 1 \\
\sum_{k=1}^{M} p(a, b_k | I) &= p(a | I)
\end{aligned}
$$

Marginalization

Let $\{b_1, b_2, \cdots, b_M\}$ be an exhaustive set of mutually exclusive logical statements.

$$p(a \mid I) = \sum_{k=1}^{M} p(a, b_k \mid I)$$

This technique is called MARGINALIZATION.

Using the Sum Rule, one can MARGINALIZE over one of the parameters to obtain the probability of the remaining parameters.

This allows one to get rid of uninteresting parameters thus reducing the dimensionality of the problem.

Marginalization

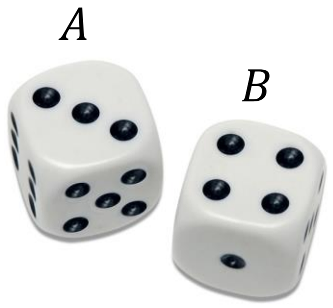
Consider rolling two six-sided dice (A and B), each with probabilities $p(\, a_i \,|I) = p(\, b_k \,|I) = \frac{1}{6}$

$$p(\, a_i, b_k \mid I) \; = p(\, a_i \mid I)\, p(\, b_k \mid I)$$

b / a	**1**	**2**	**3**	**4**	**5**	**6**	
1	1/36	1/36	1/36	1/36	1/36	1/36	
2	1/36	1/36	1/36	1/36	1/36	1/36	
3	1/36	1/36	1/36	1/36	1/36	1/36	
4	1/36	1/36	1/36	1/36	1/36	1/36	
5	1/36	1/36	1/36	1/36	1/36	1/36	
6	1/36	1/36	1/36	1/36	1/36	1/36	

$$p(\, a_i, b_k \mid I) \; = p(\, a_i \mid I)\, p(\, b_k \mid I)$$

$$= \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

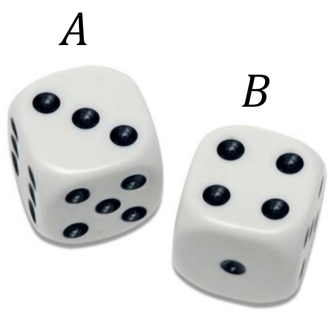


Marginalization

Consider rolling two six-sided dice (A and B), each with probabilities $p(\, a_i \,|I) = p(\, b_k \,|I) = \frac{1}{6}$

$$p(\, a_i, b_k \mid I) \;= p(\, a_i \mid I)\, p(\, b_k \mid I)$$

b / a	**1**	**2**	**3**	**4**	**5**	**6**	$p(\, a_i \mid I)$
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
3	1/36	1/36	1/36	1/36	1/36	1/36	1/6
4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
6	1/36	1/36	1/36	1/36	1/36	1/36	1/6



$$p(\, a_i, b_k \mid I) \;= p(\, a_i \mid I)\, p(\, b_k \mid I) = \frac{1}{6}\cdot\frac{1}{6} = \frac{1}{36}$$

SUM OVER THE POSSIBLE FACES OF DIE B

$$\sum_{k=1}^{M} p(a_i, b_k \mid I) \;= p(\, a_i \mid I)$$

This is called MARGINALIZATION because it used to be computed by summing and writing the result in the MARGIN of the paper.