ISTA410/INFO510 Bayesian Modelling and Inference

Lecture 3 – Probability Review

Dr. Kunal Arekar
College of Information
University of Arizona, Tucson

September 3, 2025

Basic terminology and rules

Sample Space of *outcomes* (often denoted by Ω)

{H, T}

{1, 2, 3, 4, 5, 6}

An outcome is just ONE element of the sample space

A "generic" outcome is often denoted by ω

and we can say things like, e.g., "for each $\omega \in \Omega$…"

Basic terminology and rules

Sample Space of *outcomes* (often denoted by Ω)

{H, T}

{1, 2, 3, 4, 5, 6}

An outcome is just ONE element of the sample space
A "generic" outcome is often denoted by ω
and we can say things like, e.g., "for each $\omega \in \Omega$…"

Event (subset of Ω) …does or does not contain (is true or false for) a particular outcome

odd {1, 3, 5}, even {2, 4, 6}, prime {2, 3, 5}

Basic terminology and rules

Sample Space of *outcomes* (often denoted by Ω)

{H, T}

{1, 2, 3, 4, 5, 6}

An outcome is just ONE element of the sample space
A "generic" outcome is often denoted by ω
and we can say things like, e.g., "for each $\omega \in \Omega$…"

Event (subset of Ω) …does or does not contain (is true or false for) a particular outcome

odd {1, 3, 5}, even {2, 4, 6}, prime {2, 3, 5}

Semantics of Set Operations

Equivalence between "set" and "proposition" representations.

1. Set E: outcomes s.t. proposition E is true.
2. Union, $E \cup F$: logical OR between propositions E and F.
3. Intersection, $E \cap F$: logical AND
4. Complement, E^C: logical negation

Basic terminology and rules

Sample Space of *outcomes* (often denote

{H, T}

{1, 2, 3, 4, 5, 6}

An outcome is just ONE element of the sample space

A “generic” outcome is often denoted by ω
and we can say things like, e.g., “for each $\omega \in \Omega$…”

Event (subset of Ω) …does or does not contain (is true or false for) a particular outcome

odd {1, 3, 5}, even {2, 4, 6}, prime {2, 3, 5}

Denote the **collection of measurable events** (ones we want to assign probabilities to) by S.

S must include $\varnothing$ and Ω

These special events represent the cases where “nothing” among all the choices happens (impossible), and “something” happens (certain).

Reason for being technical: It is important to be tuned into **what** a particular probability is **about** (precisely!).

Basic terminology and rules

Sample Space of *outcomes* (often denoted by Ω)

{H, T}

{1, 2, 3, 4, 5, 6}

An outcome is just ONE element of the sample space
A "generic" outcome is often denoted by ω
and we can say things like, e.g., "for each $\omega \in \Omega$…"

Event (subset of Ω) …does or does not contain (is true or false for) a particular outcome

odd {1, 3, 5}, even {2, 4, 6}, prime {2, 3, 5}

Denote the **collection of measurable events** (ones we want to assign probabilities to) by S.

S must include $\varnothing$ and Ω

S is *closed* under set operations

…aka: σ-algebra

$$\alpha, \beta \in \mathrm{S} \Rightarrow \alpha \cup \beta \in \mathrm{S}, \ \alpha \cap \beta \in \mathrm{S}, \ \alpha^C = \Omega - \alpha \in \mathrm{S}, \text{ etc.}$$

Translation: We need to be able to deal with concepts such as "either A or B" happens, or "both A and B" happen.

E.g., I'll accept either an even or prime number

E.g., If I roll a 3, it is both odd and prime

Basic terminology and rules

Probability Space

A **probability space** is a sample space augmented with a function, P, that assigns a **probability** to each event, $E \subset S$.

Kolmogorov Axioms

1. $0 \leq P(E) \leq 1$ for all $E \subset S$. **Non-Negativity**
2. $P(\Omega) = 1$. **Normalization**
3. If $E \cap F = \emptyset$ then $P(E \cup F) = P(E) + P(F)$. **Additivity**

Important Consequences

1. $P(\emptyset) = 0$.
2. $P(E^{C}) = 1 - P(E)$ **Complement Rule**
3. In general, $P(E \cup F) = P(E) + P(F) - P(E \cap F)$. **General Addition Rule**

