INDIAN INSTITUTE OF INFORMATION TECHNOLOGY

Dharwad

ज्ञानेन विकासः

Probabilistic Models

Dr. Animesh Chaturvedi

Assistant Professor: **IIIT Dharwad**

Young Researcher: **Heidelberg Laureate Forum** and **Pingala Interaction in Computing**

Young Scientist: **Lindau Nobel Laureate Meetings**

Postdoc: **King's College London** & **The Alan Turing Institute**

PhD: **IIT Indore** MTech: **IIITDM Jabalpur**

LINDAU NOBEL LAUREATE MEETINGS

HEIDELBERG LAUREATE FORUM

KING'S College LONDON

The Alan Turing Institute

Indian Institute of Technology Indore

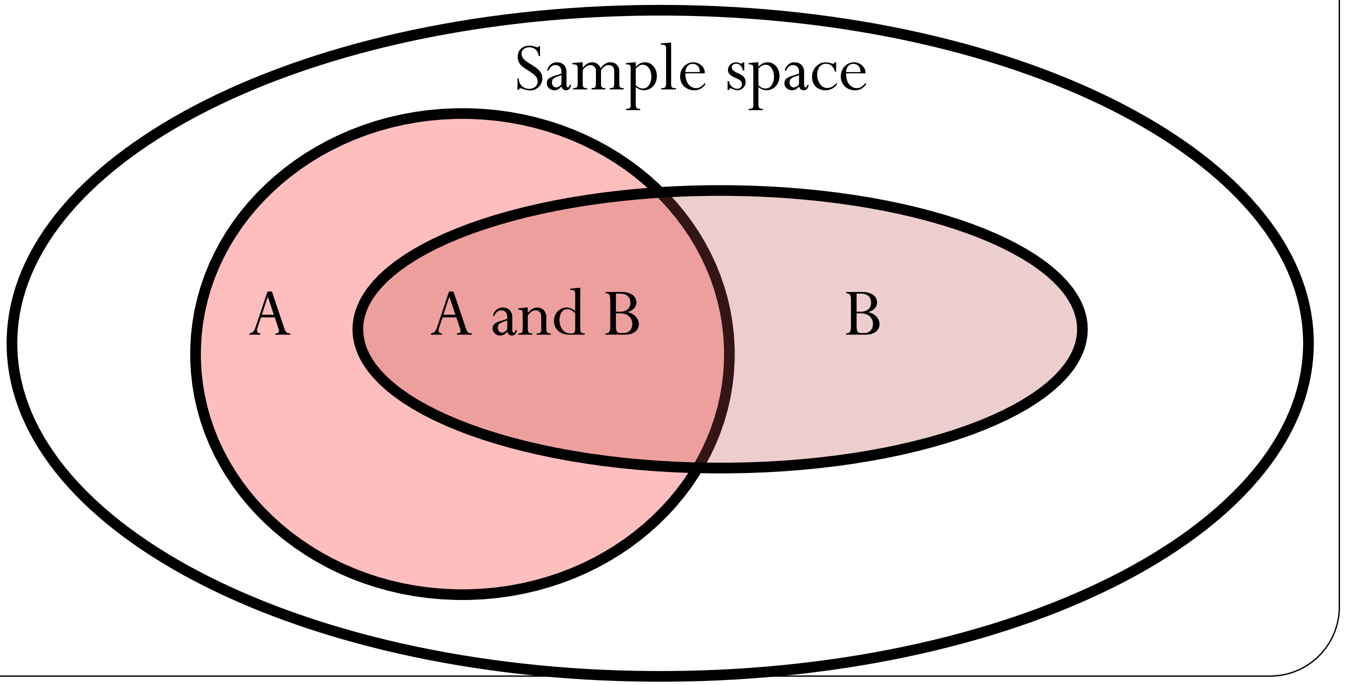
भारतीय प्रौद्योगिकी संस्थान इंदौर

PDPM Indian Institute of Information Technology, Design and Manufacturing, Jabalpur

Conditional probability

- Bayes' theorem relates the conditional and marginal probabilities of stochastic events A and B:
- P(A | B) = P(A and B) / P(B)
- P(A | B)P(B) = P(A and B)
- P(A | B) = P(B | A)P(A)/P(B)
  - Bayes' rule

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$



http://en.wikipedia.org/wiki/Bayes'_theorem

Derivation

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

$$\Pr(B \mid A) = \frac{\Pr(A \cap B)}{\Pr(A)}$$

Combining these 2 equations:

$$\Pr(A \mid B)\Pr(B) = \Pr(B \mid A)\Pr(A)$$

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}$$

Conditioning with Dependence

- X: is it raining?
  - $P(X=1) = 0.3$
- Y: are the sprinklers on?
  - $P(Y=1) = 0.4$
  - X and Y are independent
- Z: is the grass wet?
  - $P(Z=1 \mid X=0, Y=0) = 0.1$
  - $P(Z=1 \mid X=0, Y=1) = 0.8$
  - $P(Z=1 \mid X=1, Y=0) = 0.7$
  - $P(Z=1 \mid X=1, Y=1) = 0.9$

