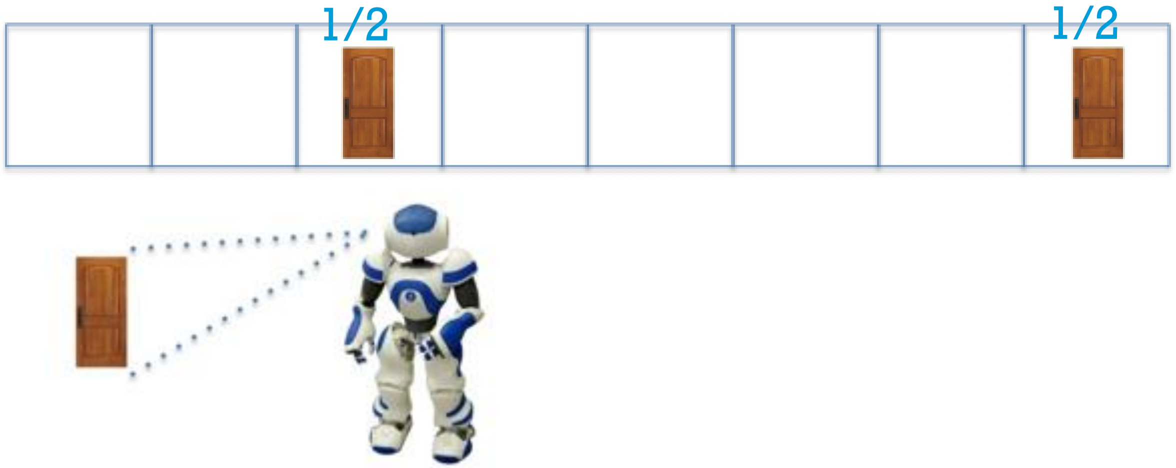
Grado en ingeniería informática
Artificial Intelligence 2021/2022

4. Uncertainty

Lara Quijano Sánchez



Universidad Autónoma de Madrid

Uncertainty in AI

❑Formalization of uncertainty through probabilities

❑Bayes theorem in AI

❑Bayesian Networks

Readings

❑ Chapter 1 of Bishop

❑ Chapter 1,2 of Jaynes

4.1 Probability as a measure of expectation

- ❑Probabilities can be viewed in two ways
  - ❑Frequencies of outcomes in a repeated experiment
  - ❑Reasonable expectations of outcomes in a single trial

Which interpretation to use for probabilistic agents?

Cox In "Probability, Frequency, and Reasonable Expectation," Am. Jour. Phys. 14, 1–13, (1946) argues for 2nd interpretation:

- ❑Let A be an assertion over the world in particular situation, characterized by I (information available):

$P(A \mid I)$: Estimate of how likely is A given I.

$P(A \mid I)=0$ A is impossible, given I.

$P(A \mid I)=1$ A is certain, given I.

The value of $P(A \mid I)$ could be different for different probabilistic agents, who have access to different information I

Probabilities

- ❑**Probability** can be interpreted as a measure of:
    - ❑proportion of times something is true
        - ❑20 students passed the exam out of 22
        - ❑a physical phenomenon
        - ❑can be experimentally measured
    - ❑degree of belief over something
        - ❑I think Real Madrid will win the Liga with a 80%
        - ❑can vary over people
        - ❑or intelligent system
- ❑**Probability calculus** does not depend on the interpretation
    - ❑Probabilities range in [0.0..1.0]
    - ❑Probability=0: false
    - ❑Probability=1: true

Probabilities and causality

- Probabilities represent logical connections, not causal connections
  - A ⇒ B should not be understood as "A is the physical cause of B“
  - By equivalence ¬B ⇒ ¬A and “¬B is the physical cause of ¬A” (?)
    - Eg. ¬BATTERY_OK ⇒ ¬WORKING (causal?)
      - “ The device does not work because the battery is not OK "
      - WORKS ⇒ BATTERY_OK (NO CAUSAL)
      - The workingof the device is not the physical cause that the battery is OK
    - Eg. Clouds are the physical cause of rain.
      - However, Clouds ⇒ Rain is incorrect.
      - The correct assertion is Rain ⇒ Clouds, which cannot be understood as "rain is the physical cause of clouds"

Probabilities and causality

- Experiment with extraction of an urn
  - Experiment 1: Urn with 1 red ball and 5 black balls.
    - The probability of drawing a red ball is 1/6.
  - Experiment 2: Urn with 1 red ball and 5 black balls.
    - A red ball is drawn from the urn and not returned to it.

      In a second extraction a ball is extracted.

      Since there is only one red ball in the urn, the probability of having observed a red ball on the second draw is 0.

      The probability of the result of the second draw depends on the result of the first.

      However, the second extraction cannot causally affect the first.

Using Probabilities in AI

- Typical tasks: decision making, classification, prediction ,. ..
  - What's true? a physical phenomenon
    - use of classical logic: propositional satisfaction, production systems, shortest path, chess, etc.
  - vs. What is more likely?
    - Use of probabilities: Bayesian networks, sequence prediction (speech recognition), classification (of language), weather forecast, video games
  - What if the selected model is wrong?
    - In classical logic:
      - incomplete model → ok
      - wrong model → problem
    - In probabilities
      - In general, it is more interesting to know the relationship between the probabilities than the exact numbers: P (e)> P (ex)?
      - could be more robust

Random variables

- Propositional logic
  - we describe states as sets of **boolean variables**: p, q, r
  - an interpretation is a **truth assignment** to those variables: p = V, q = F, r = V
- Probability theory
  - we use a set of **random variables** that can take values on a given **domain**:
    - one dice: $X \in \{1, 2, 3, 4, 5, 6\}$
    - two dices: $X \in \{1, 2, 3, 4, 5, 6\}$, $Y \in \{1, 2, 3, 4, 5, 6\}$
  - the associated **value** to a random variable is **unknown**
  - we can assign a **probability** to each value
    - dice: $P(X = 1) = 1/6 , . . . , P(X = 6) = 1/6$
  - these probabilities define a **probability distribution**

Example

- ❑Given a robot in a 100 × 100 grid with a given orientation
- ❑Define its random variables and domains
    - ❑Random variables
        - ❑$X \in \{0, . . . , 99\}, Y \in \{0, . . . , 99\}, \theta \in \{0, . . . , 359\}$

A "priori" probability

- The probability distribution of a random variable is usually represented as a (vector)
  - Example:
    - $P(X) = (P(X = 0), \ldots, P(X = 99))$
- The joint probability distribution is the distribution for several variables.
  - E.g. $P(X, Y)$, $P(X, Y, \theta)$
    - $P(X, Y) = (P(X = 0, Y = 0), P(X = 0, Y = 1), \ldots, P(X = 99, Y = 99))$

      $= (\frac{1}{10000}, \frac{1}{10000}, \ldots\ldots, \frac{1}{10000})$
- This distribution is "a priori" or unconditional, since it does not depend on any condition

Example

- Given a robot in a 100 × 100 grid with a given orientation
- Define its random variables, domains, and probability distribution
  - Random variables

    $X \in \{0, . . . , 99\}, Y \in \{0, . . . , 99\}, \theta \in \{0, . . . , 359\}$
  - Probability distribution (unknown position):

    $P(X = 0, Y = 0) = P(X = 0, Y = 1) = . . . =$

    $P(X = 0, Y = 99) = P(X = 1, Y = 0) = . . . =$

    $P(X = 99, Y = 99) = \frac{1}{100\ x\ 100}$
  - Probability distribution (unknown position and orientation):
    - $P(X = 0, Y = 0, \theta = 0) = . . .$
    - $P(X = 100, Y = 100, \theta = 360) = \frac{1}{100\ x\ 100\ x\ 360}$

Law of total probability

- Given a set of pairwise disjoint events A_i such that their union is the whole sample space and another event B:
  - $P\ (B) = \sum_{i=1}^{n} P(\ B, Ai) = \sum_{i=1}^{n} \mathrm{P\ (B|Ai)\ P(Ai)}$

- Thus, if we have a random variable A with possible disjoint values $a_1, \ldots, a_n$ and an event B:
  - $P\ (B) = \sum_{i=1}^{n} P(\ B, A = ai) = \sum_{i=1}^{n} \mathrm{P\ (B|A= a_i)\ P(A= a_i)}$
- Example
  - $P\ (X = 0) = \sum_{i=1}^{99} P(X = 0, Y = i) = \mathrm{P(X = 0, Y = 0) + P(X = 0, Y = 1) + \ldots + P(X = 0, Y = 99)} = \frac{1}{10000} + \frac{1}{10000} + \cdots + \frac{1}{10000} = \frac{100}{10000} = \frac{1}{100}$

Conditional probability

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

- P (A | B) can be interpreted as the updated probability of A, once B has been observed
  - Examples
    - $P(X = 0)$?
    - $P(X = 0 | X < 10)$?
    - $P(X = 0 | Y = 0)$?
  - $0 \leq P(A \mid I) \leq 1$
  - $P(True \mid I) = 1, \qquad P(False \mid I) = 0$
  - Sum rule:
    - $P(A) + P(\neg A) = P(A) + P(\bar{A}) = 1$

Conditional probability

- Conditional probability
  - $P(A\,|B) = \frac{P(A \wedge B)}{P(B)}$ if P (B) ≠ 0
- Product Rule
  - P (A, B) = P (A $\wedge$ B) = P(A|B)P(B) =P(B|A) P(A)
- Bayes Rule
  - $P(A\,|B) = \frac{\mathrm{P(B|A)P(A)}}{\mathrm{P(B)}}$ = αP(B|A)P(A)
- Sometimes obtaining P(B|A) is easier than P(A|B)
  - it is usually easier to ask an expert P(Effect|Cause) than P(Cause|Effect)

Independence

- A and B are independent if any of these three cases:
  - $P(A|B) = P(A)$;
  - $P(B|A) = P(B)$; or
  - $P(A, B) = P(A)P(B)$
- Example
  - P(RobotX, Orientation, Dice) = P(RobotX)P(Orientation)P(Dice)
  - Reduction in the distribution size:
    - $100 \times 360 \times 6 = 216000 \sim 100 + 360 + 6 = 466$
- Smaller description implies
  - more efficient algorithms
  - less data (probabilities) to be specified

Uncertainty in AI

❑Formalization of uncertainty through probabilities

❑Bayes theorem in AI

❑Bayesian Networks

4.2 Bayes theorem in AI

❑Bayes theorem

$$P(H\,|D) = \frac{\mathrm{P(D|H)P(H)}}{\mathrm{P(D)}}$$

❑ Law of total probability

❑If we have the variables $H_1, . . . , H_n$ and an event D

$$P\,(D) = \sum_{i=1}^{n} \mathrm{P\,(D\,|\,H_i)\,P(H_i)}$$

❑ H: Hypothesis

❑ D: Data

❑ P (H): A priori probability of hypothesis H.