Random Variables

Random variables

- Defined by functions mapping outcomes to values
- A random variable is a way of reporting an attribute of an outcome
- By choice, whatever we are interested in
- Typically denoted by uppercase letters (e.g., X)
- Generic values are corresponding lower case letters
- Shorthand: P(x) = P(X=x)
- Value “type” is arbitrary (typically categorical or real)

Example

- Outcomes are student grades (A,B,C)
- Random variable G=f_{GRADE}(student)

$$P('A') = P(G = 'A') = P(\left\{w \in \Omega : f_{GRADE}(w) = 'A'\right\})$$

We sometimes use sets, but usually R.Vs.: $\overbrace{P(A \cap B \cap C)}^{\text{Sets}} \equiv \overbrace{P(A, B, C)}^{\text{R. Vs.}}$

Random Variables

Random Variable

- Formally, a **random variable** is a function, X that assigns a number to each outcome in S (e.g., dead $\rightarrow 0$, alive $\rightarrow 1$).
- Key consequence: a random variable divides the sample space into **equivalence classes**: sets of outcomes that share some property (differ only in ways irrelevant to X)

Example

- Let $S =$ all sequences of 3 coin tosses.
- We can define a r.v. X that counts number of heads.
- Then HHT and HTH are equivalent in the eyes of X:

$$X(HHT) = X(HTH) = 2$$

Random Variables

Distribution of a Random Variable

- The expression $P(X = x)$ refers to the probability of the event $E = \{\omega \in S : X(\omega) = x\}$.
- Sometimes we can obtain it by breaking it down into simpler, mutually exclusive events and adding their probabilities (Kolmogorov axiom 3)

Example

- S = all sequences of 3 coin tosses.
- $X(\omega) = \#$ of heads in ω.

$$
\begin{aligned}
\{X = 2\} &= \{HHT\} \cup \{HTH\} \cup \{THH\} \\
P(X = 2) &= P(HHT) + P(HTH) + P(THH) \\
&= \frac{1}{8} + \frac{1}{8} + \frac{1}{8}
\end{aligned}
$$

Random Variables

Distribution of a Random Variable

- Similarly, $P(X < x)$ is the probability of the event $E = \{\omega \in S : X(\omega) < x\}$.
- Can sometimes obtain it the same way as we did above.

Example

- $S =$ all sequences of 3 coin tosses.
- $X(\omega) =$ # of heads in ω.

$$\begin{aligned}
\{X < 2\} &= \{TTT\} \cup \{TTH\} \cup \{THT\} \cup \{HTT\} \\
P(X < 2) &= P(TTT) + P(TTH) + P(THT) + P(HTT) \\
&= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}
\end{aligned}$$

Random Variables

Distribution of a Random Variable

Example, continued

- Notice that in this example we could also have written

$$\{X < 2\} = \{X = 0\} \cup \{X = 1\}$$
$$P(X < 2) = P(X = 0) + P(X = 1)$$

  which is useful if we have already calculated $P(X = x)$ for each value of x.
- This always works if X is always an integer.

Joint Probability

Joint Probability

- We have already seen the concept of *intersecting events*: $A \cap B$ is the event that occurs when *both* A and B are true *at the same time.*
- $P(A \cap B)$ is called the **joint probability** of A and B.
- If A is $\{X = x\}$ and B is $\{Y = y\}$, then $A \cap B$ means $X = x$ and $Y = y$ *at the same time.*
- If X and Y are discrete, $P(X = x, Y = y)$, for different combinations of x and y, characterize the **joint distribution** of X and Y.

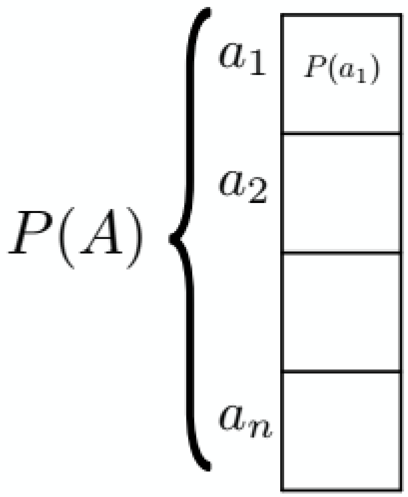
We write $P(x, y)$ for $P(\{ w \in \Omega : X(w) = x \text{ and } Y(w) = y \})$