Not wet

	Raining	Not raining
Sprinklers	0.012	0.056
No sprinklers	0.054	0.378

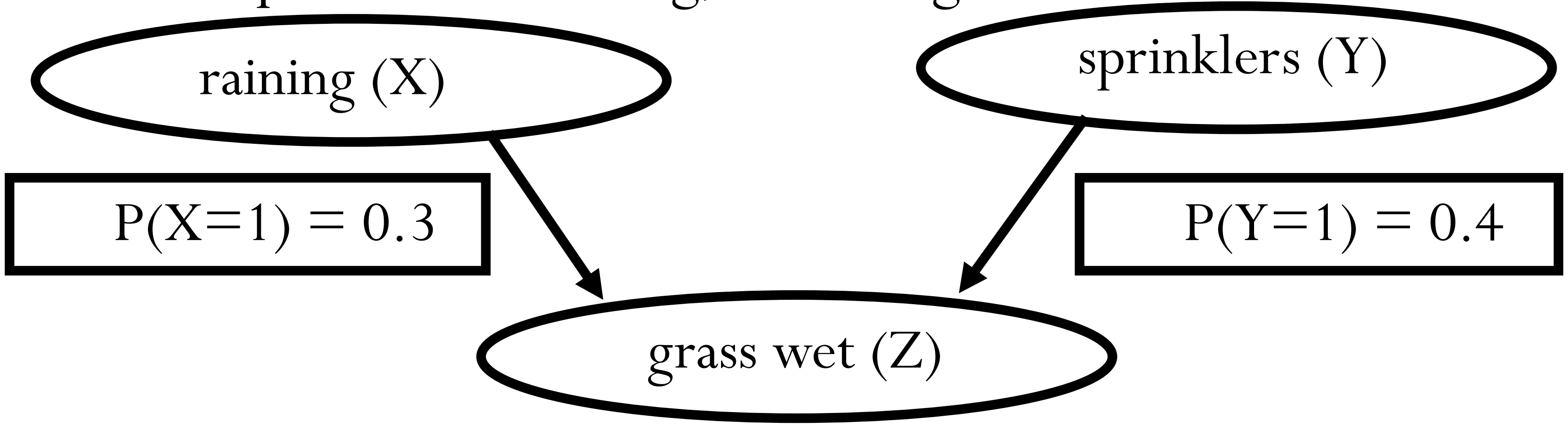
Wet

	Raining	Not raining
Sprinklers	0.108	0.224
No sprinklers	0.126	0.042

- Conditional on Z=1, X and Y are **not** independent
- If you know Z=1, rain seems likely; then if you also find out Y=1, this “explains away” the wetness and rain seems less likely