Probability of the hypothesis, **before** looking at the data.

❑ P (D | H): Likelihood of the hypothesis given the data.

❑ P (D): Evidence of the data.

It is independent of the hypothesis and functions as a normalization factor.

❑ P (H | D): A posteriori probability of the hypothesis.

Probability of the hypothesis, **after** observing the data.

Inference

- ❑Main task:
  - ❑Compute probabilities of events *e* given some evidence *o*: P(e|o)
- ❑Example:
  - ❑Compute posterior distribution given evidence
  - ❑Choose an action to achieve high reward given some evidence
  - ❑Decision making with optimal utility
  - ❑Classification
  - ❑Diagnosis

Inference

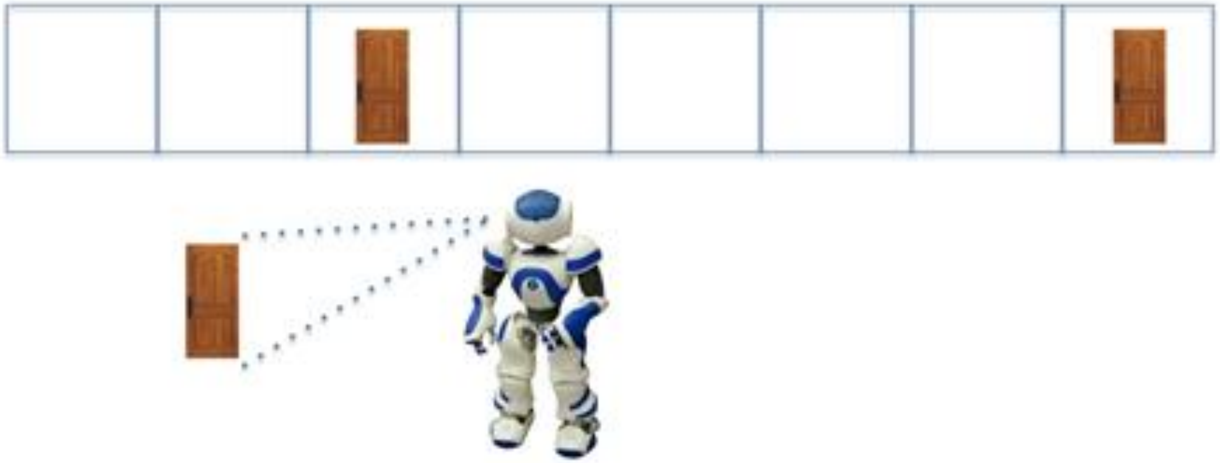
- □ Compute posterior distribution given evidence P(X |o)

1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

- □ $P\ (X) = (\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$

Inference

- ❑Compute posterior distribution given evidence P (X| o)



- ❑What is the probability distribution of the position of the robot (X) given that it observed a door (o =door), P(X |o = door)?
  - ❑ P (X| o= door) = $(0,0,\frac{1}{2},0,0,0,0,\frac{1}{2})$
  - ❑P (X=0| o= door) =$\frac{P(door|X=0)P(x=0)}{P\ (door)} = \frac{0\ x\ \frac{1}{8}}{\frac{2}{8}} = 0$
  - ❑P (X = 2| o= door) =$\frac{P(door|X=2)P(x=2)}{P\ (door)} = \frac{1\ x\ \frac{1}{8}}{\frac{2}{8}} = 0.5$

Inference

- Classification
  - given some observations to which class do they belong?
  - compares P (Class = 1 | o) versus P (Class = 2 | o)
- examples:
  - given some customer data (average money in the bank, home location, monthly earnings), determine whether it will correctly repay a mortgage (Class = 1) or not (Class = 2)
  - given some image (number of pixels of a given luminosity, number of lines), determine if it belongs to a cat (Class = cat), a dog (Class = dog), or something different (Class other)
- Computing (Naïve Bayes)
  - Class = arg max$_{c \in \text{Classes}}$ P (Class = c | o)
- Diagnosis
  - probability of a disease I = 1 given the results of the analysis o, P (I = 1 | o)

Inference from the data

- ❑Maximum likelihood (ML):

  Selects the hypothesis that maximizes the likelihood of the hypothesis given the data

  $$H^*_{ML} = \arg\max_H P(D \mid H)$$

  - ❑ML does not use information from the prioris (equivalent to assuming a uniform priori, P (H) = constant)

- ❑Maximum posterior (MAP)

  Selects the hypothesis that maximizes the posterior probability

  $$H^*_{MAP} = \arg\max_H P(H \mid D) = \arg\max_H P(D \mid H)P(H)$$

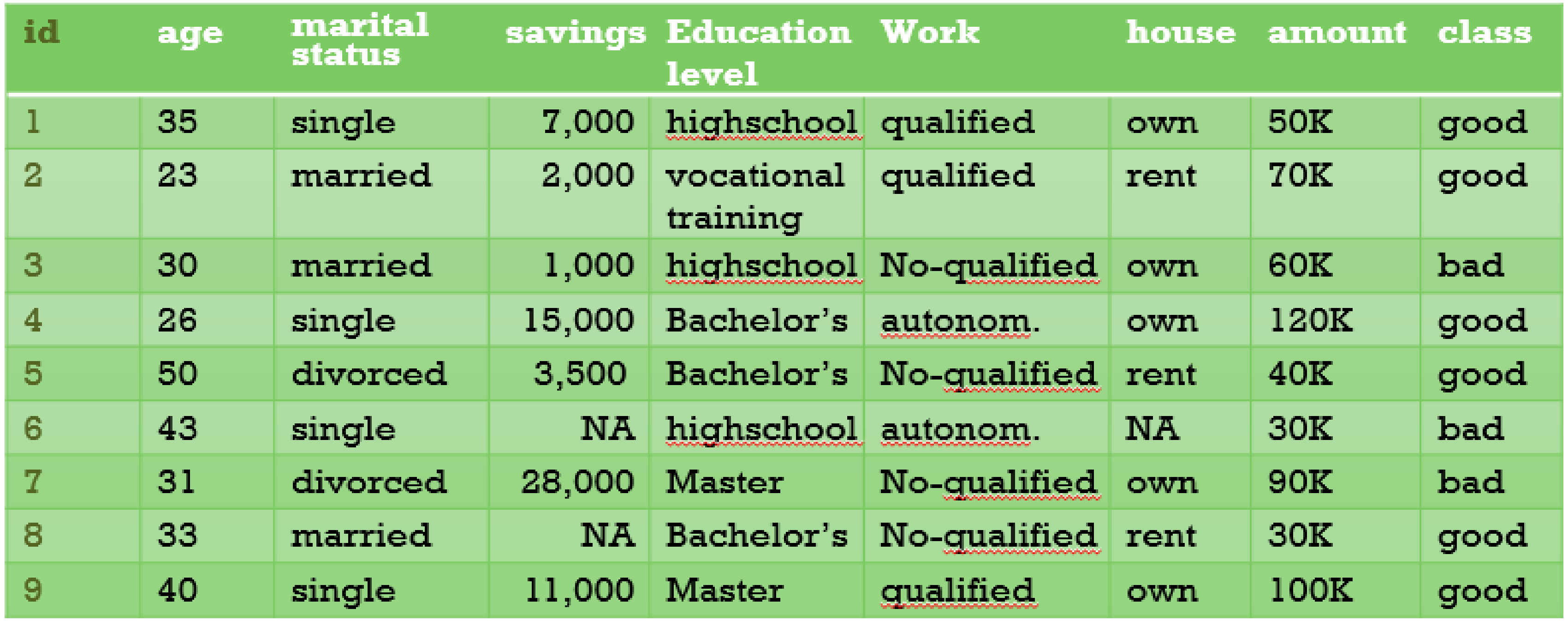
- ❑Bayesian Inference
  - ❑Average over all hypotheses with probabilities

Bayesian Inference

- Bayesian decision theory is based on two assumptions
  - The decision problem can be described in probabilistic terms
  - All probabilities of the problem are known or at least can be estimated
- Decisions are made based on observed data
- Notation
  - Set of classes: $C = \{c_1, c_2, . . . , c_m\}$
  - Attribute set: $A = \{a_1, a_2, . . . , a_n\}$
  - Instance (attribute values): $X = \{x_1, x_2, . . . , x_K\}$
  - Conditional probabilities:
    - $P(c_j \mid X)$ Probability of observing class c_j given instance X
    - $P(X \mid c_j)$ Probability of observing instance X given class c_j

Bayesian classifiers

- ❑ Observations



id	age	marital status	savings	Education level	Work	house	amount	class
1	35	single	7,000	highschool	qualified	own	50K	good
2	23	married	2,000	vocational training	qualified	rent	70K	good
3	30	married	1,000	highschool	No-qualified	own	60K	bad
4	26	single	15,000	Bachelor's	autonom.	own	120K	good
5	50	divorced	3,500	Bachelor's	No-qualified	rent	40K	good
6	43	single	NA	highschool	autonom.	NA	30K	bad
7	31	divorced	28,000	Master	No-qualified	own	90K	bad
8	33	married	NA	Bachelor's	No-qualified	rent	30K	good
9	40	single	11,000	Master	qualified	own	100K	good

- ❑ Decision problem
  - ❑ P(class = **good** | work = qualified, savings = 50 - 100K,home = own, age= 35-40)
  - ❑ P(class = **bad** | work = qualified, savings = 50 - 100K,home = own, age= 35-40)
  - ❑ ¿**good** o **bad**?