Alternatively, $P\big((X = x) \cap (Y = y)\big)$

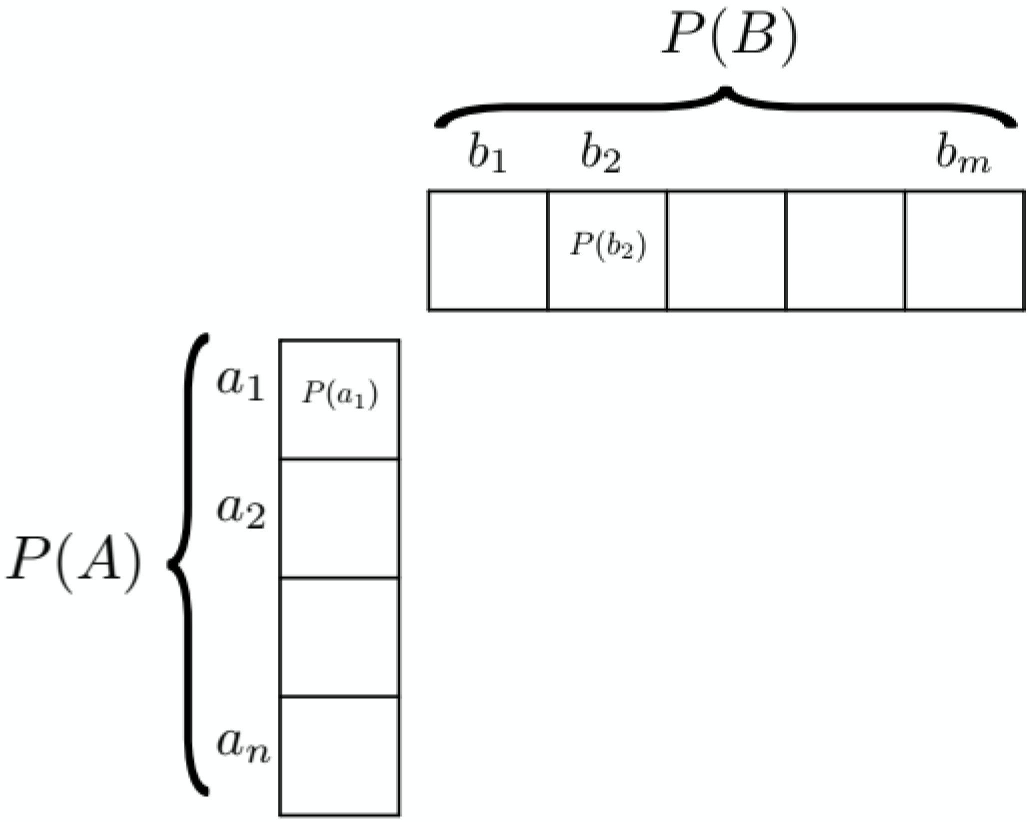
Note that the comma in the usual form, $P(x, y)$, is read as "and".