Rain and sprinklers example

- sprinklers is independent of raining, so no edge between them



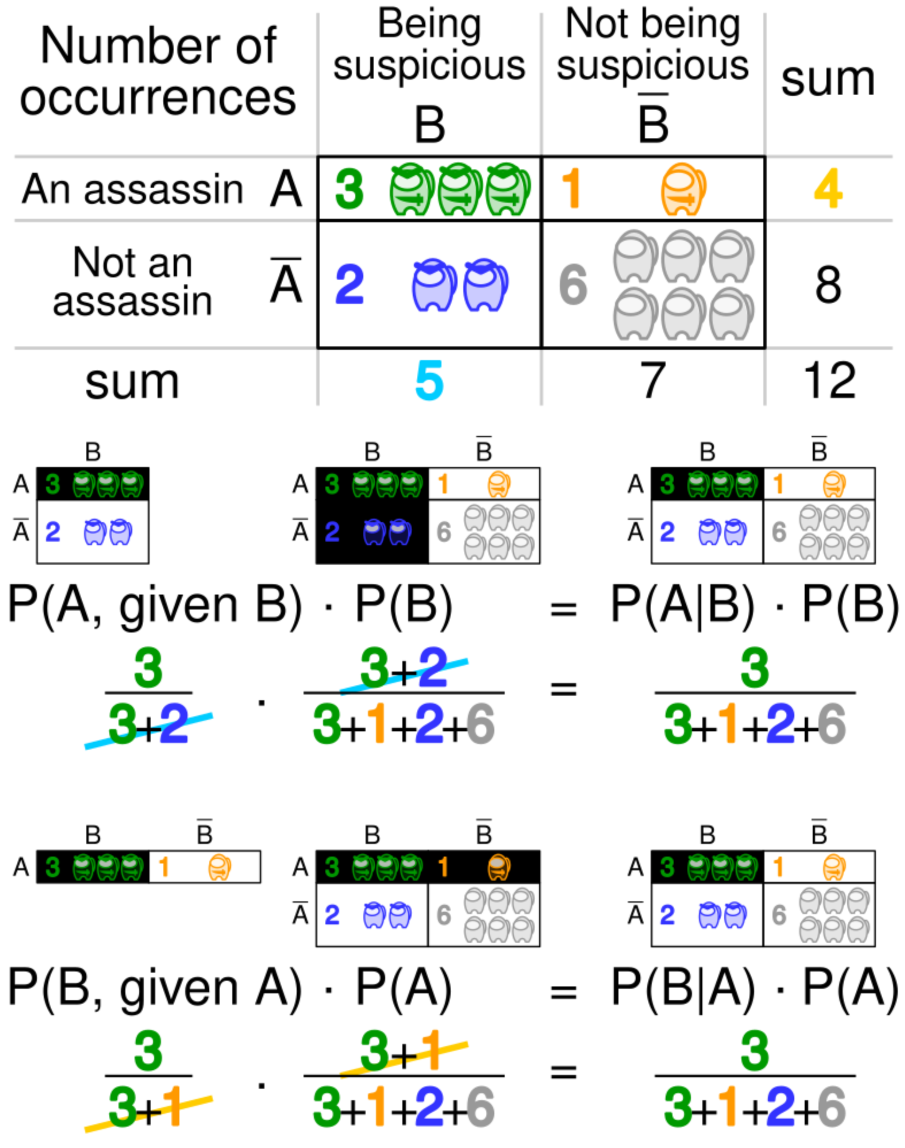
Each node has a conditional probability table (CPT)

$P(Z=1 \mid X=0, Y=0) = 0.1$
$P(Z=1 \mid X=0, Y=1) = 0.8$
$P(Z=1 \mid X=1, Y=0) = 0.7$
$P(Z=1 \mid X=1, Y=1) = 0.9$

Example

Example 1

- 2 cookie bowls
  - Bowl 1: 10 chocolate-chip, 30 plain
  - Bowl 2: 20 chocolate-chip, 20 plain
- Buck picks a plain cookie from one of the bowls, but which bowl?
  - $\Pr(A) = \text{Bowl 1} = 0.5,\ 1 - \Pr(A) = \text{Bowl 2}$
  - $\Pr(B) = \text{Plain cookie} = 50/80 = 0.625$
  - $\Pr(B \mid A) = 30/40 = 0.75$
  - $\Pr(A \mid B) = 0.75 \text{ x } 0.5/0.625 = 0.6$



Example 2

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

$$\therefore \quad P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Bayes' Theorem for a given parameter θ

- $p\ (\theta \mid \text{data}) = p\ (\text{data} \mid \theta)\ p\ (\theta)\ /\ p\ (\text{data})$

  $\downarrow \qquad \downarrow \qquad \downarrow$

  1/p(data) is basically a normalizing constant

- Posterior ∝ likelihood x prior
- Prior is the probability of the parameter and represents what was thought before seeing the data.
  Prior Distribution – use probability to quantify uncertainty about unknown quantities (parameters)
- Likelihood is the probability of the data given the parameter and represents the data now available.
  Likelihood – relates all variables into a "full probability model"
- Posterior represents what is thought given both prior information and the data just seen.
  Posterior Distribution – result of using data to update information about unknown quantities (parameters)
- It relates the conditional density of a parameter (posterior probability) with its unconditional density (prior, since depends on information present before the experiment).

Bayesian inference

- Prior information $p(\theta)$ on parameters θ
- Likelihood of data given parameter values $f(y \mid \theta)$
- Posterior distribution is proportional to likelihood $\times$ prior distribution.
- Not generally necessary to compute this integral.