Example I: Is it raining?

- Today's weather prediction is 20 % chance of rain.

$$P(H = \text{"}rain\text{"}) = 0.2$$
$$P(H = \text{"}no\ rain\text{"}) = \mathbf{0.8}$$

$H^*_{prior} = \text{"}no\ rain\text{"}$

- The agent is in a windowless room and cannot directly determine whether its is raining or not. However, the agent detects that someone has just entered the room carrying an umbrella.
- The agent knows that the probability of someone carrying an umbrella is 70 % if it is raining and 10 % if it is not

$$P(D = \text{"}umbrella\text{"} \mid H = \text{"}rain\text{"}) = \mathbf{0.7}$$

$$P(D = \text{"}umbrella\text{"} \mid H = \text{"}no\ rain\text{"}) = 0.1$$

$H^*_{ML} = \text{"}rain\text{"}$

MAP solution

Use Bayes Theorem to compute posteriors

- Priors

$$P(H = \text{"rain"}) = 0.2 \quad P(H = \text{"no rain"}) = 0.8$$

- Likelihoods

$$P(D = \text{"umbrella"} \mid H = \text{"rain"}) = 0.7; \; P(D = \text{"umbrella"} \mid H = \text{"no rain"}) = 0.1$$

- Evidence

$$P(D = \text{"umbrella"}) = P(D = \text{"umbrella"} \mid H = \text{"rain"})\, P(H = \text{"rain"}) + P(D = \text{"umbrella"} \mid H = \text{"no rain"}) P(H = \text{"no rain"})$$

$$= 0.7 \times 0.2 + 0.1 \times 0.8 = 0.22$$

- Posteriors

$$\frac{0.7 \times 0.2}{0.7 \times 0.2 + 0.1 \times 0.8}$$

$$P(H = \text{"rain"} \mid D = \text{"umbrella"}) = \frac{P(D=\text{"umbrella"} \mid H=\text{"rain"})\; P(H=\text{"rain"})}{P(D=\text{"umbrella"})} = \mathbf{0.64}$$

$$P(H = \text{"no rain"} \mid D = \text{"umbrella"}) = 0.36$$

$$H^*_{MAP} = \text{"rain"}$$

Example 2: taxi

- ❑ There has been a car accident related to a taxi and the taxi driver has fled. There are two taxi companies in the city: green (85%) and blue (15%). What is the probability that the taxi in the accident is from the blue company?
  - ❑ Answer P(H = blue) = 0.15 **P(H = green) = 0.85** (priors)

Example 2: taxi

- What if there is a witness (80% who is telling the truth) who says that the taxi responsible for the accident was from the blue company?
  - H = blue: "The accident was caused by a taxi from the blue company“
  - D = blue: "The witness says the taxi was blue",

    "15% of the city's taxis are blue" + "The degree of reliability of the witness is 80%“
  - Priors
    - P(H = blue) = 0.15   P(H = green) = 0.85
  - Likelihoods
    - $P(D=$“blue" | $H=$“blue")=0.8
    - $P(D=$“green" | $H=$“green")=0.8 => $P(D=$“blue" | $H=$“green")=0.2

$$H^*_{ML} = \text{"}blue\text{"}$$

  - Posteriors
    - P(H=blue|D=blue) = $\dfrac{P(D=\text{“blue" | }H=\text{“blue")P(H = blue)}}{P(D=blue)} = \dfrac{0.8\ x\ 0.15}{P(D=blue)} = \dfrac{0.12}{P(D=blue)}$
    - P(H= green| D= blue)= $\dfrac{P(D=\text{“blue" | }H=\text{"green")P(H = green)}}{P(D=blue)} = \dfrac{0.2\ x\ 0.85}{P(D=blue)} = \dfrac{0.17}{P(D=blue)}$
  - Normalization
    - P(H=blue|D=blue) + P(H= green| D= blue) = 1

$$\frac{0.12}{P(D=blue)} + \frac{0.17}{P(D=blue)} = 1 \rightarrow P(D = blue) = 0.29$$

  - Result
    - P(H=blue|D=blue) = 0.41
    - **P(H= green| D= blue) = 0.59**

$$H^*_{MAP} = \text{"}green\text{"}$$

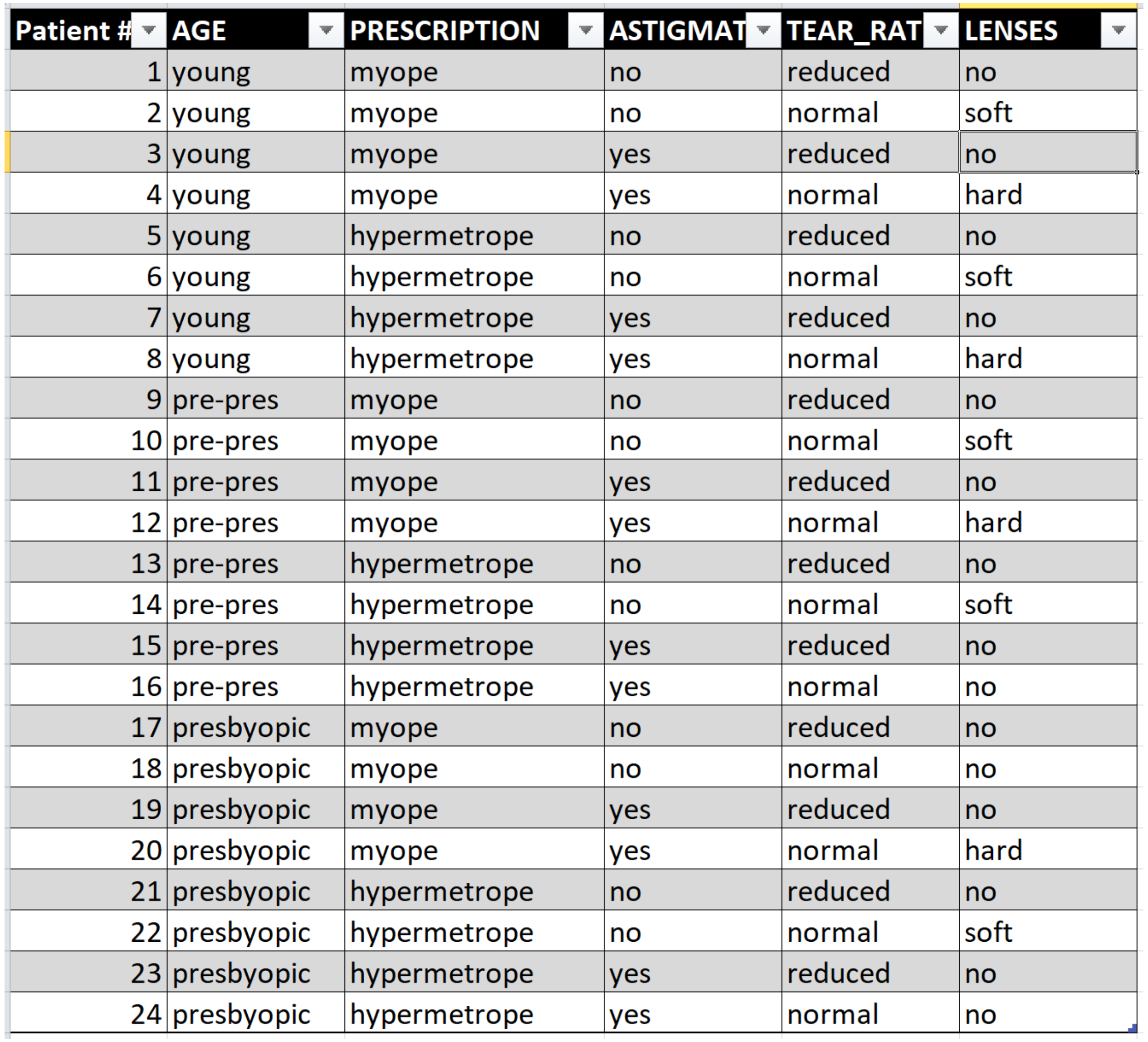
Example 3: contact lenses recommendation

Attributes (data):

- ❑ Age (a)
- ❑ Prescription (p)
- ❑ Astigmatism (as)
- ❑ Tear rate (tr)

Class (hypothesis):

- ❑ Which type of lenses should the patient wear? (c)



Patient #	AGE	PRESCRIPTION	ASTIGMAT	TEAR_RAT	LENSES
1	young	myope	no	reduced	no
2	young	myope	no	normal	soft
3	young	myope	yes	reduced	no
4	young	myope	yes	normal	hard
5	young	hypermetrope	no	reduced	no
6	young	hypermetrope	no	normal	soft
7	young	hypermetrope	yes	reduced	no
8	young	hypermetrope	yes	normal	hard
9	pre-pres	myope	no	reduced	no
10	pre-pres	myope	no	normal	soft
11	pre-pres	myope	yes	reduced	no
12	pre-pres	myope	yes	normal	hard
13	pre-pres	hypermetrope	no	reduced	no
14	pre-pres	hypermetrope	no	normal	soft
15	pre-pres	hypermetrope	yes	reduced	no
16	pre-pres	hypermetrope	yes	normal	no
17	presbyopic	myope	no	reduced	no
18	presbyopic	myope	no	normal	no
19	presbyopic	myope	yes	reduced	no
20	presbyopic	myope	yes	normal	hard
21	presbyopic	hypermetrope	no	reduced	no
22	presbyopic	hypermetrope	no	normal	soft
23	presbyopic	hypermetrope	yes	reduced	no
24	presbyopic	hypermetrope	yes	normal	no