Here events are being defined by assignments of random variables

Joint Probability

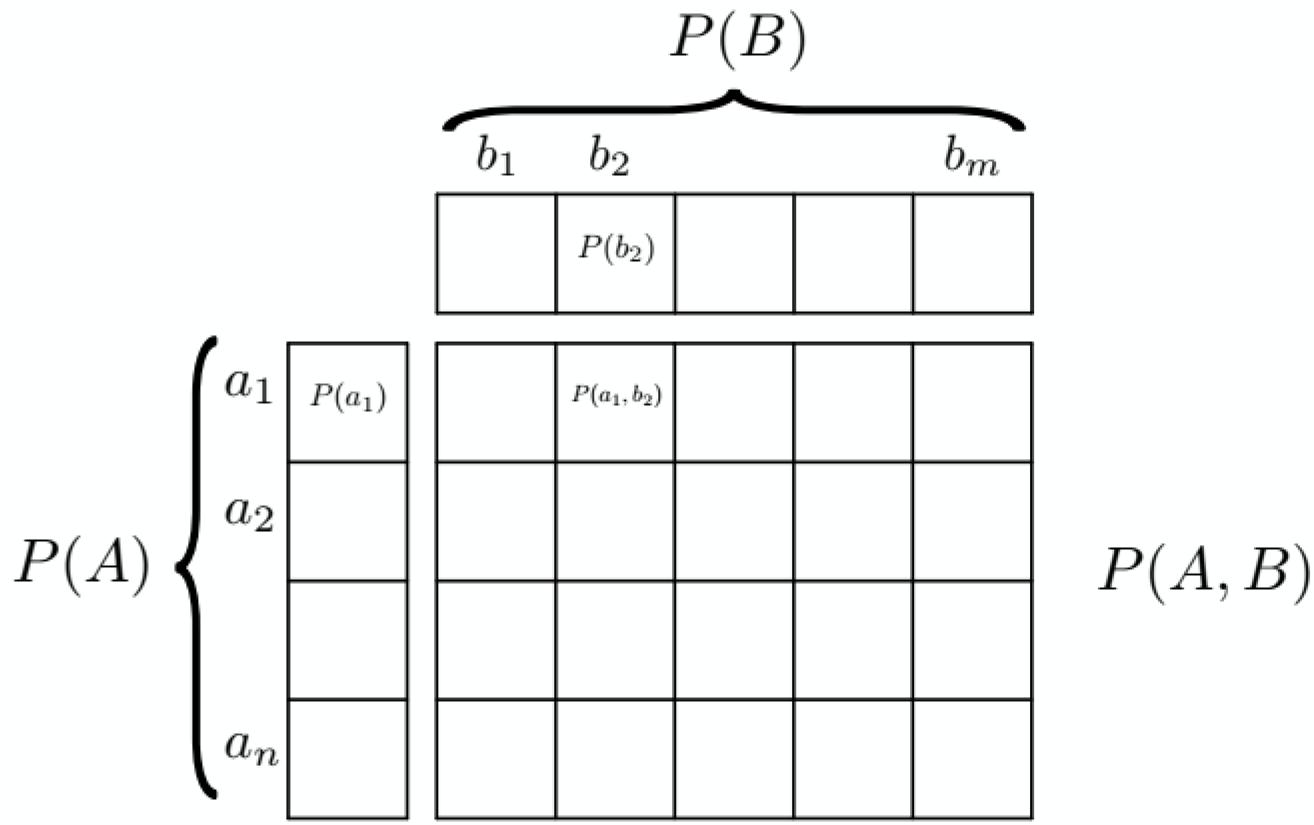


Joint Probability



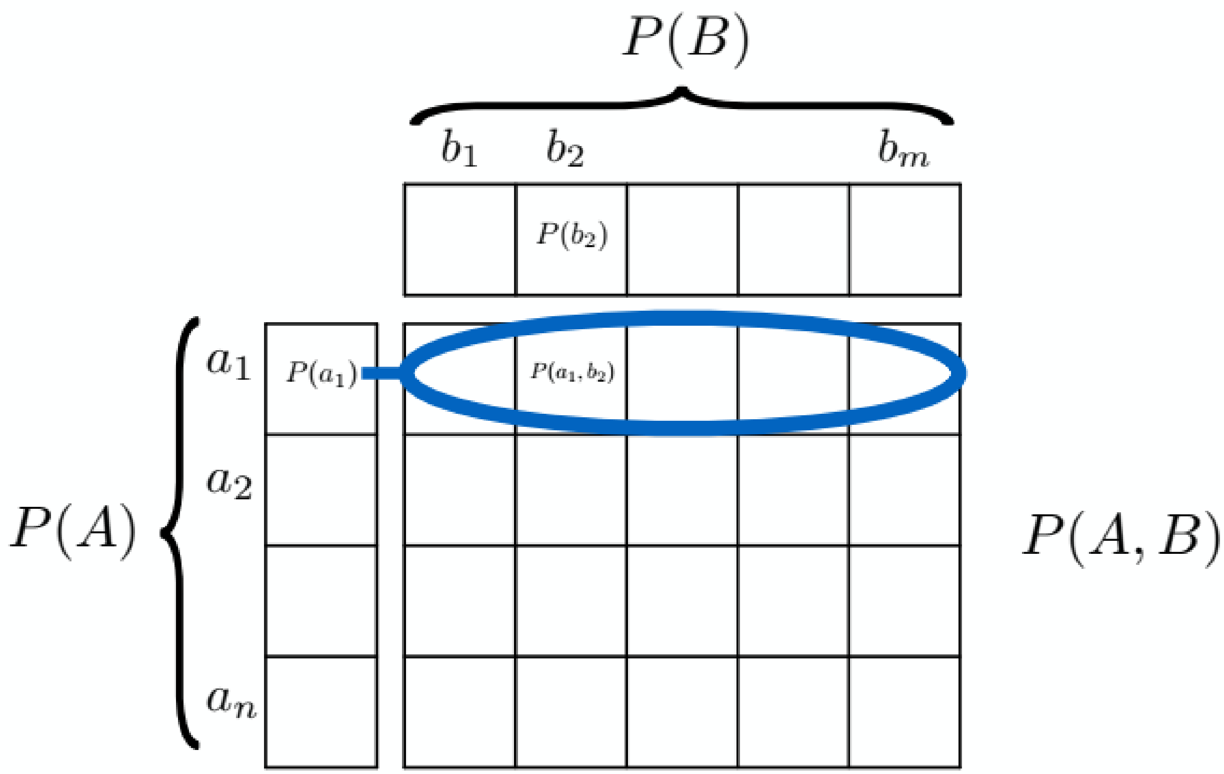
Joint Probability

Joint Probability



Joint Probability

Joint Probability



Marginalization: $$P(A) = \sum_{b \in B} P(A, B)$$

Marginalization: $$P(A) = \sum_{b \in B} P(A, B)$$