$$p(\theta \mid y) = \frac{f(y \mid \theta)\, p(\theta)}{f(y)}$$

or

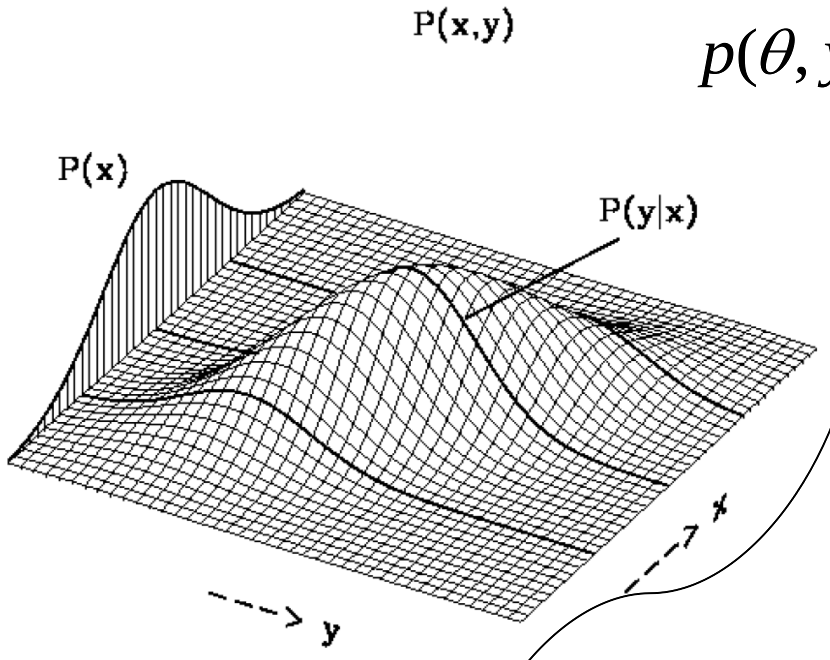
$$\pi(\theta \mid y) \propto f(y \mid \theta)\, p(\theta)$$

or

$$f(y) = \int_{-\infty}^{\infty} f(y \mid \theta)\, p(\theta)\, d\theta$$

Bayesian inference

- To make probability statements about θ given y we begin with a model



$$p(\theta, y) = p(\theta)\,p(y \mid \theta) \longleftarrow \text{joint prob. distribution}$$

$$p(\theta \mid y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)\,p(y \mid \theta)}{p(y)}$$

where

$$p(y) = \sum\nolimits_{\theta} p(\theta) p(y \mid \theta) \longleftarrow \text{discrete case}$$

or

$$p(y) = \int_{\Theta} p(\theta) p(y \mid \theta) \longleftarrow \text{continuous case}$$

posterior prior likelihood

$$\boxed{p(\theta \mid y) \propto p(\theta)\,p(y \mid \theta)}$$

Maximum A Posterior

- Based on Bayes Theorem, compute the *Maximum A Posterior* (MAP) hypothesis for the data
- Best hypothesis for some space H given observed training data D.

$$h_{MAP} \equiv \underset{h \in H}{\operatorname{argmax}} P(h \mid D)$$

$$= \underset{h \in H}{\operatorname{argmax}} \frac{P(D \mid h) P(h)}{P(D)}$$

$$= \underset{h \in H}{\operatorname{argmax}} P(D \mid h) P(h)$$

- *H*: set of all hypothesis.
- Drop *P(D)* as the probability of the data is constant (and independent of hypothesis).

Maximum Likelihood

- Assume that all hypotheses are equally probable a priori,
  - i.e., $P(h_i) = P(h_j)$ for all h_i, h_j belong to H.

- This is called assuming a *uniform prior*. It simplifies computing the posterior:

$$h_{ML} = \arg\max_{h \in H} P(D \mid h)$$

- This hypothesis is called the *maximum likelihood hypothesis.*

Summary

- Bayesian methods use probability models for quantifying uncertainty in inferences based on statistical data analysis. Bayesian estimation

1. **Priors over the parameters** start with the ***formulation of a model*** that we hope is adequate to describe the situation of interest.
2. **Posterior distributions *observe the*** of belief (probability).
3. **New priors over the parameters** evaluate the fit of the model. If necessary, we compute predictive distributions for future observations.

Prejudices or scientific judgment?

The selection of a prior is subjective and arbitrary.	It is reasonable to draw conclusions in the light of some reason.

References

- Vincent Conitzer, CPS 270: Artificial Intelligence, Introduction to probability, Department of Computer Science, Duke, http://www.cs.duke.edu/courses/fall08/cps270/

- Raymond J. Mooney, CS 343: Artificial Intelligence, University of Texas at Austin

תודה רבה
Hebrew

Ευχαριστώ
Greek

Спасибо
Russian

Danke
German

धन्यवादः
Sanskrit

Merci
French

ধন্যবাদ
Bangla

நன்றி
Tamil

شكرا
Arabic

ಧನ್ಯವಾದಗಳು
Kannada

Thank You
English

നന്ദി
Malayalam

多謝
Traditional Chinese

Grazie
Italian

ధన్యవాదాలు
Telugu

આભાર
Gujarati

ਧੰਨਵਾਦ
Punjabi

धन्यवाद
Hindi & Marathi

多谢
Simplified Chinese

Gracias
Spanish

https://sites.google.com/site/animeshchaturvedi07

Obrigado
Portuguese

ありがとうございました
Japanese

ขอบคุณ
Thai

감사합니다
Korean