Bayes inference I

- Example we observe : Patient: myopic + normal tear rate
- Class priors: $P(c = "n") = \mathbf{\frac{15}{24}}; P(c = "s") = \frac{5}{24}; P(c = "h") = \frac{4}{24};$

$H^*_{prior} = 'n'$

- Likelihoods:

$$P(p = "m", tr = "n" | c = "n") = \frac{1}{15}$$

$$P(p = "m", tr = "n" | c = "s") = \frac{2}{5}$$

$$P(p = "m", tr = "n" | c = "h") = \mathbf{\frac{3}{4}}$$

$H^*_{ML} =' h'$

- Evidence: $P(p = "m", tr = "n") = \frac{1}{15} \times \frac{15}{24} + \frac{2}{5} \times \frac{5}{24} + \frac{3}{4} \times \frac{4}{24} = \frac{1}{4}$
- Posteriors:

$H^*_{MAP} =' h'$

$$P(c = "n" | p = "m", tr = "n") = \frac{P(p = "m", tr = "r" | c = "n") P(c = "n")}{P(p = "m", tr = "r")} = \frac{\frac{1}{15} \times \frac{15}{24}}{P(p = "m", tr = "r")} = \frac{1}{6}$$

$$P(c = "s" | p = "m", tr = "n") = \frac{P(p = "m", tr = "r" | c = "s") P(c = "s")}{P(p = "m", tr = "r")} = \frac{\frac{2}{5} \times \frac{5}{24}}{P(p = "m", tr = "r")} = \frac{1}{3}$$

$$P(c = "h" | p = "m", tr = "n") = \frac{P(p = "m", tr = "r" | c = "h") P(c = "h")}{P(p = "m", tr = "r")} = \frac{\frac{3}{4} \times \frac{4}{24}}{P(p = "m", tr = "r")} = \mathbf{\frac{1}{2}}$$

Bayes inference II

- ❑ Example given we observe: Patient: myopic + reduced tear rate
- ❑ Class priors: $P(c = "n") = \frac{\mathbf{15}}{\mathbf{24}}; P(c = "s") = \frac{5}{24}; P(c = "h") = \frac{4}{24};$
- ❑ Likelihoods:

$$P(p = "m", tr = "r" | c = "n") = \frac{\mathbf{6}}{\mathbf{15}}$$

$$P(p = "m", tr = "r" | c = "s") = \frac{0}{5}$$

$$P(p = "m", tr = "r" | c = "h") = \frac{0}{4}$$

$H^*_{prior} = 'n'$

$H^*_{ML} = 'n'$

- ❑ Evidence: $P(p = "m", tr = "r") = \frac{6}{15} \times \frac{15}{24} + \frac{0}{5} \times \frac{5}{24} + \frac{0}{4} \times \frac{4}{24} = \frac{6}{24}$
- ❑ Posteriors:

$H^*_{MAP} = 'n'$

$$P(c = "n" | p = "m", tr = "r") = \frac{P(p = "m", tr = "r" | c = "n") P(c = "n")}{P(p = "m", tr = "r")} = \frac{\frac{6}{15} \times \frac{15}{24}}{P(p = "m", tr = "r")} = \mathbf{1}$$

$$P(c = "s" | p = "m", tr = "r") = \frac{P(p = "m", tr = "r" | c = "s") P(c = "s")}{P(p = "m", tr = "r")} = \frac{\frac{0}{5} \times \frac{5}{24}}{P(p = "m", tr = "r")} = 0$$

$$P(c = "h" | p = "m", tr = "r") = \frac{P(p = "m", tr = "r" | c = "h") P(c = "h")}{P(p = "m", tr = "r")} = \frac{\frac{0}{4} \times \frac{4}{24}}{P(p = "m", tr = "r")} = 0$$

Bayes inference III

- Example given we observe: Patient: hypermetrope+ normal tear rate
- Class priors: $P(c = "n") = \frac{\mathbf{15}}{\mathbf{24}}; P(c = "s") = \frac{5}{24}; P(c = "h") = \frac{4}{24};$
- Likelihoods:

$$P(p = "h", tr = "n" | c = "n") = \frac{2}{15}$$

$$P(p = "h", tr = "n" | c = "s") = \frac{\mathbf{3}}{\mathbf{5}}$$

$$P(p = "h", tr = "n" | c = "h") = \frac{1}{4}$$

$H^*_{prior} = 'n'$

$H^*_{ML} = 'n'$

- Evidence: $P(p = "h", tr = "n") = \frac{2}{24} + \frac{3}{24} + \frac{1}{24} = \frac{1}{4}$
- Posteriors:

$H^*_{MAP} = 'n'$

$$P(c = "n" | p = "h", tr = "n") = \frac{P(p = "h", tr = "n" | c = "n") P(c = "n")}{P(p = "h", tr = "n")} = \frac{\frac{2}{15} \times \frac{15}{24}}{P(p = "h", tr = "n")} = \frac{1}{3}$$

$$P(c = "s" | p = "h", tr = "n") = \frac{P(p = "h", tr = "n" | c = "s") P(c = "s")}{P(p = "h", tr = "n")} = \frac{\frac{3}{5} \times \frac{5}{24}}{P(p = "h", tr = "n")} = \frac{\mathbf{1}}{\mathbf{2}}$$

$$P(c = "h" | p = "h", tr = "n") = \frac{P(p = "h", tr = "n" | c = "h") P(c = "h")}{P(p = "h", tr = "n")} = \frac{\frac{1}{4} \times \frac{4}{24}}{P(p = "h", tr = "n")} = \frac{1}{6}$$

Classifier ML vs. Bayes classifier

- **Maximum Likelihood Classifier**: Assigns the class that maximizes the likelihood (probability of the conditional observation to the class) [ML]
- **Bayes classifier**: Assigns the class whose posterior probability (given the observation) is maximum [MAP]

prescripción	**lagrimeo**	**clase predicha (ML)**	**clase predicha (Bayes)**
miope	*normal*	*duras*	*duras* [50%]
miope	*reducido*	*no*	*no* [100%]
hipermétrope	*normal*	*blandas*	*blandas* [50%]
hipermétrope	*reducido*	*no*	*no* [100%]

- Uniform Priors ⇒ ML Classifier = Bayes Classifier
- In general, the predictions of the ML classifier may be different than those of the Bayes classifier.
- Bayes is optimal (minimizes the error).

Bayesian classifiers: Naïve Bayes

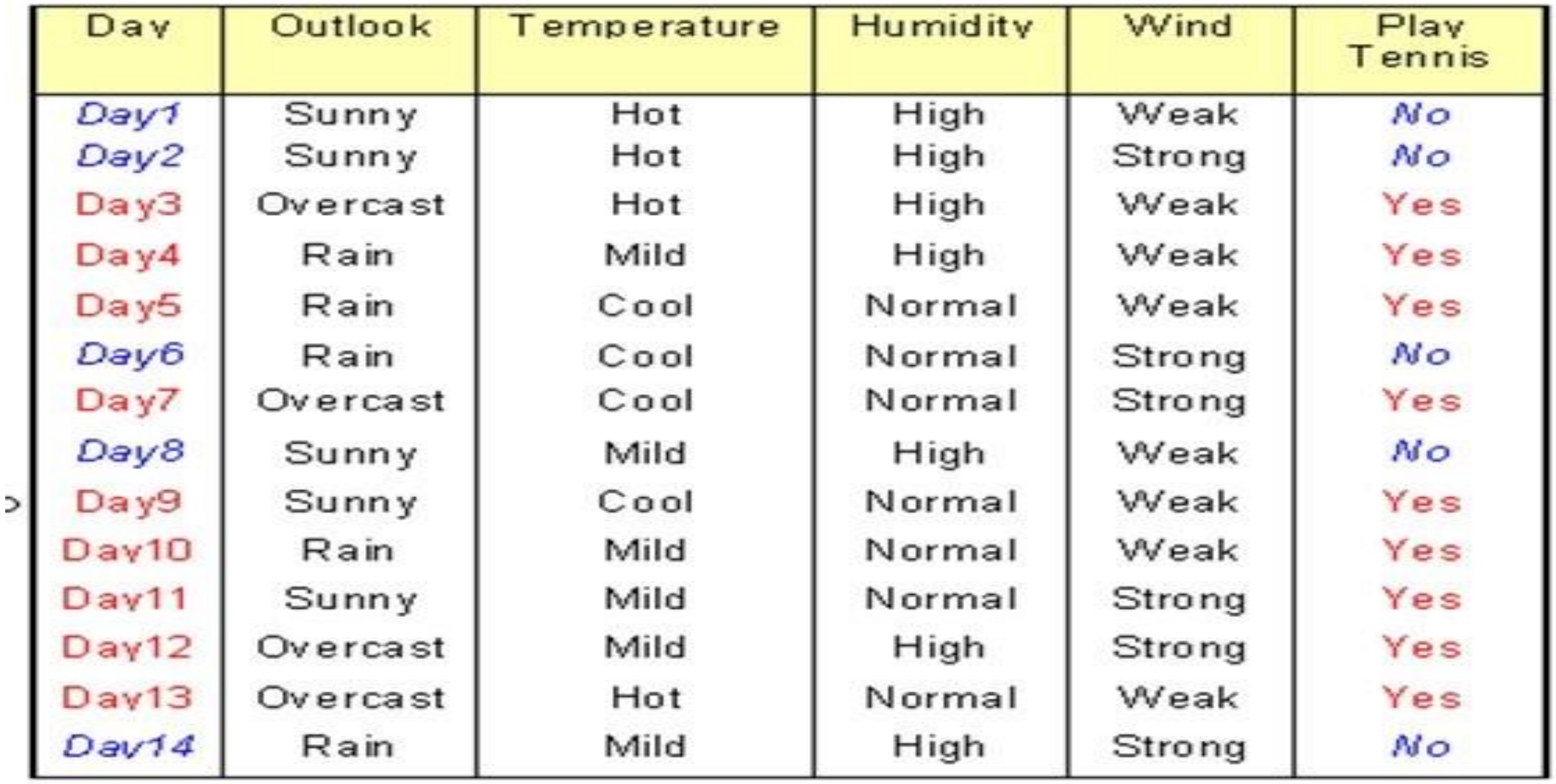
Bayes theorem

$$P\left(c_j \middle| x_j\right) = \frac{P\left(x_j \middle| c_j\right) P(c_j)}{P(x_j)}$$