way to calculate the probability of a single event (like $P(A)$) by summing over all possible outcomes of another event (like B)

A|B	B = 1	B = 2	B = 3
A = 1	0.1	0.2	0.1
A = 2	0.2	0.1	0.3

P (A = 2)

P (A = 2) = P (A = 2, B = 1) + P (A = 2, B = 2) + P (A = 2, B = 3)

P (A = 2) = 0.2 + 0.1 + 0.3

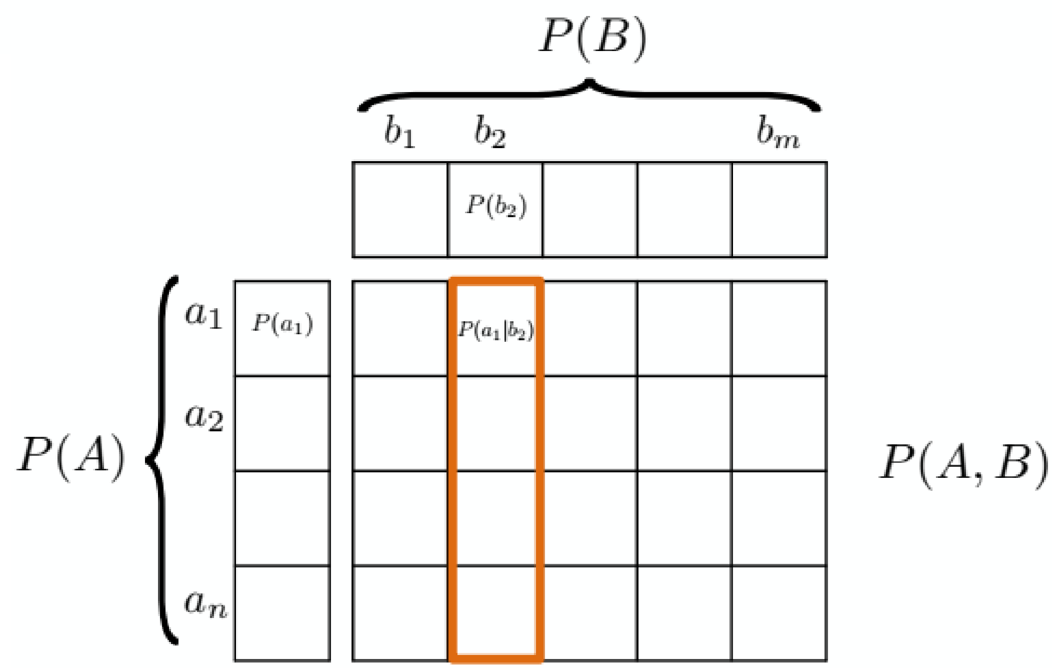
P (A = 2) = 0.6

Conditional Probability

"probability in context"

Conditional probability (definition)