- ❑ The a priori probability of an instance X_i is independent of the value of the class, so P(x_i) is generally not calculated → **Naïve**
- ❑The idea of the classifier is to choose the most probable class according to the posterior probability P (c_j | x_i) = P (H|D)

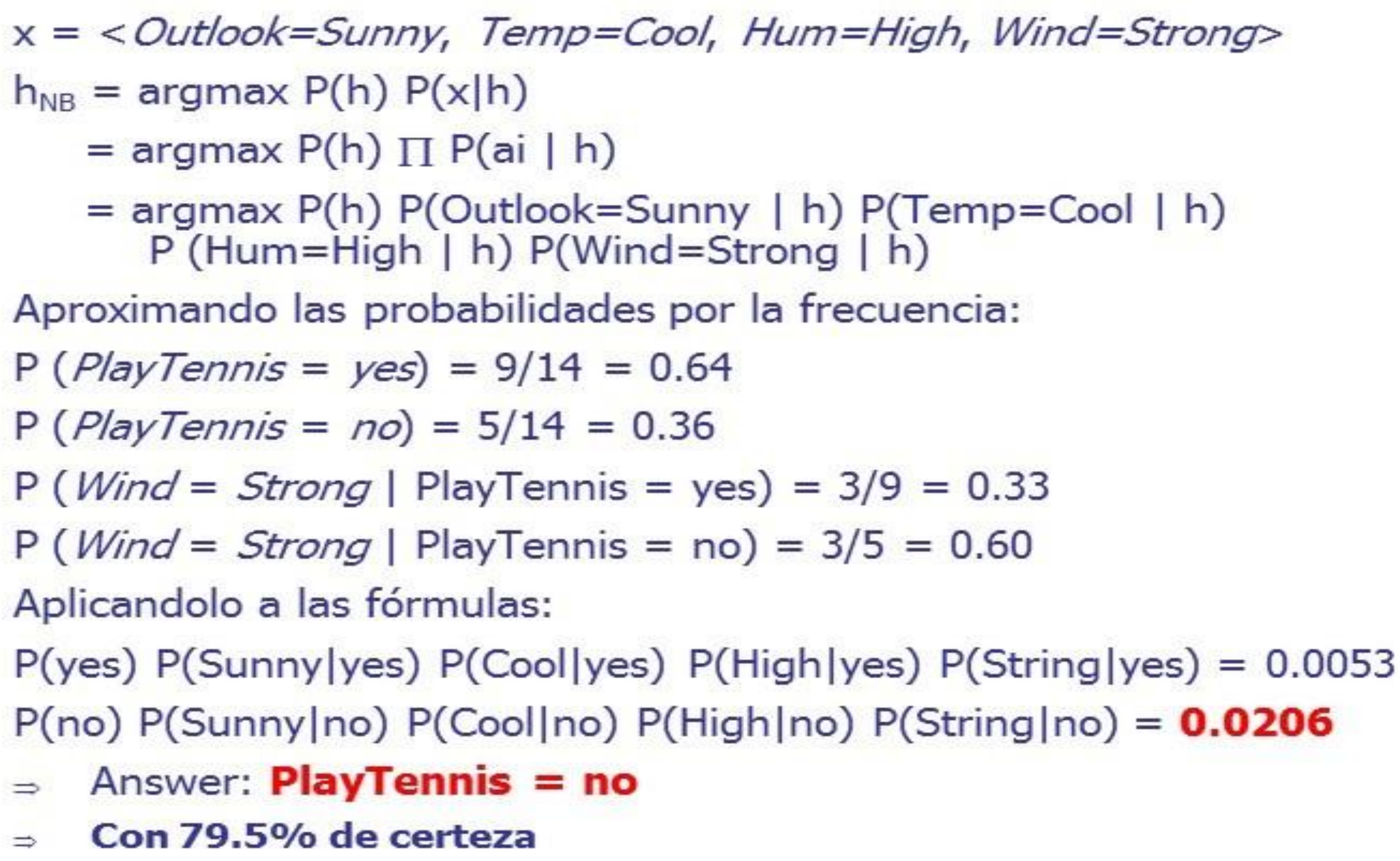
$$BayesianClassifier(x_i) = \underset{c_j \in C}{\text{argmax}}\, P\left(x_i \middle| c_j\right) P\left(c_j\right) = \underset{c_j \in C}{\text{argmax}}\, P\left(c_j\right) \prod_{i \in A}^{n} P\left(x_i \middle| c_j\right)$$

Naïve Bayes - Example



Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Naïve Bayes - Example



x = <*Outlook=Sunny, Temp=Cool, Hum=High, Wind=Strong*>

h_{NB} = argmax P(h) P(x|h)

= argmax P(h) $\prod$ P(ai | h)

= argmax P(h) P(Outlook=Sunny | h) P(Temp=Cool | h)
P (Hum=High | h) P(Wind=Strong | h)

Aproximando las probabilidades por la frecuencia:

P (*PlayTennis* = *yes*) = 9/14 = 0.64

P (*PlayTennis* = *no*) = 5/14 = 0.36

P (*Wind* = *Strong* | PlayTennis = yes) = 3/9 = 0.33

P (*Wind* = *Strong* | PlayTennis = no) = 3/5 = 0.60

Aplicandolo a las fórmulas:

P(yes) P(Sunny|yes) P(Cool|yes) P(High|yes) P(String|yes) = 0.0053

P(no) P(Sunny|no) P(Cool|no) P(High|no) P(String|no) = **0.0206**

⇒ Answer: **PlayTennis = no**

⇒ **Con 79.5% de certeza**

Naïve Bayes: myopic + normal tear rate

- Class priors: $P(c = "n") = \frac{\mathbf{15}}{\mathbf{24}}; P(c = "s") = \frac{5}{24}; P(c = "h") = \frac{4}{24};$
- Likelihoods:

$$P(p = "m", tr = "n" | c = "n") \approx P(p = "m" | c = "n") P(tr = "n" | c = "n") = \frac{7}{15} \times \frac{3}{15}$$

$$P(p = "m", tr = "n" | c = "s") \approx P(p = "m" | c = "s") P(tr = "n" | c = "s") = \frac{2}{5} \times \frac{5}{5}$$

$$P(p = "m", tr = "n" | c = "h") \approx P(p = "m" | c = "h") P(tr = "n" | c = "h") = \frac{3}{4} \times \frac{4}{4}$$

- Evidence: $P(p = "m", tr = "n") \approx \frac{1}{15} \times \frac{3}{15} \times \frac{15}{24} + \frac{2}{5} \times \frac{5}{5} \times \frac{5}{24} + \frac{3}{4} \times \frac{4}{4} \times \frac{4}{24} = \frac{4}{15}$

$Norm = 4/15$

- Posteriors:

$$P(c = "n" | p = "m", tr = "n") = \frac{P(p = "m", tr = "r" | c = "n") P(c = "n")}{P(p = "m", tr = "r")} \approx \frac{\frac{1}{15} \times \frac{3}{15} \times \frac{15}{24}}{Norm} = 0.22$$

Sample = 0.17

$$P(c = "s" | p = "m", tr = "n") = \frac{P(p = "m", tr = "r" | c = "s") P(c = "s")}{P(p = "m", tr = "r")} \approx \frac{\frac{2}{5} \times \frac{5}{5} \times \frac{5}{24}}{Norm} = 0.31$$

Sample= 0.33

$$P(c = "h" | p = "m", tr = "n") = \frac{P(p = "m", tr = "r" | c = "h") P(c = "h")}{P(p = "m", tr = "r")} \approx \frac{\frac{3}{4} \times \frac{4}{4} \times \frac{4}{24}}{Norm} = \mathbf{0.47}$$

Sample= **0.50**

Same decision!

Naïve Bayes

- Advantages
  - In spite of the strong conditional independence assumption, it works surprisingly well in many real-world problems.
  - Even if dependences exist their effects can cancel out.
  - Fast training & prediction.
- Drawbacks
  - The probability estimates are not reliable

https://scikit-learn.org/stable/modules/naive_bayes.html

Estimation of probabilities

❑ Frequency estimates of probabilities can be unreliable, especially when the samples are small

	c = “no”	c = “soft”	c = “hard”
tr = “normal”	3	5	4
tr = “reduced”	12	0	0

$$P(tr = "r" | c = "n") = \frac{12}{15}$$

$$P(tr = "r" | c = "s") = \frac{0}{5}$$

$$P(tr = "r" | c = "h") = \frac{0}{4}$$

This probability is zero. Therefore, all products that involve this term (e.g. in NB) will be zero, which is not reasonable.

Laplace correction

- Avoids zero probability estimates
- More robust estimates
- Asymptotically small

❑ Let $\mathrm{P}(a_i = x_j|c = c_l)$ be the frequency estimate of the probability that attribute ai takes the value v_j (out of K possible values) in examples of class $c = c_1$

$$P(a_i = x_j|c = c_l) = \frac{\text{Nº}\ x_{jl}}{\text{Nº}\ cl}$$

Number of examples of class c_l that have x_j

Number of examples of class c_l

❑ The Laplace corrected estimate of this probability is

$$P(a_i = x_j|c = c_l) = \frac{\text{Nº}\ x_{jl} + \frac{m}{K}}{\text{Nº}\ cl + m}$$

Add m fictitious examples, evenly distributed for the K possible values of attribute a_i

❑ Typically, $m = K$: $$P(a_i = x_j|c = c_l) = \frac{\text{Nº}\ x_{jl} + 1}{\text{Nº}\ cl + K}$$

Laplace correction

- Include fictitious examples (one per possible value of the attribute)

	c = “no”	c = “soft”	c = “hard”
tr = “normal”	3+1	5+1	4+1
tr = “reduced”	12+1	0+1	0+1

$$P(tr = "r" | c = "n") = \frac{12+1}{15+2} = \frac{13}{17}$$

$$P(tr = "r" | c = "s") = \frac{0+1}{5+2} = \frac{1}{7}$$

$$P(tr = "r" | c = "h") = \frac{0+1}{4+2} = \frac{1}{6}$$

Avoids zeros in the probability estimates at cost of introducing some (asymptotically small) bias.

So far we know

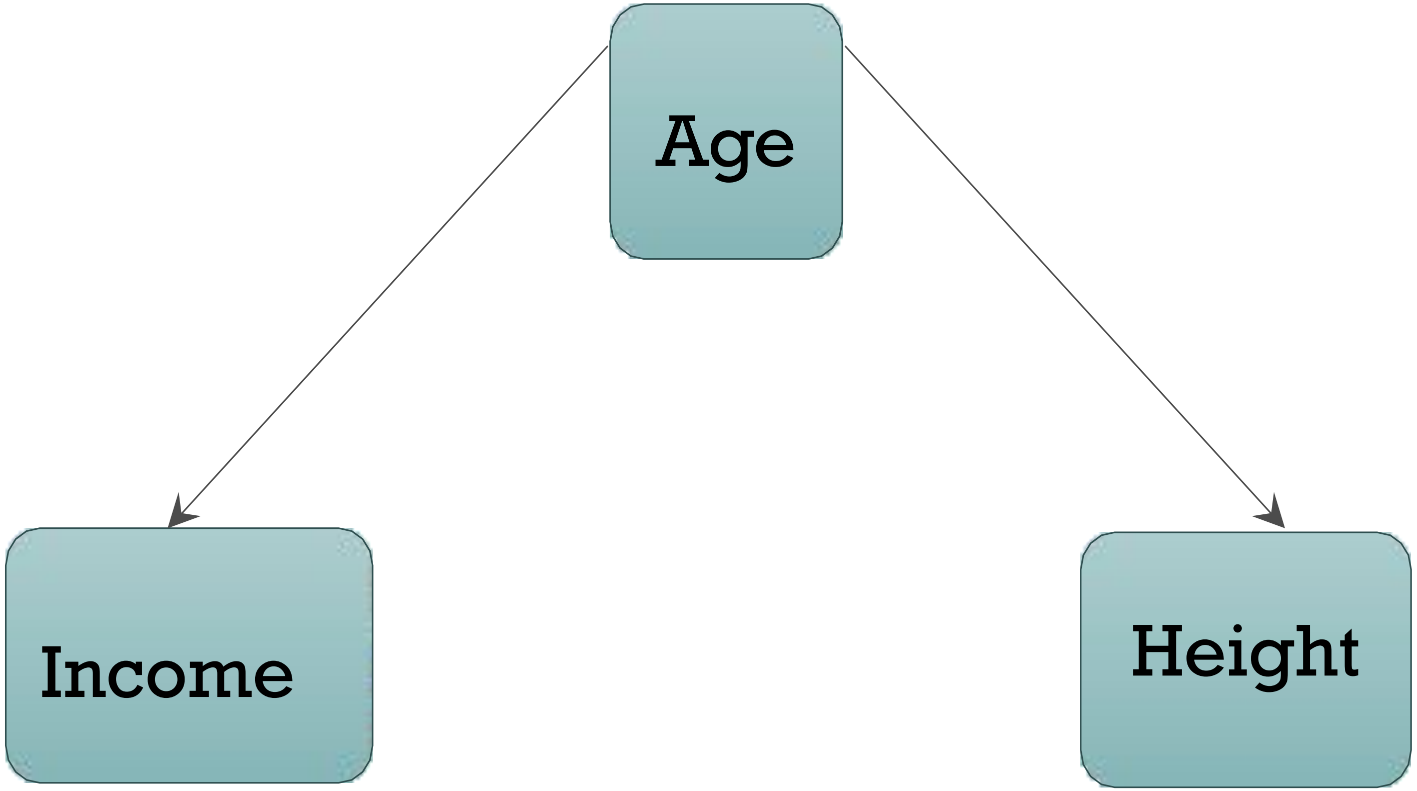
- Representation in domains with random variables + probability distribution
- Given the probability distribution for all possible events, we can solve queries P (Variables | Observation)
- The distribution size is exponential in the number of variables.
- Independence could allow us to reason more efficiently
- But… How can we use probabilities more efficiently?
    - Answer: Bayesian networks

Uncertainty in AI

- Formalization of uncertainty through probabilities
- Bayes theorem in AI
- Bayesian Networks

Conditional independence

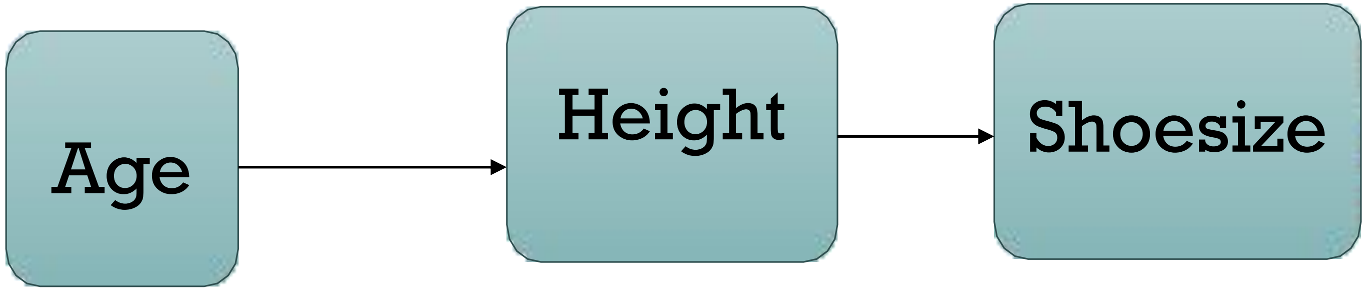
- P(Income|Height,Age) = P(Income|Age)



- Income and Height are conditionally independent “given” Age

Conditional independence

- ❑P(Shoesize|Height,Age) = P(Shoesize|Height)



- ❑Age and Shoesize are conditionally independent “given” Height

Conditional independence

- X and Y are conditionally independent given Z if
  - P (X, Y | Z) = P (X | Z) P (Y | Z)
- As well:
  - P (X | Y, Z) = P (X | Z)
- Often reduces the number of parameters from exponential in n (number of variables) to linear in n
- Conditional independence is an efficient probabilistic reasoning tool
  - less parameters
  - less computation
- It is represented by the missing axis

4.3 Definition of a Bayesian network

- ❑A set of nodes
  - ❑each node represents a random variable
  - ❑variables can be either discrete or continuous
- ❑A set of edges
  - ❑an edge from node X to node Y: X has a direct influence on Y
  - ❑it is a Direct Acyclic Graph (DAG)
- ❑Probability distributions
  - ❑each node X has a Conditional Probability Table (CPT) that defines the effects of its parents

$$P(Node|Parents(Node))$$

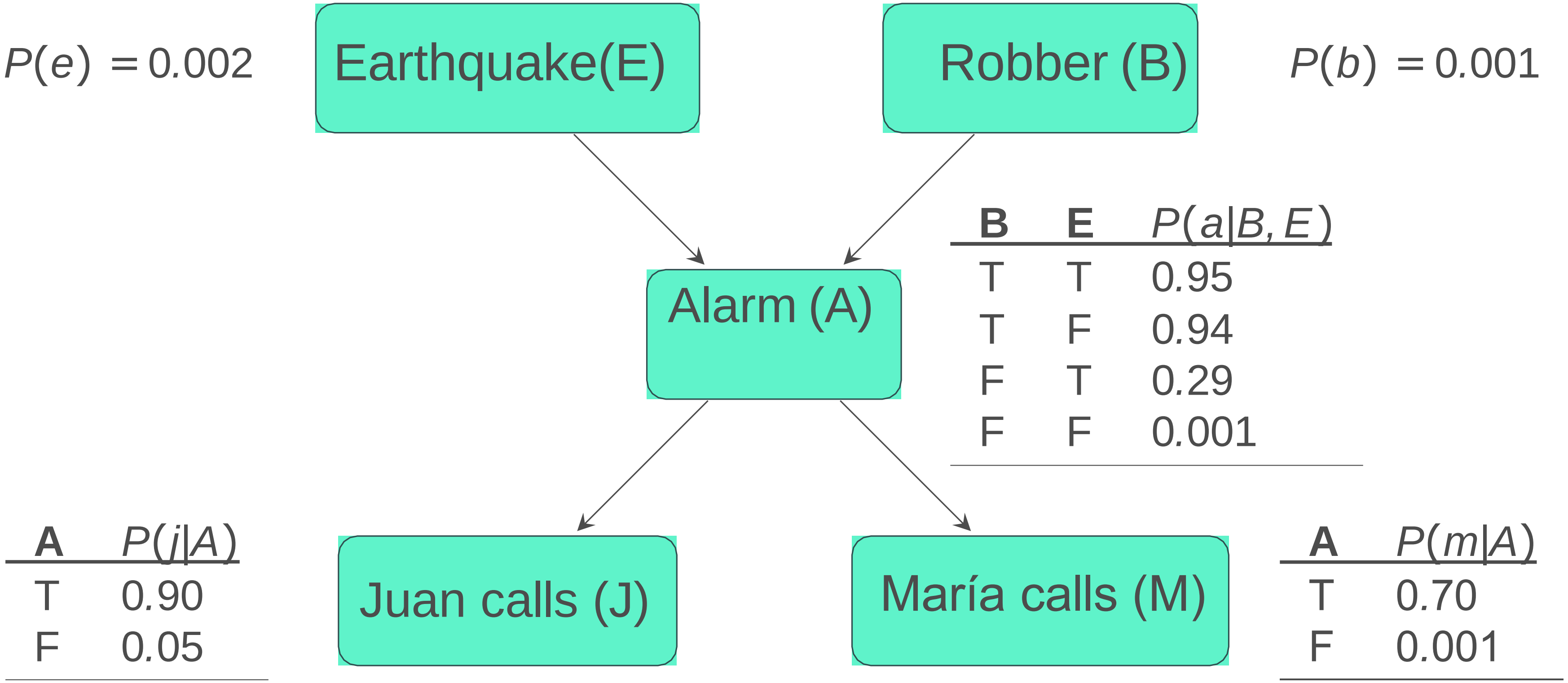
  - ❑parents of node X are the only edges directed to X
  - ❑if a node does not have parents, it is the "a priori" probability