$$P\left(A|B\right) \equiv \frac{P(A \cap B)}{P(B)}$$



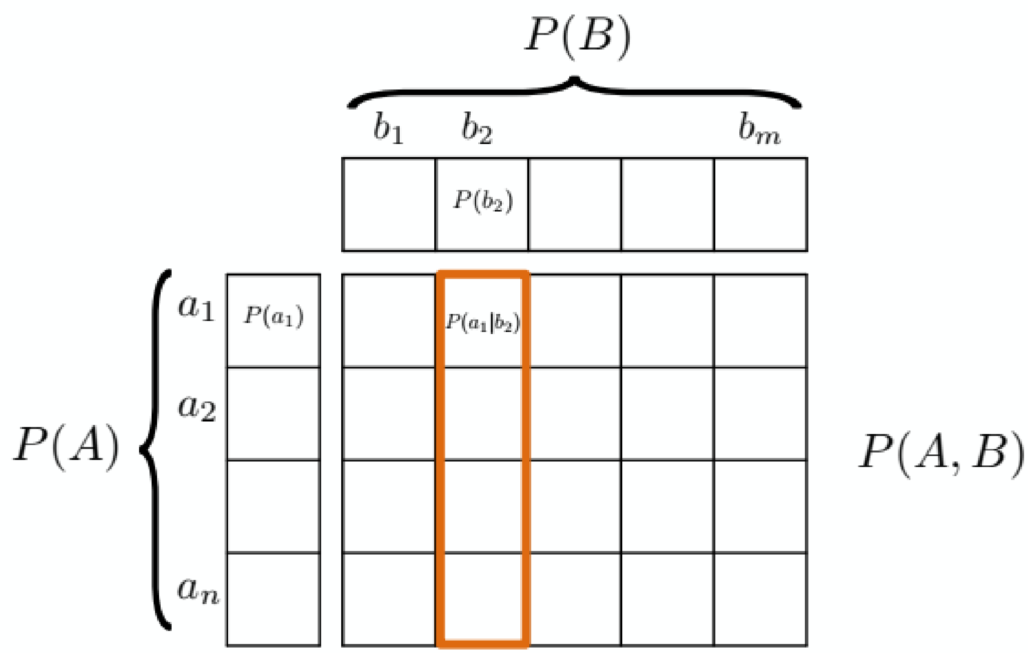
Conditional Probability

"probability in context"

Conditional probability (definition)

$$P(A|B) \equiv \frac{P(A \cap B)}{P(B)}$$

Example, what is the probability that you have rolled 2, given that you know you have rolled a prime number?



Product Rule

"probability in context"

Conditional probability (definition)

$$P(A|B) \equiv \frac{P(A \cap B)}{P(B)}$$

Applying a bit of algebra,

$$P(A \cap B) = P(B)P(A|B)$$

Chain (Product) Rule

"probability in context"

Conditional probability (definition)

$$P\left(A|B\right) \equiv \frac{P(A \cap B)}{P(B)}$$

Applying a bit of algebra,

$$P(A \cap B) = P(B)P(A|B)$$

In general, we have the **chain** (**product**) rule:

Product $$P\left(A_1 \cap A_2\right) = P(A_1)P(A_2|A_1)$$

Chain $$P\left(A_1 \cap A_2 \cap \quad .... \quad A_N\right) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \; .... \; P(A_N|A_1 \cap A_2 \cap \quad .... \quad A_{N-1})$$ *

Bayes Rule

Going back to the definition of conditional probability

$$P\left(A|B\right) \equiv \frac{P(A \cap B)}{P(B)}$$

Applying a little bit more algebra,

$$P\left(A \cap B\right) = P(A)P(B|A)$$

and $$P\left(A \cap B\right) = P(B)P(A|B)$$

and thus $$P(B)P(A|B) = P(A)P(B|A)$$

and we get $$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$ **Bayes rule** *

Bayes Rule

Going back to the definition of conditional probability

$$P(A|B) \equiv \frac{P(A \cap B)}{P(B)}$$

Applying a little bit more algebra,

$$P(A \cap B) = P(A)P(B|A)$$

and $$P(A \cap B) = P(B)P(A|B)$$

and thus $$P(B)P(A|B) = P(A)P(B|A)$$

and we get $$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Common to represent denominator as marginalization of numerator:

$$P(B) = \sum_{a \in A} P(A,B)$$
$$= \sum_{a \in A} P(A)P(B|A)$$

Bayes rule *

Example Using Bayes Rule (**KF**, example 2.2)

- Suppose a TB test is 95% accurate

$$P(\text{positive} \mid \text{TB}) = 0.95$$

$$P(\text{negative} \mid \neg\text{TB}) = 0.95$$

Example Using Bayes Rule (**KF**, example 2.2)

- Suppose a TB test is 95% accurate

$$P(\text{positive} \mid \text{TB}) = 0.95$$

$$P(\text{negative} \mid \neg\text{TB}) = 0.95 \qquad P(\text{positive} \mid \neg\text{TB}) = 0.05$$

Example Using Bayes Rule (**KF**, example 2.2)

- Suppose a TB test is 95% accurate

$$P(\text{positive} \mid \text{TB}) = 0.95$$

$$P(\text{negative} \mid \neg\text{TB}) = 0.95$$

$$P(\text{positive} \mid \neg\text{TB}) = 0.05$$

- What is $P(\text{TB} \mid \text{positive})$?