$$P(Node)$$

Example of an alarm

- We have an anti-theft system at home with an alarm
- It detects robbers, but the alarm also fires with some earthquakes
- There are two neighbours (Juan and Maria) that will call us if they hear the alarm
- Juan always calls when he hears the alarm, but he sometimes is confused with some door bell
- Maria hears music very loud, so sometimes she cannot hear the alarm
- Modelling
    - Earthquake: E E=T ~ e E=F ~ ¬ e
    - Robbery: B B=T ~ b B=F ~ ¬ b
    - Alarm: A (a, ¬ a)
    - Juan calls: J (j, ¬ j)
    - María calls: M (m, ¬ m)

Complete BN for the Alarm example



$P(e) = 0.002$

Earthquake(E)

Robber (B)

$P(b) = 0.001$

Alarm (A)

B	E	$P(a\|B,E)$
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

Juan calls (J)

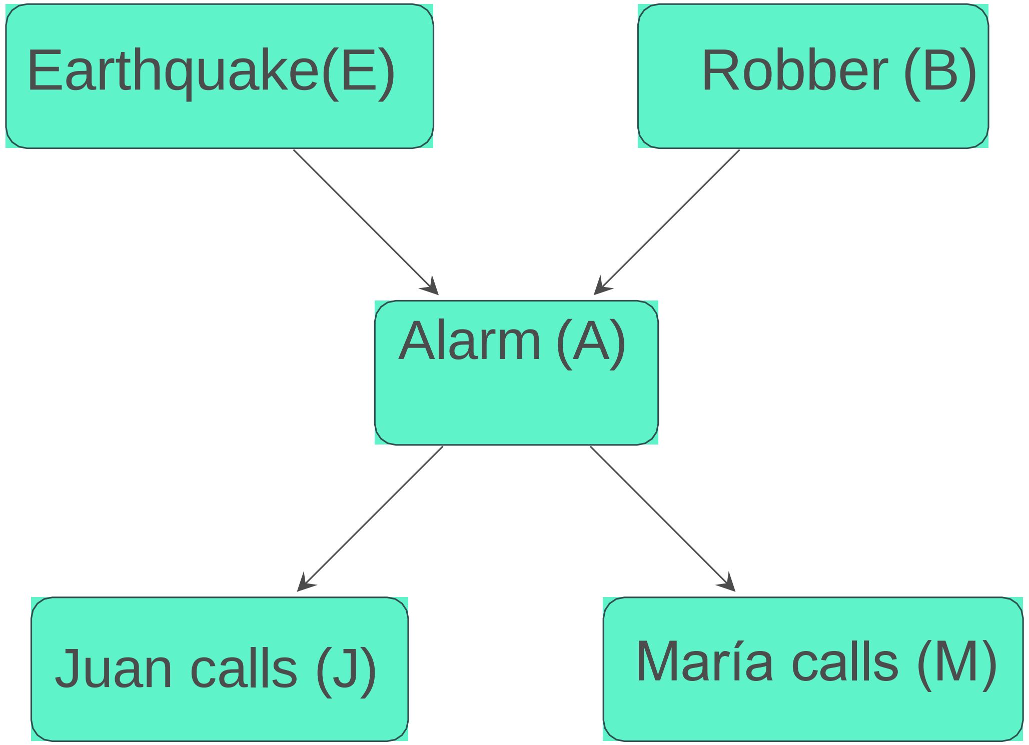
María calls (M)

A	$P(j\|A)$
T	0.90
F	0.05

A	$P(m\|A)$
T	0.70
F	0.001

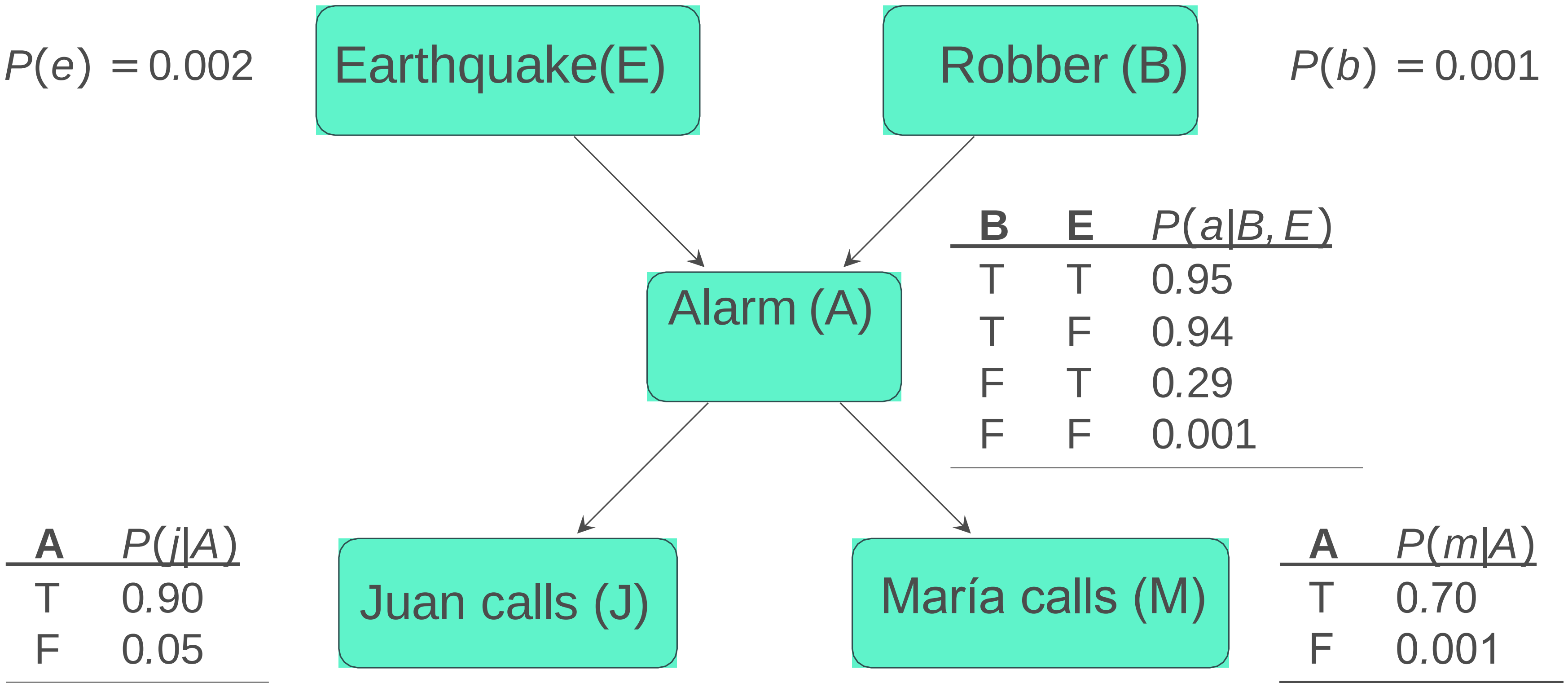
- ❑We only provide P(e) given that $P(\neg e) = 1 - P(e)$
- ❑Also, $P(\neg a \mid b, \neg e) = 1 - P(a \mid b, \neg e)$
- ❑The topology of this BN reflects the direct causes of its variables:
  - ❑a robber can fire the alarm
  - ❑an earthquake can fire the alarm
  - ❑the alarm can cause Maria to call
  - ❑the alarm can cause Juan to call

BN for the example



- ❑There is no dependency between Earthquake and Robbery
- ❑But, there is dependency between Alarm and the other two variables:
  - ❑P(Alarm|Earthquake, Robbery) ƒ= P(Alarm|Earthquake)
  - ❑P(Alarm|Earthquake, Robbery) ƒ= P(Alarm|Robbery)
- ❑There is conditional independence between Juan calling and variables Earthquake and Robbery, given the Alarm variable. And the same for María
  - ❑P(Juan|Alarm, Earthquake, Robbery) = P(Juan|Alarm)
  - ❑P(Maria|Alarm, Earthquake, Robbery) = P(Maria|Alarm)

BN are compact



$P(e) = 0.002$

$P(b) = 0.001$

B	**E**	$P(a\|B,E)$
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	$P(j\|A)$
T	0.90
F	0.05

A	$P(m\|A)$
T	0.70
F	0.001

- The explicit joint distribution would require $2^5 - 1 = 31$ parameters
- The BN uses 1 + 1 + 4 + 2 + 2 = 10 parameters

Semantics of BNs

- Global semantics: the joint probability distribution is the product of local distributions

$$\mathrm{P}(\mathrm{X}_1, \ldots, \mathrm{X}_n) = \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1}) = \prod_{i=1}^{n} P(X_i \mid \mathrm{Parents}(X_i))$$

- Origin: Chain rule

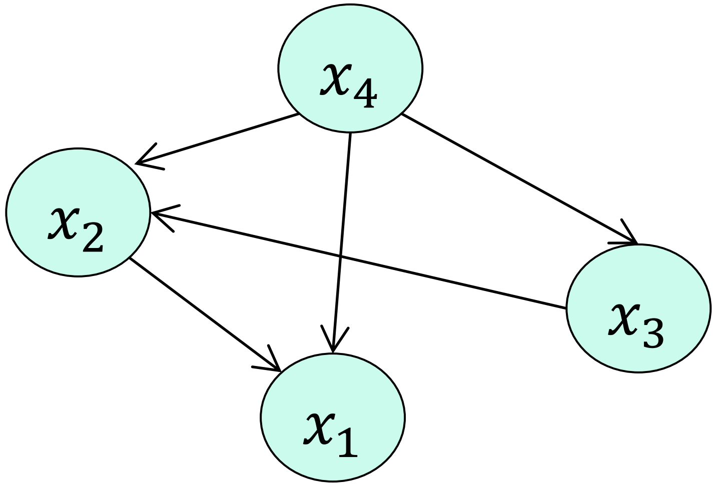
$$P(X_1, X_2, X_3, X_4) = P(X_1 \mid X_2, X_3, X_4)\, P(X_2, X_3, X_4) =$$

$$P(X_1 \mid X_2, X_3, X_4)\, P(X_2 \mid X_3, X_4)\, \mathrm{P}(X_3, X_4) =$$

$$P(X_1 \mid X_2, X_3, X_4)\, P(X_2 \mid X_3, X_4)\, \mathrm{P}(X_3 \mid X_4)\, \mathrm{P}(X_4)$$

Interpretation of the graph

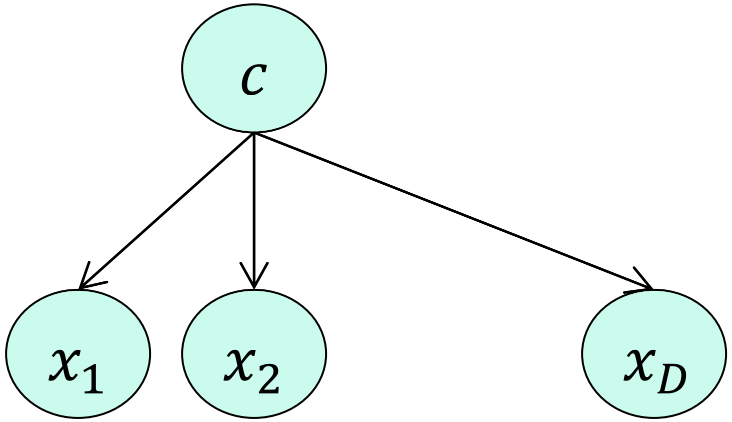
$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{Parents}(X_i))$$



- ❑ $P(x_1, x_2, x_3, x_4) = P(x_1 | x_2, x_4) P(x_2 | x_3, x_4) P(x_3 | x_4) P(x_4)$

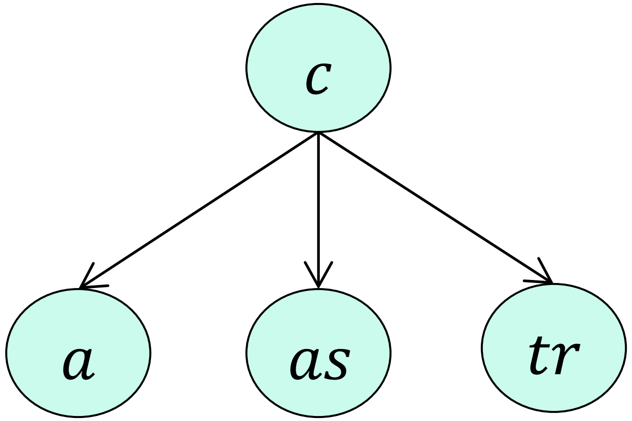
Naïve Bayes graph

$$P(\mathrm{x}, c) = \prod_{i=1}^{N} P(x_i|c)P(c)$$



Young + astigmatism + normal tear rate

Prioris $P(c='n')=\frac{15}{24}$ $P(c='s')=\frac{5}{24}$ $P(c='\ h')=\frac{4}{24}$



Conditional marginals

	no	soft	hard
a =‘y’	4	2	2
Total	15	5	4

	no	soft	hard
as = ‘y’	8	0	4
Total	15	5	4

	no	soft	hard
tr =‘normal’	3	5	4
Total	15	5	4

$P(a="y", as="y"; tr="n")\sim$ Norm =0.1011

- Naïve Bayes: We assume independence

$$P(c="n"|a="y",as="y";tr="n")=\frac{P(a="y",as="y";tr="n"|c="n")P(c="n")}{P(a="y",as="y";tr="n")}$$

$$\approx\frac{P(a="y"|c="n")P(as="y"|c="n")P(tr="n"|c="n")P(c="n")}{Norm}=\frac{\frac{4}{15}\times\frac{8}{15}\times\frac{3}{15}\times\frac{15}{24}}{Norm}=0.18$$

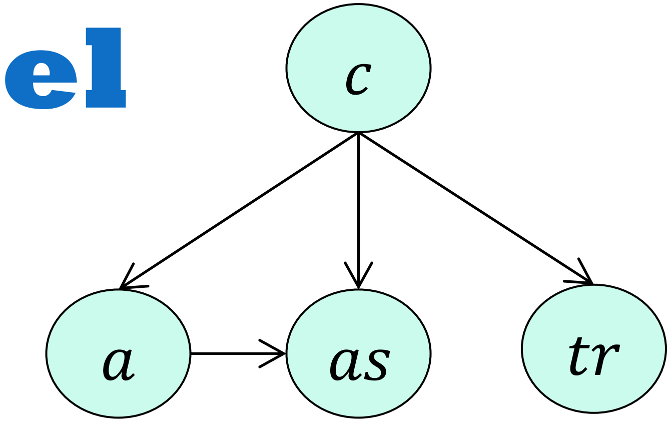
$$P(c="s"|a="y",as="y";tr="n")=\frac{P(a="y",as="y";tr="n"|c="s")P(c="s")}{P(a="y",as="y";tr="n")}$$

$$\approx\frac{P(a="y"|c="s")P(as="y"|c="s")P(tr="n"|c="n")P(c="s")}{Norm}=\frac{\frac{2}{5}\times\frac{0}{5}\times\frac{5}{5}\times\frac{5}{24}}{Norm}=0.00$$

$$P(c="h"|a="y",as="y";tr="n")=\frac{P(a="y",as="y";tr="n"|c="h")P(c="h")}{P(a="y",as="y";tr="n")}$$

$$\approx\frac{P(a="y"|c="h")P(as="y"|c="h")P(tr="n"|c="h")P(c="h")}{Norm}=\frac{\frac{2}{4}\times\frac{4}{4}\times\frac{4}{4}\times\frac{4}{24}}{Norm}=\mathbf{0.82}$$

A more sophisticated model



Prioris $P(c='n') = \frac{15}{24}$ $P(c='s') = \frac{5}{24}$ $P(c='\,h') = \frac{4}{24}$

Conditional marginal: **young + astigmatism + normal**

	no	soft	hard
a ='y'	4	2	2
Total	15	5	4

	C=nor age=young	C= Soft age=young	C= Hard age=young
as = 'y'	2	0	2
Total	4	2	2

	no	soft	hard
tr ='normal'	3	5	4
Total	15	5	4

$P(a = \text{"y"}, as = \text{"y"}; tr = \text{"}n\text{"})\sim$ Norm =0.1011

- ❑ Bayesian network

$$P(c = \text{"}n\text{"} | a = \text{"y"}, as = \text{"y"}; tr = \text{"}n\text{"}) = \frac{P(a = \text{"y"}, as = \text{"y"}; tr = \text{"}n\text{"} | c = \text{"}n\text{"})P(c = \text{"}n\text{"})}{P(a = \text{"y"}, as = \text{"y"}; tr = \text{"}n\text{"})}$$

$$\approx \frac{P(a = \text{"y"} | c = \text{"}n\text{"})P(as = \text{"y"} | a = \text{"y"}, c = \text{"n"})P(tr = \text{"}n\text{"} | c = \text{"}n\text{"})P(c = \text{"}n\text{"})}{Norm} = \frac{\frac{4}{15} \times \frac{2}{4} \times \frac{3}{15} \times \frac{15}{24}}{Norm} = 0.17$$

$$P(c = \text{"}s\text{"} | a = \text{"y"}, as = \text{"y"}; tr = \text{"}n\text{"}) = \frac{P(a = \text{"y"}, as = \text{"y"}; tr = \text{"}n\text{"} | c = \text{"}s\text{"})P(c = \text{"}s\text{"})}{P(a = \text{"y"}, as = \text{"y"}; tr = \text{"}n\text{"})}$$

$$\approx \frac{P(a = \text{"y"} | c = \text{"}s\text{"})P(as = \text{"y"} | a = \text{"y"}, c = \text{"}s\text{"})P(tr = \text{"}n\text{"} | c = \text{"}n\text{"})P(c = \text{"}s\text{"})}{Norm} = \frac{\frac{2}{5} \times \frac{0}{2} \times \frac{5}{5} \times \frac{5}{24}}{Norm} = 0.00$$

$$P(c = \text{"}h\text{"} | a = \text{"y"}, as = \text{"y"}; tr = \text{"}n\text{"}) = \frac{P(a = \text{"y"}, as = \text{"y"}; tr = \text{"}n\text{"} | c = \text{"}h\text{"})P(c = \text{"}h\text{"})}{P(a = \text{"y"}, as = \text{"y"}; tr = \text{"}n\text{"})}$$

$$\approx \frac{P(a = \text{"y"} | c = \text{"}h\text{"})P(as = \text{"y"} | a = \text{"y"}, c = \text{"}h\text{"})P(tr = \text{"}n\text{"} | c = \text{"}h\text{"})P(c = \text{"}h\text{"})}{Norm} = \frac{\frac{2}{4} \times \frac{2}{2} \times \frac{4}{4} \times \frac{4}{24}}{Norm} = \mathbf{0.83}$$