Example Using Bayes Rule (**KF**, example 2.2)

- Suppose a TB test is 95% accurate

$$P(\text{positive} \mid \text{TB}) = 0.95$$

$$P(\text{negative} \mid \neg\text{TB}) = 0.95$$

$$P(\text{positive} \mid \neg\text{TB}) = 0.05$$

- What is $P(\text{TB} \mid \text{positive})$?
  - The naive approach:
    - If the test result is wrong 5% of the time, then probability subject is infected is 0.95.
    - I.e., 95% of subjects with positive results have TB
  - What does the Bayesian say?

Example Using Bayes Rule (**KF**, example 2.2)

- Suppose a TB test is 95% accurate

$$P(\text{positive} \mid \text{TB}) = 0.95$$

$$P(\text{negative} \mid \neg\text{TB}) = 0.95$$

$$P(\text{positive} \mid \neg\text{TB}) = 0.05$$

- What is $P(\text{TB} \mid \text{positive})$?

$$P(\text{TB} \mid \text{pos}) = \frac{P(\text{pos} \mid \text{TB}) P(\text{TB})}{P(\text{pos})}$$

Example Using Bayes Rule (**KF**, example 2.2)

- Suppose a TB test is 95% accurate

$$P(\text{positive} \mid \text{TB}) = 0.95$$

$$P(\text{negative} \mid \neg\text{TB}) = 0.95$$

$$P(\text{positive} \mid \neg\text{TB}) = 0.05$$

- What is $P(\text{TB} \mid \text{positive})$?
  - $P(\text{TB}) = ?$
  - $P(\text{pos}) = ?$

$$P(\text{TB} \mid \text{pos}) = \frac{P(\text{pos} \mid \text{TB}) P(\text{TB})}{P(\text{pos})}$$

Example Using Bayes Rule (**KF**, example 2.2)

- Suppose a TB test is 95% accurate

$$P(\text{positive} \mid \text{TB}) = 0.95$$

$$P(\text{negative} \mid \neg\text{TB}) = 0.95$$

$$P(\text{positive} \mid \neg\text{TB}) = 0.05$$

- What is $P(\text{TB} \mid \text{positive})$?

$$P(\text{TB} \mid \text{pos}) = \frac{P(\text{pos} \mid \text{TB}) P(\text{TB})}{P(\text{pos})}$$

  - Suppose 1 in 1000, so $P(\text{TB}) = 0.001$
  - $P(\text{pos}) = ?$

Example Using Bayes Rule (**KF**, example 2.2)

- Suppose a TB test is 95% accurate

$$P(\text{positive} \mid \text{TB}) = 0.95$$

$$P(\text{negative} \mid \neg\text{TB}) = 0.95$$

$$P(\text{positive} \mid \neg\text{TB}) = 0.05$$

- What is $P(\text{TB} \mid \text{positive})$?

$$P(\text{TB} \mid \text{pos}) = \frac{P(\text{pos} \mid \text{TB}) P(\text{TB})}{P(\text{pos})}$$

  - Suppose 1 in 1000, so $P(\text{TB}) = 0.001$
  - $P(\text{pos}) = P(\text{pos}|\text{TB})P(\text{TB}) + P(\text{pos}|\neg\text{TB})P(\neg\text{TB})$

Example Using Bayes Rule (**KF**, example 2.2)

- Suppose a TB test is 95% accurate

$$P(\text{positive} \mid \text{TB}) = 0.95$$

$$P(\text{negative} \mid \neg\text{TB}) = 0.95$$

$$P(\text{positive} \mid \neg\text{TB}) = 0.05$$

- What is $P(\text{TB} \mid \text{positive})$?

$$P(\text{TB} \mid \text{pos}) = \frac{P(\text{pos} \mid \text{TB}) P(\text{TB})}{P(\text{pos})}$$

  - Suppose 1 in 1000, so $P(\text{TB}) = 0.001$
  - $P(\text{pos}) = P(\text{pos}|\text{TB})P(\text{TB}) + P(\text{pos}|\neg\text{TB})P(\neg\text{TB})$

    $= (0.95 * 0.001) + (0.05 * 0.999)$

    $= 0.0509$

- Now plug in to Bayes rule:

Example Using Bayes Rule (**KF**, example 2.2)

- Suppose a TB test is 95% accurate

$$P(\text{positive} \mid \text{TB}) = 0.95$$

$$P(\text{negative} \mid \neg\text{TB}) = 0.95 \qquad P(\text{positive} \mid \neg\text{TB}) = 0.05$$

- What is $P(\text{TB} \mid \text{positive})$?

$$P(\text{TB} \mid \text{pos}) = \frac{P(\text{pos} \mid \text{TB}) P(\text{TB})}{P(\text{pos})}$$

  - Suppose 1 in 1000, so $P(\text{TB}) = 0.001$
  - $P(\text{pos}) = P(\text{pos}|\text{TB})P(\text{TB}) + P(\text{pos}|\neg\text{TB})P(\neg\text{TB})$

$$= (0.95 * 0.001) + (0.05 * 0.999)$$

$$= 0.0509$$

- Now plug in to Bayes rule: $\frac{0.95 \times 0.001}{0.0509} \approx 0.0187$

Example Using Bayes Rule (**KF**, example 2.2)

- Suppose a TB test is 95% accurate

$$P(\text{positive} \mid \text{TB}) = 0.95$$

$$P(\text{negative} \mid \neg\text{TB}) = 0.95 \qquad P(\text{positive} \mid \neg\text{TB}) = 0.05$$

- What is $P(\text{TB} \mid \text{positive})$? $\quad P(\text{TB} \mid \text{pos}) = \dfrac{P(\text{pos} \mid \text{TB})P(\text{TB})}{P(\text{pos})}$
  - Suppose 1 in 1000, so $P(\text{TB}) = 0.001$
  - $P(\text{pos}) = P(\text{pos}|\text{TB})P(\text{TB}) + P(\text{pos}|\neg\text{TB})P(\neg\text{TB})$

    $= (0.95 * 0.001) + (0.05 * 0.999)$

    $= 0.0509$

- Now plug in to Bayes rule: $\dfrac{0.95 \times 0.001}{0.0509} \approx 0.0187$

The bottom line: although a subject with a positive test is much more likely to be TB-infected than is a random subject (by almost 20 times)…

… **fewer than 2 percent** of those subjects are TB-infected.

Basic rules (so far)

Marginalization

$$P(A) = \sum_{b \in B} P(A, B)$$ *

Conditional probability (definition)

$$P(A|B) \equiv \frac{P(A \cap B)}{P(B)}$$ *

Chain (Product) Rule

$$P(A_1 \cap A_2) = P(A_1)P(A_2|A_1)$$ *

$$P(A_1 \cap A_2 \cap \; .... \; A_N) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \; .... \; P(A_N|A_1 \cap A_2 \cap \; .... \; A_{N-1})$$

Bayes Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$ *

Normalization

Often we will deal with quantities or functions which are *proportional* to probabilities (they don't currently sum to 1, so need to scale them).

To convert such quantities to probabilities we *normalize*.

If $p(x) \propto P(X = x)$ then $P(X = x) = \dfrac{p(x)}{\sum_x p(x)}$

Example: $P(X \mid Y) \propto P(X, Y)$

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

Probabilistic Queries

Organize variables into

Evidence (observed), **E**

Query (what you want to know), **Y**

Hidden (leftover), **X** (for completeness) – latent variable

Generic Query: P(**Y**|**E**)

This leads to a *distribution* over **Y** given the evidence Note that **X** is marginalized out

We can use this to make a decision

Simplest is most probable, i.e., $\underset{\mathbf{Y}}{\text{Argmax}}\, P(\mathbf{Y}, \mathbf{E})$

“Maximum *a posteriori*”

MAP Query (most probably configuration of variables):

$$MAP(\mathbf{W} \mid \mathbf{E}) = \underset{\mathbf{W}}{\text{Argmax}}\, P(\mathbf{W}, \mathbf{E}) \qquad (\mathbf{W} = \mathbf{Y} \cup \mathbf{X})$$

Assume this table is conditioned on E

		Y: y_1	y_2	
X	x_1	0.04	0.30	0.34
	x_2	0.36	0.30	0.66
		0.40	0.60	

Argmax $P(x,y)$ is (x_2, y_1)

Argmax $P(x)$ is (x_2)

Argmax $P(y)$ is (y_2)

Arg max $P(x,y)$ is **not necessarily** (Arg max $P(x)$, Arg max $P(y)$)